Epidemiology: An epistemological perspective

Alfredo Morabia

To the Epistemology Group of the Wade Hampton Frost Reading Room, Baltimore, 1986–1988

1. Introduction

1.1. A contextual name

The term "epidemiology" is a source of confusion about the nature of this discipline. For the public, "epidemiology" evokes a medical discipline that deals with large-scale outbreaks of infectious diseases. This was indeed its meaning in the first treatises which included "epidemiology" in their titles. In the 16th century the Spanish physician Angelerio published a study on plague entitled "*Epidemiología*" and in 1802 another Spanish physician, Villalba, wrote a compilation of epidemics and outbreaks over 13 centuries entitled "*Epidemiología Española*" (meaning Spanish epidemiology) (Pan American Health Organization, 1988, p. 3–4).

The term epidemiology was also quite accurate when the discipline made its first steps. It reflected the particular historical context in 19th century England, when epidemics of infectious diseases, and in particular cholera, were the main scourges whose causes had to be identified. The *London Epidemiologic Society*, created in 1850, assembled scientists, public health practitioners and physicians to unite their efforts in the fight against "epidemics". Today, epidemiology is still associated with the fight against infection, in all types of contexts, including emerging diseases (e.g., severe acute respiratory syndrome or SARS), bioterrorism (e.g., criminal dissemination of anthrax bacteria), and even digital viruses! Isn't it meaningful that a bioinformatician, Alberto-Laszlo Barabasi, describes "computer security experts" as "a new breed of epidemiologists who vigilantly monitor the health of our online universe", protecting it from international viruses capable of causing life-threatening emergencies (Barabasi, 2002, p. 141)?

Even though the name continues to evoke the fight against infectious plagues, the domain of epidemiology has enormously expanded and is not restricted to specific types of diseases. If we had to name the discipline today, we would probably give it a different name. Physics, chemistry, or medical specialties such as cardiology or neurology have names that are unambiguous, because they describe the subject of the

A History of Epidemiologic Methods and Concepts, edited by Alfredo Morabia

discipline. The name "epidemiology", on the other hand, has more to do with the circumstances in which the discipline was born than with the substance of the discipline in its present state.

What is then the subject of epidemiology? We can rephrase the question and ask: Why would one hire an epidemiologist today rather than a statistician, a sociologist, a clinician, etc.? An epidemiologist is expected to have learned a particular set of methods and concepts taught in epidemiology classes which are needed to identify determinants of health and disease: describe states of health in populations, investigate outbreaks of diseases, compare groups, use, with increasing degrees of complexity, the concepts of bias, confounding and interaction and be familiar with the epidemiologic approaches to causal inference. This would however still be an insufficient reason. It is like saying that we need cardiologists because they know how to use a stethoscope. Most physicians use stethoscopes and most public health professions are familiar with basic epidemiology and use it. However, they do not necessarily master it. Thus, what defines an epidemiologist is probably the ability to *adapt* this particular set of methods and concepts to specific research questions. This ability allows them, in more exceptional circumstances, to make methods and concepts *evolve* when encountering new types of problems.

The subject of epidemiology is therefore the investigation of causes of health-related events in populations. A name more closely reflecting this subject would be "population health etiology", etiology meaning "science of causation".

1.2. Historical contribution of epidemiology

Science allows us to understand how our world is and how it works. We identify causal links and these indicate ways to act upon the world and modify it. What is then the historical contribution of epidemiology to knowledge?

Epidemiology is a recent scientific discipline. It has roots in the 17th century but it is really a 19th century science. Its mission has historically been to identify determinants of human diseases (and later health), mostly at the population level. Epidemiologic discoveries can be used for improving human health. Probably one of the most important discoveries, for its scientific and public health impacts, has been the demonstration that cigarette smoke *caused* lung cancer in smokers, and that preventing exposure to cigarette smoke could prevent occurrence of lung cancer.

To identify causes we can act upon, we need methods, that is, strategies or experiments, which are organized in such a way that their results can reveal stable relations and laws. Experiments can be of different sorts. In some (typically laboratory) experiments, the researcher can manipulate exposure to assess changes in outcome. This is not the main type of experiments in epidemiology. Epidemiologists usually compare groups of people differing on carefully selected characteristics, but they cannot manipulate exposure. For example, they cannot allocate a certain number of cig-

arettes to be smoked per day. The fundamental observation establishing that cigarette smoke caused lung cancer was that smokers tended to develop lung cancer *more* frequently *than* non-smokers. Because they are mostly based on observations (as opposed to interventions), epidemiologic experiments have considerable complexity. If lung cancer is more common in smokers than in non-smokers, this does not mean yet that smoking causes lung cancer. Smokers and non-smokers could differ on one or several characteristics, which are the true causes of lung cancer (i.e., confounding). The disease could have multiple causes, whose pathways overlap with other pathways (i.e., interaction). The experiment itself is prone to errors, which can interfere with the sound interpretation of its results (i.e., biases). Therefore comparing groups would be a very naive endeavor if it were not supported by a theory that allowed epidemiologists to design experiments, organize the facts and interpret the observations in a way that takes into account the complexity of the matter studied. I will refer to the elements of this theoretical framework as *concepts*. We will see that confounding, interaction and bias are examples of concepts. We did not simply observe them. They are intellectual constructions that were (and are) refined over time.

Thus, historically, the specific contribution of epidemiology has been the progressive constitution of a coherent ensemble of methods and concepts, aimed to assess health determinants. We will see that it was based on two principles: population thinking and group comparisons.

1.3. Theme of this essay

This essay is about the genesis of epidemiology as a scientific discipline. Its theme is that current epidemiologic concepts and methods have evolved since the 18th century in a series of relatively well-defined steps to constitute an integrated theory based on two essential principles: 1) population thinking and 2) group comparisons.

Population thinking, as opposed to individual thinking, is a mode of conceptualizing issues for a whole group of people defined in a specific way (e.g., geographically, socially, biologically). The entire group is the population. In 1950, John E. Gordon, Professor of Epidemiology and Preventive Medicine at the Harvard School of Public Health, expressed the essence of population thinking when he stressed that each population has its own individuality:

> *"The study of disease as a mass phenomenon differs from the study of disease in the individual primarily in respect to the unit of investigation. It is early appreciated that the herd, the crowd or the community is not a simple aggregate of the persons comprising that grouped population, but that each universe of people is an entity, a composite that possesses as much individuality as does a person."* (Gordon, 1950, p. 198).

The second principle, *group comparison*, consists in contrasting what is observed in the presence of exposure to what would have occurred had the group of interest not been exposed to the postulated cause. Differences in event occurrence between groups can logically be interpreted as being caused by the exposure. This is the main mode of knowledge acquisition in epidemiology. It relies on population thinking.

This essay addresses questions such as: how did epidemiologists integrate into their population thinking measures of disease occurrence of growing theoretical complexity and abstraction? How did simple ratios and proportions evolve into risks and rates, and later cumulative incidences and incidence densities? How did simple group comparisons eventually lead to a unified theory of study designs distinguishing cohort from case-control studies? Historical examples of the theoretical innovations and refinements illustrate the answers.

Even though there is a historical thread to its argument, this essay is more about the *epistemology* than about the history of epidemiology. Epistemology is a discipline that deals with the evolution of knowledge. This essay focuses more on how epidemiologic ideas evolved than on the description of the historical contexts in which these evolutions occurred or the identification of the exact moments at which they occurred, who had the original idea or published it first, etc. This approach is, I believe, analogous to annotated anthologies of articles and books (Pan American Health Organization, 1988; Greenland 1987a) or to the James Lind Library enterprise (http://www.jameslindlibrary.org).

Taken in isolation, population thinking and group comparisons can be found in other disciplines. Population thinking belongs to demography, statistics, and biology (Mayr, 1985). Group comparisons can be found in sociology or anthropology. But the blending of population thinking and group comparisons in an integrated theory to appraise health-related causal relations characterizes epidemiology. Indeed, the juncture of population thinking and group comparisons was the critical element that led to the birth of epidemiology in the 18th century. Over a period of less than 300 years, the theory of epidemiology has become quite rich. It comprises methods for group-comparisons (i.e., contrasts of exposed *vs.* unexposed to potential risk factors, and affected *vs.* unaffected by specific conditions) and two sets of concepts. One set rigorously expresses health-related phenomena occurring at the population level (e.g., prevalence, incidence, risks or rates). Another set of concepts is related to the design and interpretation of group-comparisons (e.g., confounding, interaction, bias, causal inference).

The material assembled in this essay demonstrates that epidemiology is a dynamic scientific discipline. Its methods and concepts have evolved across time, and will most likely continue to do so. This thesis would be refuted if it was shown that 1) the apparent evolution described below is a fallacy, that is, the whole corpus was present from the inception of epidemiology and has only been repeatedly re-invented; 2) epidemiology has now reached its definitive form and will not evolve beyond its current state of formalization.

2. Population thinking

2.1. Definitions

Predicting the experiences of a whole group of people distinguishes population thinking from other modes of reasoning. Under certain assumptions, population predictions can be made with a measurable degree of certainty. We can predict the number of new cases of disease in a population, but we cannot predict if a given individual will become sick. What will happen to an individual or the way an individual will behave in the future cannot be predicted with certainty.

Population thinking leads, however, to reliable predictions at the population level, which can then be applied to individuals. Suppose that 150 cases of breast cancer occur per 100,000 women and per year in a population of 200,000 women. We can predict with certainty that, if the rate remains constant, 300 new cases of the disease (plus or minus a certain number of cases reflecting the imprecision of the estimate) would be diagnosed each year. We cannot however precisely predict whether a specific woman, among the 200,000 women "at risk" for the disease, will develop breast cancer. At the individual level we can formulate "probabilities": each woman in this population has an annual risk of [300 ÷ 200,000] = 0.15%. This probability statement is based on what we observed for the group to which the woman belongs.

The relevance of population thinking to medical practice is not straightforward. Clinicians have opposed it in the past and still tend to avoid it. Medicine is the art of individual thinking. A skilled physician is one who is able to make the best prediction in terms of diagnosis and prognosis for the individual patient and adapt the management and treatment to the unique characteristics of an essentially unpredictable person. Because medicine is the art of individual thinking, we need physicians and cannot replace them by computers. But it has been a major and difficult conceptual leap for physicians to realize that something useful could be learned for the individual from populations.

Thus, there is a contradiction between population and individual thinking. For all of us, it takes a certain change in perspective to realize that populations don't behave as if they were simply the collection of unique and unpredictable individuals. Even though we don't understand exactly why that is so, populations have, to use Gordon's expression, their individuality. For example, heavy drinkers represent a larger fraction of some populations than others. When heavy drinking is common, the whole population tends to drink, on average, more alcohol compared to populations in which heavy drinking is less common. Heavy drinkers do not appear to be a well-defined, proportionally constant subgroup of people in every society. Their frequency varies and can even be predicted from the average alcohol intake of the population they belong to. This phenomenon was first reported by the French demographer Sully Ledermann in the 1950s (Ledermann, 1956) with respect to alcohol consumption. It has been popularized in epidemiology by Geoffrey Rose (1926–1993), Emeritus Pro-

fessor of Epidemiology at the London School of Public Health and Tropical Medicine, in a paper entitled: "The population mean predicts the number of deviant individuals" (Rose and Day, 1990).

From where do populations get their individuality? How does the community influence individual behaviors? There are probably no simple answers to these questions but it is clearly established that populations are more than collections of individuals. Some populations tolerate more obesity, excess alcohol intake, smoking, etc. in society. Some populations are physically more active than others. Some societies are more egalitarian than other. The crucial point is that the statistical laws that govern populations provide information that can be useful for the individuals belonging to these populations.

2.1.1. Ratios, risks, rates and odds

In order to think at the population level, we need to be able to describe the occurrence and evolution of events in populations. This requires appropriate measures. How frequent is the disease? How will it evolve in the future? At what speed will this evolution take place?

We will review how the intuitive adoption of population thinking eventually became a theory comprised of a set of well-defined concepts. This will take us from the 17th century to modern times. But before we get to these examples, we need to define some of the terms that are indispensable for exploring epidemiology's past. The number of concepts used in epidemiology is relatively limited, but the wealth of terms found in the epidemiologic literature can be confusing. An astounding effort of homogenization has been made by the *International Epidemiology Association* (Last, 2001), but we are still far from a consensual usage of a minimum terminology.

The words risk, rate, ratio, and odds are measures of event occurrence that differ by the nature of their numerator and denominator. I will use the definitions of these terms that Regina C. Elandt-Johnson, statistician from the Department of Biostatistics, University of North Carolina, gave in the very influential commentary she wrote in the October 1975 issue of the *American Journal of Epidemiology* (Elandt-Johnson, 1975).

Rates, risks, ratios and odds are measures (M) computed by dividing one quantity by another. The dividend is the numerator and the divisor is the denominator.

$$M = a \div b, \text{ where } M = \text{measure}, a = \text{numerator and } b = \text{denominator}$$

In a *ratio*, the numerator and denominator are two separate and distinct quantities, which are not included in one another. For example, dividing the number of deaths (numerator) by the number of births (denominator) is a ratio. The etymology of the word "ratio" is interesting. In Latin, it means "reason". Its original usage in mathe-

matics may be related to the fact that a ratio yielded a rational (an integer divided by another, e.g., 4 ÷ 2, as opposed to an irrational, e.g., square root of 2) number. But the term ratio now relates more to the principle of *comparing* two quantities.

A *risk* is a proportion, that is, a measure in which the denominator includes the numerator. For example, the risk of developing lung cancer is the proportion of a group of people at risk (denominator) who newly develop lung cancer (numerator) over a specified period of time (e.g., the risk of lung cancer can be 10% over 20 years in heavy smokers).

A *rate* is a measure of change in one quantity per unit of another quantity. In epidemiology, a rate is often used as synonym for incidence rate, which is the change in risk per unit of time. For example, a risk of 10% over 20 years can, if constant, be expressed as a rate of 0.5 per 100 and per year [rate = risk ÷ time = 10% ÷ 20 years = 0.5% per year].

To contrast the frequency of occurrence of an event to that of nonoccurrence we use the *odds*. The *odds of disease* are computed by dividing the risk by its complement: a risk of 10% over 20 years corresponds to the odds of 1 to 9 [odds = risk ÷ (100%-risk) = 10% ÷ 90% = 1 over 9]: the disease has 1 chance to occur *vs.* 9 not to occur. Similarly, the *odds of exposure* is obtained by dividing the percentage of exposed by the percentage of unexposed.

2.1.2. Prevalence, incidence, mortality, case fatality

Different concepts express whether a count (usually of people) is the result of past events or if it is a prediction for the future. In this essay I shall use the following terminology. The *prevalence* is the proportion of people in the total population suffering from a given disease (or exposed to a given factor) at a given point in time. The trait (disease, exposure, etc.) may be long existing or recent. Thus, prevalence measures a state of health resulting from events that occurred in the distant or recent past. The *incidence* is the proportion of *new* cases occurring in a population *at risk* of disease over a specified period of time (i.e., excluding prevalent cases or people not susceptible of contracting the disease). It is a synonym of risk. In contrast to prevalence it is a predictive statement about cases-to-be in a population still free of the disease. *Mortality* indicates the proportion of deaths in general, whereas *case fatality* is reserved for the deaths occurring among people who are diseased.

When incidence, mortality and case fatality are expressed per *unit of time*, they will be called incidence rate, mortality rate and case fatality rate.

2.2. Origin of population thinking

At the dawn of the 17th century, emerging modern European states became interested in collecting population data and using them to guide their policy. The wealth and

power of modern states depended on the education, health, income, political involvement and other characteristics of the population they governed. In England and France, the devastations of plague epidemics stimulated the process of population data collection, in which health indicators represented an important component. Hence, the etymology of the word "statistics": systematic data collection for the state. As the historian of public health, George Rosen (1910–1977) has put it,

> *"Initially, those who undertook to use the statistical approach concerned themselves chiefly with what might be called the bookkeeping of the state. Efforts were made to ascertain the basic quantitative data of national life in the belief that such knowledge could be used to increase the power and prestige of the state (...) The father of "political arithmetic" was William Petty (1623–1687), physician, economist and scientist, who invented the term and was keenly alive to the importance of a healthy population as a factor in national opulence and power. Repeatedly, Petty urged the collection of numerical data on population, education, diseases, revenue and many other related topics."* (Rosen, 1958, p. 111).

To the best of our current knowledge, the book of John Graunt (1620–1674) entitled "*Natural and Political Observations made upon the Bills of Mortality*" (Graunt, 1662) may be the first solid contribution to "public health statistics". According to a 17th century biographer (Aubrey, 2004), John Graunt was by profession a haberdasher, who eventually went bankrupt. He was also admitted as fellow of the Royal Society and pioneered the analysis of the Bills of Mortality (ancestors of the death certificates, systematically collected in England since 1603) to find uniform and predictable mass phenomena.

Kenneth J. Rothman, Professor of epidemiology at Boston University, has written a laudatory commentary on Graunt's contribution:

> *"With this book Graunt added more to human knowledge than most of us can reasonably aspire to in a full career. Graunt was the first to report, and to document, that more boys than girls are born. He presented one of the first life-tables. He reported the first time-trends for many diseases, taking into account changes in population size. He described new diseases, and noted others that seemed to increase over time only because of changes in classification. He offered the first reasoned estimate of the population of London, demonstrating its rapid growth and showing that most of the growth came from immigration. He proffered epidemiologic evidence refuting the theory that the plague spreads by contagion. (He also refuted the notion that plague epidemics are coincident with the reign of a new king.) He showed that the large population decreases in plague years were offset by large increases in births in subsequent years. He showed that physicians have twice as many female as male patients, but that more males than females die. He produced the first hard evidence about the frequencies of various causes of death.*

And, presaging our present-day paranoia, he tried to allay unwarranted anxiety about risks that were feared far out of proportion to their likelihood of occurrence." (Rothman, 1996, p. 37).

It should be kept in mind that the history of English "statistics" has apparently been more studied than that of other countries, even in Europe. Thus, it is much less known that the Swiss physician Felix Platter (1536–1614) had shown, before Graunt, that the plague appeared to regulate the population size of the City of Basel in the northern part of Switzerland (Mattmueller, 2004).

The growing interest in population data, probabilities and population thinking reached medicine too. Compiling the mass of data generated by the activity of hospitals and infirmaries could be used to improve medical activity (Troehler, 2000, p. 15). We find in 18th century England early attempts to evaluate the *average* effect of specific therapies in groups of patients. In the 19th century, some physicians clearly expressed the need for aggregated data:

"... that it is impossible to appreciate each case with mathematical exactness, and it is precisely on this account that enumeration becomes necessary." (Louis, 1836, p. 60).

And for population thinking:

"To ascertain the cause of cholera, we must consider it not only in individual cases but also in its more general character as an epidemic." (Snow, 1849, p. 746).

Population thinking in the domain of health first appears in the 18th century and is unambiguously expressed by scientists of the 19th century. Let us review now the evolution of the measures and concepts which have contributed to population thinking in epidemiology.

2.3. Early ratios, proportions and rates

The first measures used to express the occurrence of disease in populations were ratios, proportions and probably primitive mortality rates.

2.3.1. Eighteenth century

Plaque, a lethal disease caused by *Tersinia Pestis* and propagated by fleas and rats, has constituted a significant demographic factor in late medieval and early modern times in all parts of Europe (McNeil, 1976, p. 151). The data in Table 1 are from chapter IV ("Of the plague") of John Graunt's "*Natural and Political Observations Made upon the Bills of Mortality*"(Graunt, 1662). The table shows the overall num-

Table 1 – Proportions and ratios in the work of John Graunt. The data are extracted from chapter IV ("Of the plague") of John Graunt's "Natural and Political Observations upon the Bills of Mortality" (Graunt, 1662, pp. 33–36).

Year	Deaths ("Died"or "buried")	"Whereof plague"	Other causes	Plague mortality "proportion"	Births ("christened")	Death to birth ratios
1592	25,886	11,503	14,383	2 to 5	4,277	6 to 1
1603	37,294	30,561	6,733	4 to 5	4,784	8 to 1
1625	54,265*	35,417	18,848	7 to 10	6,983	8 to 1
1636	23,359	10,400	10,400	2 to 5	9,522	5 to 2

* The table in Graunt's book says 51,758, which is probably a typographical error. The "Table of burials and christenings", appended in page 75 of the *Observations,* indicates a total of 54,265 deaths. Graunt uses sometimes 54,265, and sometimes 51,758, in his calculations.

bers of deaths, deaths due to the plague and the number of "christened", that is, births, in London for the years 1592, 1603, 1625 and 1636. The table also reports the "proportions" of all deaths due to plague, and the ratios of "buried to christened", that is, deaths to births. Graunt assembled the numbers and made the calculations in Table 1 to address the following question:

> *"In which of [these years] was the greatest Mortality of all Diseases in general, or of the Plague in particular?"* (Graunt, 1662, p. 33).

a) Proportions
Graunt uses proportions (column 5 of Table 1) to show that the greatest mortality from plague occurred in 1603, as 80% (4 to 5) died of plague, which is greater than the 70% (7 to 10) which occurred in 1625.

> *"For if the Year 1625 had been as great a Plague-Year as 1603 there must have died not only 7 to 10 but 8 to 10 which in those great numbers makes a vast difference (...) We must therefore conclude the Year 1603 to have been the greatest Plague-Year of this age."* (Graunt, 1662, p. 34).

b) Ratios
Graunt notes some inconsistency in the Bills. The year of greatest mortality from the plague (1603) is different from the year of greatest overall mortality (1625). For that purpose, Graunt computes the ratio of the number of deaths (i.e., burials) over the number of births (i.e., christenings) (last column of Table 1). This ratio is 8 to 1 both

in 1603 and 1625. There was apparently no "errour in the Accompts" for the overall mortality in 1625. However, compared to the years before (1622) or after (1626) the plague, there was in 1625 an excess of 11,000 deaths from causes other than the plague. This excess could be explained by misclassification of plague deaths into deaths from other causes. Graunt thus added 11,000 to the 35,417 plague deaths of 1625, making a total of 46,417, which is about "four to five" of the whole 54,265, almost the same as 1603

> *"... thereby rendering the said year 1625 to be as great a Plague-year as that of 1603 and no greater, which answers to what we proved before, viz. that the Mortality of the two Years was equal."* (Graunt, 1662, p. 35).

c) Rates

Graunt observes that the mortality from plague varies from one epidemic to another and makes "sudden jumps" within the evolution of the same epidemic. In order to describe this mortality variation, Graunt uses a primitive form of mortality rates. The time unit is *year* to compare one epidemic with the other

> *"The Plague of 1636 lasted twelve Years, in eight whereof there died 2000 per annum one with another, and never under 300."* (Graunt, 1662, p. 36).

The sudden jumps of deaths occurring within the same epidemic are given per *week*:

> *"... the sudden jumps, which the Plague hath made, leaping in one Week from 118 to 927: and back again from 993 to 258: and from thence again the very next Week to 852."* (Graunt, 1662, p. 36).

Of course, deaths are not divided by the number of people at risk, and these rates may not have been accurate if the population of London varied substantially during plague years, when the wealthiest fled out of the city. But these deaths per year or per week play the role of mortality rates. On their basis, Graunt can go beyond the mere description of the overall burden of deaths due to each plague epidemic. Deaths per year or per week decompose the overall mortality from plague into small units of time allowing Graunt to describe the variation in intensity of the epidemic. Indeed, Graunt concluded that such sudden changes in mortality had to be determined by some external causes, related to the environment, and could not be due to causes internal to the human constitution:

> *"The which effects must surely be rather attributed to change of the Air, then of the Constitution of Mens bodies, otherwise then as this depends upon that."* (Graunt, 1662, p. 36).

Table 2 – Mortality after 7 weeks. London, 1854: "The following is the proportion of deaths to 10,000 houses, during the first seven weeks of the [1854] epidemic, in the population supplied by the Southwark and Vauxhall Company, in that supplied by the Lambeth Company, and in the rest of London." Source: Table IX, in (Snow, 1855, p. 53).

	Number of houses	Death from cholera	Deaths in each 10,000 houses
Southwark and Vauxhall Company	40,046	1,263	315
Lambeth Company	26,107	98	37
Rest of London	256,423*	1,422	59**

* There seems to be some inconsistency between the table and the text relative to the number of households in the rest of London. "*The number of houses in London at the time of the last census was 327,391. If the houses supplied with water by the Southwark and Vauxhall Company, and the deaths from cholera occurring in these houses, be deducted, we shall have in the remainder of London 287,345 houses ...*" (Snow, 1855, p. 50). Thus, [327,391–40,046 –26,107 =] 261,238, which is different from the 287,345 given elsewhere in the text and from the 256,423 in the table.

** [1,422 ÷ 256,423 =] 55 per 10,000, not 59 as reported by Snow.

2.3.2. Nineteenth century

Cholera had long been endemic in Bengal, India. It was a frightening disease, which killed its victims sometimes within hours, by radical dehydration from diarrhea, vomiting and fever. Ruptured capillaries made the skin turn black and blue, hence the popular name of the disease: the blue death or, in French, *la mort bleue*. In the early 19th century, cholera made recurrent world excursions, which brought it several times to London.

John Snow (1813–1858) was an English anesthesiologist, convinced that cholera was a contagious disease. He had been studying the recurring outbreaks of cholera in England and published in 1849 the hypothesis that polluted water was one of the means of cholera transmission (Vinten-Johansen et al., 2003; Shephard, 1995). When cholera returned to London in July 1854, John Snow used the opportunity to test his hypothesis. I will describe the study itself in detail later (section 3.3.2), but focus here on the measures of disease occurrence used by Snow.

a) Ratios

Table 2 reproduces the most famous results of John Snow's investigation on the mode of transmission of cholera from the 1855 edition of his book "*On the Mode of Communication of Cholera*" (Snow, 1855). They were collected during the first seven weeks of the epidemic of cholera that hit London in July 1854. Snow uses a ratio to

Table 3 – Mortality after 14 weeks. London, 1854: "By adding the number of deaths, which occurred in the first seven weeks of the epidemic, we get the numbers in the subjoined table (No. XI), where the population of the houses supplied by the two water companies is that estimated by the Registrar General." Source: Table XI, in (Snow, 1855, p. 55).

	Population in 1851	**Death by cholera in 14 weeks end Oct 14 [1854]**	**Deaths in 10,000 livings**
Southwark and Vauxhall Company	266,516	4,093	153
Lambeth Company	173,748	461	26
London	2,362,236	10,367	43

quantify the impact of the epidemics on, respectively, the clients of two water supply companies, the Southwark and Vauxhall Company, and the Lambeth Company, and the rest of London. The numerators are the numbers of deaths observed in each of the three groups. The denominators are the numbers of households supplied by water companies. Snow refers to this ratio by saying, inappropriately, that it is "the proportion of deaths to 10,000 houses" (Snow, 1855, p. 86).

However, to interpret the ratio of deaths to households as a proportion or a risk, Snow would have had to assume that the average size of the households was similar across London. Let us imagine that the Southwark and Vauxhall Company supplied poor and crowded house blocks, in which the average household was 8.4 times larger than in the more well off house blocks supplied by the Lambeth company. In that situation, the actual mortality risk from cholera would be identical for the two companies, as [1263 ÷ (40,046 × 8.4)] is equivalent to [98 ÷ 26,107]. Indeed, According to John Eyler (Eyler, Part Ia), the fact that Snow did not know the number of clients at risk of cholera fed the initial skepticism towards his conclusions.

b) Proportions

In Table XI of "*On the Mode of Communication of Cholera*" (see Table 3 above), Snow does present real proportions. The numerator is the number of deaths while the denominator is the number of people living in the houses supplied by the companies. This denominator had its own limitations as it was based on an already three-year old census. The office of the Registrar General computed these proportions.

Note that apparently 6 to 7 people lived in each household supplied by both companies. The deaths had tripled for the Southwark and Vauxhall, almost quintupled for the Lambeth company in seven weeks. Still, mortality over 14 weeks (Table 3)

Table 4 – Duration, mortality (i.e., risk) and force of mortality (i.e., rate) for cholera and phthisis. Source: (Farr, Part II).

Disease	Mean duration (in days)	Mortality (% of all the sick)	Force of mortality (= Mortality rate per 100 sick a year)
Cholera	7	46	2415
Phthisis	730	90–100	50

was almost half of that over 7 weeks based on households (Table 2) and differences were less important.

2.4. Risks and rates

It has taken about 150 years to sort out the properties of risks and rates, clarify their interpretation and produce a theory of their mathematical relationships. We will review here three episodes of this process.

2.4.1. Burden of life destruction and force of mortality

As Superintendent of the General Register Office, England's center for vital statistics, William Farr (1807–1883) was responsible for collecting and reporting information on causes of death (Susser and Adelstein, 1975). In the pamphlet entitled "*On Prognosis*", reproduced *in extenso* in this book (Farr, Part II), Farr illustrates the need for different types of measure of disease occurrence by contrasting an acute infectious disease, cholera, with a chronic infectious disease, phthisis (i.e., tuberculosis). He invokes the following paradox:

> "*Cholera destroys in a week more than phthisis consumes in a year. Phthisis is more dangerous than cholera; but cholera, probably, excites the greatest terror.*" (Farr, Part II).

Table 4 shows that almost every tuberculosis patient will die from the disease. The case fatality risk of phthisis is 90–100%. Cholera kills only one of two persons who are affected: its case fatality risk is 46.2%.

Half of the people who get cholera but almost none of those with phthisis will survive. Between cholera and phthisis, it would seem reasonable to prefer cholera, but people fear cholera more than tuberculosis. Why is it so? Farr notes that mortality is

insufficient to characterize the "form and nature of diseases". We need two different measures of disease occurrence:

> *"Diseases may be examined (1) in their tendency to destroy life, expressed by the deaths out of a given number of cases; and (2) in their mean relative 'force of mortality', expressed by the deaths out of a given number sick at a given time."* (Farr, Part II).

Let us consider each of these two ways of examining a disease. For the first parameter, "the tendency to destroy life", Farr gives as synonyms the "probability of death", "mortality" and "death percent". If 990 patients died out of 2,142 cases of cholera, "mortality" is 46.2%. Farr does not use the word "risk", but risk is the term that we would commonly use today. More specifically, this is a "case fatality risk". It expresses the probability that patients with cholera will *die* from their disease. Deaths are in the numerator and sick people are in the denominator.

The second parameter, "force of mortality", is the "quantity eliminated daily by death out of a given constant quantity (e.g., 100) sick". Farr also refers to it as the "mean *rate* of dying per unit of sick time". To compute the force of mortality, Farr divides the number of deaths by the product of the number of persons sick and the average duration during which they were sick. If 2,142 cases of cholera have been sick an average of 7 days each, this corresponds to a total of [7 × 2,142 =] 14,994 days of sickness, or sick person-days. Sick-person days divided by 365 days in a year gives 41 years of sickness or 41 sick person-years. Thus, if 990 die out of 41 sick person-years of cholera, the "force of mortality" is [(990 ÷ 41) × 100) =] 2,415 per 100 sick person-years. The modern synonym of "force of mortality" is mortality rate, and in this example specifically, it is a "case fatality rate". It is the proportion of the cases that will die from their disease *per unit of time*: 2,415 per 100 patients per *year* or 6.6 per 100 patients per *day*.

Distinguishing these two measures of death occurrence allows Farr to explain the paradoxical terror generated by cholera. The data are shown in Table 4. Almost all patients died from tuberculosis (mortality risk = 90–100%), but the death rate is small (50 per 100 per year) and the average duration of the disease is long (2 years). Tuberculosis kills slowly. On the other hand, less than half of the sick will die from cholera (mortality risk = 46%), but the death rate is huge (2,415 per 100 per year) and the average duration of the disease is short (7 days). Cholera appears abruptly, kills rapidly and disappears. Viewed as such, cholera is more frightful.

Why did Farr use the word "force" to characterize a rate? We can speculate that this is in relation to the concept of physical force. Farr must have been familiar with the concept of force defined by the physicist Isaac Newton (1643–1727) in his "*Principia*" (Newton, 1687):

> *"An impressed force is an action exerted upon a body, in order to change its state, either of rest, or of moving uniformly forward in a right line. This force con-*

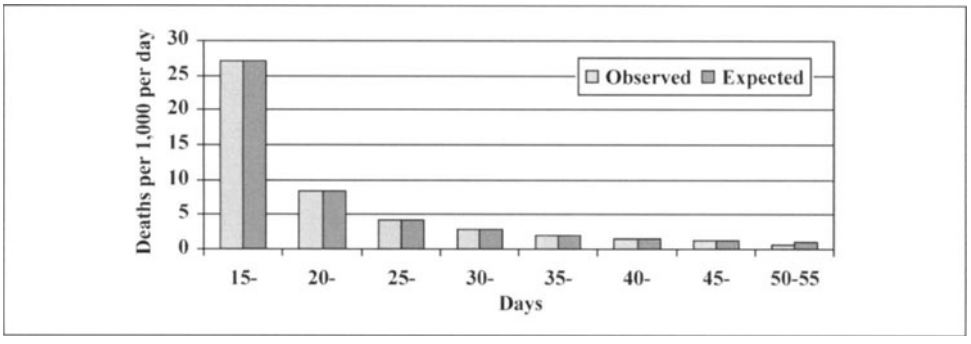


Figure 1
Evolution of the observed and expected death rates from smallpox. Source: William Farr, On Prognosis (Farr, Part II).

> *sists in the action only; and remains no longer in the body when the action is over."* (Cited by Einstein and Imfeld, 1966, p. 11).

A force can be represented by a vector, which has a direction and a velocity. The velocity, that is, the distance covered per unit of time, is by definition a rate. The force of mortality, like a vector, has a velocity and a direction. Mortality rates can go up or down.

Farr notes that predicting the direction in which risk will evolve is crucial for prognosis. The sign of the force indicates whether the rate increases or decreases over time. Indeed, Farr gives the data needed to compute the force of mortality on the 18th, 19th day, etc. of duration of smallpox (Gerstman, Part II). Using the word "*rate*", Farr notes that:

> *"The rate of mortality [from smallpox] increased from the 5–10 days to 10–15 when it attained a maximum (31.18); it decreased in a determined progression from the next period (15–20 days) to the end."* (Farr, Part II).

Farr was mostly interested in the declining part of the rate curve (see Figure 1), which demonstrated some mathematical regularity:

> *"The decrease begins to take place in geometrical progression; but the tendency to decrease is met by another force that neutralizes part of its effect."*
> (Farr, Part II).

Again, the use of the force of mortality had a very important clinical implication. In the case of cholera, early treatment was essential because half of the deaths happened in the first 24 hours:

"What the practitioner does he should do quickly."
(Farr, Part II).

2.4.2. The fallacy resulting from neglect of the period of exposure to risk

We speak of a 5-year-risk or a 10-year risk. Whether the risk is over 5 or 10 years is critical for its interpretation. Neglecting the period of exposure to risk can also lead to invalid interpretation of a study result. The British epidemiologist Austin Bradford Hill (1897–1991) described the potential fallacy resulting from neglect of the period of exposure to risk in his textbook "*Introduction to medical statistics*" (Hill, 1939). As it is difficult to write more clearly than Hill, I will quote him here extensively.

"Suppose on January 1st 1936 there are 5,000 persons under observation, none of whom are inoculated; that 300 are inoculated on April 1st, a further 600 on July 1st, and another 100 on October 1st. At the end of the year there are, therefore, 1,000 inoculated persons and 4,000 still uninoculated. During the year there were registered 110 attacks amongst the inoculated persons and 890 amongst the uninoculated. If the ratio of recorded attacks to the population at the end of the year is taken, then we have rates of 110 ÷ 1,000 = 11.0 per cent amongst the inoculated and 890 ÷ 4,000 = 22.3 per cent amongst the uninoculated, a result apparently very favorable to inoculation. This result, however, must be reached even if inoculation is completely valueless, for no account has been taken of the unequal lengths of time over which the two groups were exposed. None of the 1,000 persons in the inoculated group were exposed to risk for the whole of the year but only for some fraction of it; for a proportion of the year they belong to the uninoculated group and must be counted in that group for an appropriate length of time.

The calculation should be as follows:

All 5,000 persons were uninoculated during the first quarter of the year and therefore contribute (5,000 × 1/4) years of exposure to that group. During the second quarter 4,700 persons belonged to this group – i.e., 5,000 less the 300 who were inoculated on April 1st – and they contribute (4,700 × 1/4) years of exposure to the uninoculated group. During the third quarter 4,100 persons belonged to this group – i.e.. 4,700 less the 600 who were inoculated on July 1st – and they contribute (4,100 × 1/4) years of exposure. Finally in the last quarter of the year there were 4,000 uninoculated persons – i.e., 4,100 less the 100 on October 1st – and they contribute (4,000 × 1/4) years of exposure. The "person-years" of exposure in the uninoculated group were therefore (5,000 × 1/4) + (4,700 × 1/4) + (4,100 × 1/4) + (4,000 × 1/4) = 4,450, and the attack-rate was 890 ÷ 4,450 = 20 per cent. – i.e., the equivalent of 20 attacks per 100 persons per annum. Similarly the person-years of exposure in the inoculated group are (0 × 1/4) + (300 × 1/4) + (900 × 1/4) +

Table 5 – Hypothetical example illustrating the fallacy resulting from neglect of the period of exposure to risk. Source: Table XVII, in (Hill, 1939, p. 130).

Inoculated at each point of time	Inoculated		Uninoculated	
	Exposed to risk in each quarter of the year [A]	Attacks at 5 per cent per quarter [B = A × 0.05]	Exposed to risk in each quarter of the year [C]	Attacks at 5 per cent per quarter [D = C × 0.05]
Jan. 1st, 0	0	0	5,000	250
April 1st, 300	300	15	4,700	235
July 1st, 600	900	45	4,100	205
Oct. 1st, 100	1,000	50	4,000	200
Total at end of the year	1,000	110	4,000	890

$(1{,}000 \times 1/4) = 550$, for there were no persons in this group during the first three months of the year, 300 persons during the second quarter of the year, 900 during the third quarter, and 1,000 during the last quarter. The attack-rate was, therefore, $110 \div 550 = 20$ per cent, and the inoculated and uninoculated have identical attack-rates. Neglect of the durations of exposure to risk must lead to fallacious results and must favor the inoculated. The figures are given in tabulated form (Table XVII).

Fallacious Comparison – Ratio of attacks to final population of group. Inoculated $110 \div 1{,}000 = 11.0$ per cent. Uninoculated $890 \div 4{,}000 = 22.3$ per cent.

True Comparison – Ratio of attacks to person-years of exposure. Inoculated $110 \div (300 \times 1/4) + (900 \times 1/4) + (1{,}000 \times 1/4) = 20$ per cent. Uninoculated $890 \div (5{,}000 \times 1/4) + (4{,}700 \times 1/4) + (4{,}100 \times 1/4) + (4{,}000 \times 1/4) = 20$ per cent." (Hill, 1939 pp. 128–130).

Using the terminology adopted in this book, the risks (number of cases divided by persons at risk) were 11% in the inoculated and 22.3% in the uninoculated. Apparently, inoculation protected. But the period during which cases were ascertained was shorter for the inoculated than it was for those uninoculated, because the inoculation had been done progressively between April and October of the year of observation. Using person-years at the denominator corrected this imbalance and revealed that the rate was 20 per hundred per year, identical in both groups. The valid conclusion was that inoculation is useless.

The important concept was that a risk was always implicitly associated with a period over which it applied. A risk of 20% has a different meaning if it is expressed

over 6 months, one year or ten years. There is no doubt that this was understood before Hill. But Hill's example shows how critical this characteristic of risk can be, especially for group comparisons.

2.4.3. *Incidence density and cumulative incidence*

Olli S. Miettinen, from the Department of Epidemiology and Biostatistics at Harvard School of Public Health, revisited the relation of risk to rate 138 years after Farr in another seminal paper in the history of epidemiologic methods and concepts entitled "*Estimability and estimation in case-referent studies*" (Miettinen, 1976a). The paper addressed a problem very different from Farr's preoccupation with respect to prognosis: it had to do with the relation of case-control (which Miettinen termed case-referent) and cohort studies (see section 3.11).

Miettinen renamed the incidence rate "incidence density", and interestingly, listed as synonyms two of Farr's expressions, "force of morbidity" and "force of mortality". Miettinen also popularized the term "cumulative incidence" instead of "risk". The properties of risks and rates remained those described by Farr, but Miettinen showed that the risk could be expressed as a function of the incidence density (ID). In its simpler formulation:

$$\textit{Cumulative incidence}_{(\textit{up to time } j)} = \Sigma_{\textit{from time } i = 1 \textit{ to } j}\, ID_i$$

For example, suppose that the incidence rate of a relatively rare disease (e.g., breast cancer) changes at each year of age and that there is no cohort effect (see section 3.4.3). The risk of a woman to develop breast cancer before age 75 is the sum of the 74 age-specific incidence rates between birth and age 74. In Western societies, this cumulative incidence is about 7%. The formula found in Miettinen's paper (Miettinen, 1976a) allows for the possibility that incidence rates are stable over specific time periods, Δt (e.g., $\Delta t = 5$ for a 5-year risk). In this situation:

$$\textit{Cumulative incidence}_{(\textit{up to time } j)} = \Sigma_{\textit{from time } i = 1 \textit{ to } j}\, ID_i \times \Delta t_i$$

Miettinen's innovative concepts have reached a much larger audience than the papers in which he developed them. The original papers can be arduous for someone who is not already familiar with epidemiologic concepts and methods and does not have some mathematical background. Therefore, his concepts have usually been disseminated through the work of people who wrote didactic translations of his ideas. We owe to a group of epidemiologists and statisticians at the School of Public Health of the University of North Carolina and Yale University, Hal Morgenstern, David G. Kleinbaum and Lawrence L. Kupper a paper that translates Miettinen's 1976 "*Estimability*" paper into a more universally accessible prose (Morgenstern et al., 1980).

The paper reminded first that:

> *"(...) the concept of risk requires a specific period referent, – e.g., the 5-year risk of developing lung cancer."* (Morgenstern et al., 1980, p. 97).

When computing the risk, that is, the proportion of all the subjects at the onset who developed the disease during a given period, we assume that all subjects have been followed during the full period. What happens when this condition of complete follow-up is not met? William Farr and Bradford Hill had shown that we could avoid a bias by computing incidence rates based on person-times, instead of risks. Miettinen proposed the following solution: divide the duration of follow-up, t, into short time intervals; compute a risk for each short interval and call it incidence density (ID); sum the incidence densities over all time intervals and you get the cumulative incidence (CI) over the period t. The cumulative incidence is a measure of the risk over period t. Using Miettinen's formula given above, we can compute the cumulative incidence (= risk) as the sum of incidence densities. This measure of risk is not affected by the fact that some observations had incomplete follow-up.

Morgenstern, Kleinbaum and Kupper illustrated the relation of risk (CI) and rate (ID) by the example described in Table 6.

The question is: what is the risk of a 35-year old woman to develop breast cancer before age 55? If we take the 60,000 women in age group 35–39 followed 3 years,

Table 6 – Illustration of the estimation of risk in a dynamic population of 250,000 women free of breast cancer, aged 35 to 55y, followed up for 3 years (on average). Source: Table 1, in (Morgenstern et al., 1980).

Age (yr)	Women at risk [N]	No of incident cases [I]	Person-years [PY = N × 3]	Incidence density[1] (/100,000/yr)	5-year Risk[2] (/100)
35–39	60,000	90	180,000	50	0.250
40–44	70,000	168	210,000	80	0.399
45–49	65,000	215	195,000	110	0.550
50–54	55,000	227	165,000	138	0.686
					20-year Risk[3]
35–54	250,000	700	750,000	–	1.871

[1] Incidence density = I ÷ Person-years.

[2] Estimate of the Δt = 5-year risk for a woman at the beginning of each age category, $R_{\Delta t} \cong 1 - \exp[-ID \times \Delta t]$.

[3] Estimate of the 20-year risk for a 35 year-old woman, $R_{\Delta t} \cong 1 - \exp[-\Sigma_j ID_j \times \Delta t_j] \cong 1 - \Pi_j (1 - R_{\Delta tj})$.

they represent altogether 180,000 person-years (column 4). The incidence density in this age category is therefore [90 ÷ 180,000 =] 50 per 100,000 per year. Now, the risk of developing breast cancer for a women aged 35 before she reaches 40, that is, over a period of 5 years, is obtained, grossly, by multiplying the incidence density by 5 years, that is, 250/100,000 or 0.25% over 5 years (last column). These 5-year risks increase with age. Thus, the 20-year risk for that same woman aged 35 corresponds, grossly, to the sum of the 5-year risks across the four age categories: [0.0025 + 0.00399 + 0.0055 + 0.00686 =] 1.885%, which is close to the 1.871 per 100 obtained using the appropriate formula mentioned in the Table 6. The answer to the question is: the 20-year risk is about 1.9%.

Note that the formula used to compute the cumulative incidences is more complicated than the simple sum of incidence densities, and should be preferred if the disease is not rare. This example underlines the conceptual evolution between Farr and Miettinen, but does not fully reflect the richness of the theory developed underneath.

2.5. Prevalence and incidence

We have seen that *prevalence* measures the accumulation in the population of events (exposures or diseases) that occurred in the distant or recent past, while *incidence* is a predictive statement about cases-to-be in a population still free of the disease. The two concepts are closely related and their relationships have been explored at least under two different perspectives: a) the relation of incidence to prevalence of disease; b) the relation of (excess) incidence to prevalence of exposure.

2.5.1. Disease prevalence divided by incidence

It has been suggested that Farr had made the first description of the relation between prevalence and incidence, as follows:

> *"... in estimating the prevalence of diseases, two things must be distinctly considered; the relative frequency of their attacks, and the relative proportion of sick-time they produce. The first may be determined at once, by a comparison of the number of attacks with the numbers living; the second by enumerating several times the living and the actually sick of each disease, and thence deducing the mean proportion suffering constantly. Time is here taken into account: and the sick-time, if the attacks of two diseases be equal, will vary as their duration varies, and whatever the number of attacks may be, multiplying them by the mean duration of each disease will give the sick- time."* (Cited by Lilienfeld, 1978, p. 515).

Table 7 – Prevalence, incidence and duration of acute and chronic leukemia. Brooklyn, New York, 1948–1952. Source: Table 6, in (MacMahon et al., 1960, p. 60).

Abbreviations		Acute leukemia	Chronic leukemia
[P]	Prevalence (per million)	6.7	56.1
[I]	Incidence (per million per year)	32.4	29.0
[PI]	Duration (in years)	0.21	1.93

But this citation seems to only reiterate the distinction between death risk and death rate. Farr says that the number of deaths divided by the number of living cases gives the risk, and divided by sick person-times gives a rate. Farr uses the word prevalence as a synonym for disease occurrence. The key sentence, "time is here taken into account", is related to the computation of person-times.

The first time I found the relation of prevalence to incidence clearly described was in the textbook of epidemiology "*Epidemiology: Principles and Methods*" by Brian MacMahon, Thomas F. Pugh (1914–1973) and Johannes Ibsen (no dates found) (MacMahon et al., 1960, pp. 60–61) from the Department of Epidemiology at Harvard School of Public Health. The relation of prevalence to incidence is quite straightforward:

> "*... a change in point prevalence from one period to the next may be the result of changes in (1) incidence, (2) duration, or (3) both incidence and duration.*" (MacMahon et al., 1960, p. 61).

Prevalence may increase because patients survive longer with their disease. At a given moment, if incidence and duration can be deemed constant, their relation to prevalence seems to come out straight from a textbook of mechanical physics:

$$\textit{Prevalence} = \textit{Incidence} \times \textit{Average duration of disease}$$

Both incidence and duration need to be expressed in the same time units (e.g., years). Table 7 shows that both acute and chronic leukemia have similar incidence rates (about 30 per million per year), but that chronic leukemia is eight times more prevalent than acute leukemia.

Using the formula above, we can compute the average duration of the disease (D) by dividing the prevalence (P) by the incidence (I):

$$\textit{For acute leukemia: } D = P \div I = [6.7 \div 32.4] = 0.21 \textit{ years or } 2.5 \textit{ months}$$
$$\textit{For chronic leukemia: } D = P \div I = [56.1 \div 29.0] = 1.93 \textit{ years or } 23 \textit{ months}$$

These durations were close to the values of 2.4 months for acute leukemia and 20 months for chronic leukemia derived from independent follow-up of these same patients.

The conceptual link between prevalence, incidence rate and duration is perfectly illustrated in this example. It was to be shown later that the full theory was a bit more complicated. The $P = I \times D$ relation can only be assessed in populations that are stable in terms of risk and balanced in terms of in- and out-migration (Freeman and Hutchison, 1980; Miettinen, 1985). The exact relation is with the prevalence odds $[P \div (1\text{-}P)]$ rather than with the simple prevalence (Miettinen, 1985; Rothman, 1986).

2.5.2. Exposure prevalence multiplied by (excess) incidence

Geoffrey Rose introduced a new dimension of population thinking when he computed and interpreted the product of prevalence and (excess) incidence. In the previous examples, population thinking consisted in applying to an individual, information gathered in the population such as the risk of dying from cholera. If, on average, 46% of the cholera patients die from the disease in the population, we would say that any individual in this population had a 46% risk of dying when infected by *Vibrio cholerae*. In his seminal paper entitled "*Strategy of prevention: lessons from cardiovascular disease*" (Rose, 1981), Rose approached the question of the risk impact at the *population* rather than at the *individual* level:

> *"What we may call "population attributable risk" – the excess risk associated with a factor in the population as a whole – depends on the product of the individual attributable risk (the excess risk in individuals with that factor) and the prevalence of the factor in the population." (Rose, 1981, p. 1849).*

Rose demonstrated that, for diseases such as coronary heart disease or stroke, the majority of the cases occur among subjects at low risk of disease. Why is this so? Because low-risk constitutions for chronic diseases are usually much more common than high-risk constitutions. The histogram in Figure 2 (corresponding to Figure 3 of Rose's paper) shows the prevalence of various categories of serum cholesterol levels in 246 men aged 55–64 at the baseline examination of the Framingham Heart Study.

The way the numbers were obtained is shown in Table 8.

If we set, as Rose did, the point for hypercholesterolemia at 310 mg/dl (or 8 mmol/l), 3% of the population is hypercholesterolemic. If the cutoff is set at 250 mg/dl (6.5 mmol/l) as recommended today, about 25% of the population is hypercholesterolemic. The figure and the table show the mortality rates from coronary heart disease corresponding to each of the categories of serum cholesterol concentration. For example, the mortality rate is 11.19 per 1,000 per year in those with serum cholesterol of 310 mg/dl (8 mmol/l) or more. The excess mortality rate attributable to high cholesterol is obtained by subtracting the absolute mortality rate in

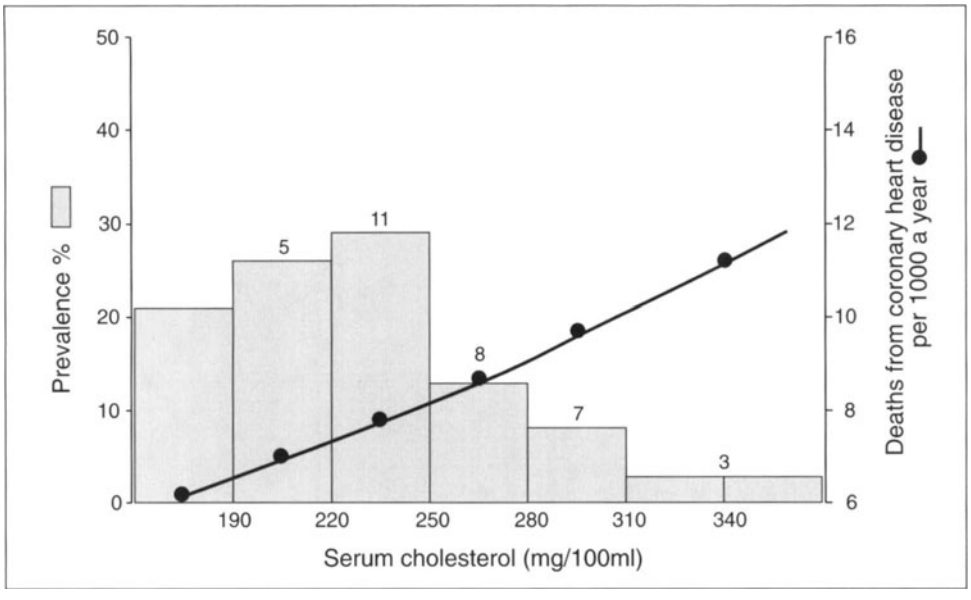


Figure 2
Prevalence of cholesterol levels, and corresponding mortality rates and excess cases. The Framingham Heart Study. Exam 1. Men aged 55–64 years. Source: Figure 3 in (Rose, 1981, p. 1849).

the subgroup with hypercholesterolemia from that with serum cholesterol below 190 mg/dl (4.92 mmol/l), that is, [11.19 – 6.22 =] 4.97/1,000 per year. The attributable risk over 10 years is therefore ten times larger, that is, 49.7/1,000. The 10y-attributable risk applied to 7 men (3% of 246 men) yields [7 × 0.0497 =] 3 deaths per 10,000 men over 10 years. If we do a similar calculation for the subgroup with total cholesterolemia between 220 and 250, we get 11 excess deaths for 10,000 at risk over 10 years. Altogether, 31 of the 34 (91%) excess deaths in 10 years will occur among people with total cholesterol <8 mmol/l, or 16 out of 34 (47%) among people with total cholesterol <6.5 mmol/l. Most cases occur in people without hypercholesterolemia.

Rose's description of Figure 2 reads like this:

> *"The risk rises fairly steeply with increasing cholesterol concentration; but out on the right, where the risk to affected individuals is high, the prevalence is fortunately low. If we want to ask, 'How many excess coronary deaths is the cholesterol-related risk responsible for in this population?' we simply multiply the excess risk at each concentration by the number of people with that concentration that are exposed to that risk. In figure 3* [Figure 2 above] *these attributable deaths*

Table 8 – Prevalence of cholesterol levels, and corresponding mortality rates, excess risk, excess deaths and cumulative proportions of excess deaths. The Framingham Heart Study. Exam 1. Men aged 55-64 years. Source: Tables 13-3-A and B, in (Kannel and Gordon, 1970).

Cholesterol	N [A]	Prevalence (%) [B = A ÷ 246]	Mortality rate (/1,000/ year) [C]	Excess mortality risk (/1,000) over 10y [D = (C– 6.22) × 10]	Excess deaths/ 10,000 over 10 y* [E = (D × A) ÷ 100]	Cumulative proportion of excess deaths [F = Σ (E ÷ 34)]
Less than 190	52	21	6.22	0	0	0
190 to 219	63	26	7.00	7.80	5	14.71
220 to 249	71	29	7.80	15.80	11	47.06
250 to 279	33	13	8.68	24.60	8	70.59
280 to 309	20	8	9.67	34.50	7	91.18
310 or more	7	3	11.19	49.70	3	100
	246	100			34	

* Rose wrote "extra deaths per *thousand* of this population over a 10-year period" but calculations based on the data he used indicate extra deaths per *ten thousand* over 10 years.

> *are shown as the numbers on top of bars. They add up to 34 extra deaths per 1,000 in this population over a 10-year period, of which only three arise at concentrations at or above 310 mg/100 ml (8 mmol/l) – which would be called high ("outside the normal range") by conventional clinical standards. The rest (90%) arise from the many people in the middle part of the distribution who are exposed to a small risk.*" (Rose, 1981, p. 1849).

Rose concluded that

> *"this illustrates a fundamental principle in the strategy of prevention. A large number of people exposed to a low risk are likely to produce more cases than a small number of people exposed to a high risk."* (Rose, 1981, p. 1849).

In this same seminal paper, Rose made another key observation, which he called the "prevention paradox":

> *"A measure that brings large benefit to the community offers little to each participating individual."* (Rose, 1981, p. 1850).

This paradox leads to another way of expressing the population attributable risk. Rose noted for example that

> *"when mass diphtheria immunization was introduced in Britain 40 years ago, even then roughly 600 children had to be immunized in order that one life would be saved – 599 "wasted" immunizations for the one that was effective."*
> (Rose, 1981, p. 1850).

How does this relate to the population attributable risk? Rose also could have said that the attributable (death) risk in non-vaccinated children was 17 per 10,000 non-vaccinated children. Hence, the vaccine would have prevented 17 deaths per 10,000 vaccinated children. Instead, he took the inverse of the attributable risk (that is, 1 over the attributable risk) to express the number of children that needed to be vaccinated in order to prevent one death: the number was [1 ÷ 0.0017 =] 588 children. The inverse of the attributable risk (1 ÷ AR) eventually became extremely popular in clinical epidemiology when repackaged under the acronym of NNT (Number needed to treat). Simple rule of thumb: NNT = 1 ÷ AR (Laupacis et al., 1988).

2.6. Risk and strength of association

In 1976 Kenneth Rothman proposed a "conceptual framework for causes" which offered the possibility of expressing the notion of risk in terms of conditions for disease causation.

> *"A cause is an act or event or a state of nature which initiates or permits, alone or in conjunction with other causes, a sequence of events resulting in an effect."*
> (Rothman, 1976, p. 588).

Rothman defined as a *sufficient* cause a *set* of causes, each of which, alone, was not sufficient to produce an effect. Figure 3 (Figure 1 of the paper), classically known today as "Rothman's causal pies", depicts three sufficient causes, each of which comprises 5 component causes.

In the paper, the different letters associated with each component cause served to explain a multitude of epidemiologic concepts, such as etiological fraction or interaction, which I do not describe here. But the contribution of the pies to the evolution of population thinking lies in their ability to conceptualize, and therefore bring to an even higher level of abstraction, the notions of "risk" and "strength of a causal risk factor". Consider that each component cause (which we may also call "risk factors") has a life of its own, and that it is only under some specific circumstances that it is united with other risk factors to form a sufficient cause, and therefore produce disease. Then:

"...the mean risk for a group indicates the proportion of individuals for whom sufficient causes are formed." (Rothman, 1976, p. 589).

This formulation of the notion of risk is tautological. If you accept the definition of the sufficient cause, then of course the risk is the probability that sufficient causes are formed. But the implication of this definition of risk opened the way to a new degree of understanding of the relation between the prevalence of risk factors, and the magnitude of the risk change they potentially incur in the population when a sufficient cause is completed:

"A component cause which requires, to complete the sufficient cause, other components with low prevalence is thereby a "weak" (component) cause. (...) On the other hand, a component cause which requires, to complete the sufficient cause, other components which are nearly ubiquitous is a "strong" (component) cause." (Rothman, 1976 p. 590).

Thus, the strength of a risk factor depends on the prevalence of the complementary risk factors needed to create a sufficient cause. This result has truly insightful implications with respect to population thinking:

"The characterization of risk factors as "strong" or "weak" has no universal basis (...) the strength of a causal risk factor (...) is dependent on the distribution in the population of the other risk factors in the same sufficient cause." (Rothman, 1976, pp. 589–590).

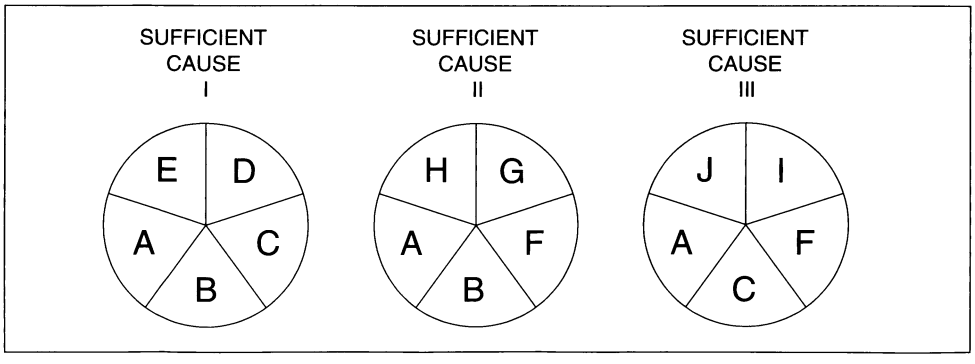


Figure 3
Conceptual scheme for the causes of a hypothetical disease. Source: Figure 1 of (Rothman, 1976).

The same risk factor can be strong in one population if its complement causes are common, and weak in another if its complement causes are rare. The textbook by Rothman and Sander Greenland, epidemiologist at University of California, Los Angeles (Rothman and Greenland, 1998, pp. 9–11) provides a numerical example, which is very useful for illustrating these concepts.

2.7. Evolution of population thinking in epidemiology

Let us at this point synthesize the lessons of the examples reviewed above. Population thinking in epidemiology has its roots in the 17th century. The premises of most measures of occurrence of disease are present in the probably first substantial contribution to population thinking (section 2.3.1). Graunt computed the proportion of deaths from the plague and the ratios of the number of births to the number of deaths. He also computed primitive weekly "rates" of plague mortality in order to compare the mortality during outbreaks of plague that lasted different numbers of years or to describe the evolution over time of a given epidemic. The distinction between rates and risks probably preceded Farr. Both measures are necessary to describe event (e.g., diseases) occurrence in a population.

John Snow used the ratio of the number of cholera deaths to the number of households provided for by each water company (section 2.3.2). This ratio could be an ambiguous measure of risk if the populations compared had different household sizes. This may explain why Farr attempted to relate cholera deaths to the number of people living in the affected neighborhoods as enumerated by the census. Farr's risks were expressed as proportions of inhabitants.

It was probably perceived that risks had to be expressed as percentages of people susceptible to getting the disease but, in public health, there could be problems in getting an appropriate denominator. Analyses in clinical settings did not suffer from this problem. Indeed, medical researchers of the 18th and 19th century commonly used proportions (Troehler, 2000; Morabia, 1996).

Simple proportions have been the most commonly used measures to describe the pattern of occurrence of acute infectious diseases in populations, which occur abruptly and have short average durations. Simple parameters suffice to describe them. The situation is very different for diseases that kill slowly and therefore last a long time. One type of disease kills quickly but lasts only days and does not accumulate in the population (e.g., cholera), while another type kills slowly but lasts for years and cases can accumulate over time (e.g., tuberculosis). Duration of disease is the factor that distinguishes these two types of disease.

Farr therefore divided the risk by the average duration of disease and this yielded a rate measure, that is, an average risk per unit of disease duration (section 2.4.1). The risk of dying from tuberculosis was 100%; the average time to death was 2 years. Thus, the death rate was [100 ÷ 2 years =] 50% per year or < 1% per week, etc.

When using risks instead of rates Farr had expressed the following caveat:

"To determine the mortality of diseases [= risk] they [the patients] should be followed from the beginning to the end; every death or recovery should be recorded; and this, though exceedingly simple, has rarely been done."
(Farr, Part II).

A measure of risk is *implicitly* related to a duration! We can express a risk of the same event over 1 week, 1 year, 10 years, etc. In the 1930s, Hill (see section 2.4.2) demonstrated, using a hypothetical example, the importance of this caveat in the context of a therapeutic trial comparing the ability of a vaccine to prevent disease attacks. By dividing the number of disease attacks by the total number of subjects inoculated during the year, we implicitly compute a risk over one year. However, if some or the entire inoculated group is followed for less than a whole year, the true risk would be underestimated. If this bias has different magnitude among the inoculated and the uninoculated, it may even produce a fallacious association between inoculation and disease risk. The solution was to use person-times in the denominator because it counted everyone's exposure for an appropriate length of time.

Epidemiologists of the 19th century had all the elements to find that the accumulation of cases in the population (prevalence) varied as a function of the incidence rate and the duration of disease. But there may not have been the need to distinguish the risk (the probability of occurrence of new cases) from the prevalence (the proportion of people with the disease in the population) as the diseases studied at that time rapidly ended up in either cure or death.

The spectrum of diseases changed in the 20th century, when major infectious scourges waned and chronic illnesses, such as cardiovascular diseases, cancer or infectious diseases with low case fatality rates surged. Consider the example of tuberculosis. Its incidence rate rapidly declined between 1900 and 1950. Still, a large fraction of the living population had been exposed to the Koch bacillus at some time in their life and had subclinical infections. The risk of getting infected was becoming low, but older people were still dying of infection contracted in the past. To accurately describe this situation, one needed to clarify how prevalence was related to risk. Around 1960, textbooks indicated that prevalence equaled the product of incidence and duration of disease. In the example given by MacMahon et al. (section 2.5.1), chronic and acute leukemia had similar incidence rates, but the prevalence of chronic leukemia was higher because its time to death was on average 2 years *vs.* about 2 months for acute leukemia. In reality, this simple, mechanical physics-looking expression, Prevalence = Incidence × Duration ($P = I \times D$), serves more heuristic than practical purposes. Its exact formulation is more complicated, and it is based on the assumption that the composition and disease experience of the population remains relatively stable.

Dividing a *risk* by the *duration* of disease yielded a rate. Multiplying an incidence *rate* by the *duration* of disease yielded a *prevalence*. What about the product of *prevalence* and incidence *rate*? Geoffrey Rose systematically explored this path and his findings were astonishing (section 2.5.2). They showed, contrary to what we would intuitively expect, that most of the cases of some chronic diseases, such as coronary heart disease, originate from the majority of the population who are at low risk for the trait. The rule was that a small risk applied to a large number of people generates an abundance of cases. The fraction of these cases that can be attributed to a given risk factor was obtained by computing the product of prevalence of exposure and the attributable (or excess) risk. This finding had a major implication for prevention: an efficient prevention strategy should consider targeting the mass of the population and not only the minority that is at high risk for the trait (Rose, 1981).

By the end of the 1960s, the distinction between risks and rates remained essentially conceptual: the number of incident cases was either divided by the number of persons at risk to form a risk, or it was divided by the number of person-times to form an incidence rate. This distinction was sufficient in practice but lacked mathematical rigor. The latter came from Miettinen's expression stipulating, in its simpler formulation, that the cumulative incidence over a time interval was the sum of the incidence densities computed over all the time sub-sections within the whole time interval (section 2.4.3). Considering that cumulative incidence is a synonym for risk and incidence density a synonym for incidence rate, we must acknowledge that quite a theoretical distance had been covered between Farr's *On Prognosis* (Farr, 2003) and Miettinen's *Estimability* (Miettinen, 1976a).

The developments have allowed epidemiologists to study more complex questions, more rigorously too. The future chapters of the evolution of population thinking in epidemiology are currently being written. A likely scenario is that the new concepts will become increasingly abstract and therefore difficult to illustrate using simplified examples as I have done here.

3. Group comparisons

3.1. Definition

Population thinking is indispensable for comparing groups, which is, as I will argue later, the main mode of knowledge acquisition in epidemiology. To compare groups we use measures of occurrence of events in populations. We compare prevalence, risks, rates, and odds.

The role of the comparison is to contrast what is observed in the presence of exposure to what would have occurred had the group of interest not been exposed to the postulated cause. Differences in frequency of disease occurrence between groups can be interpreted logically (albeit not always correctly) as being caused by

the exposure. There are two main study designs used in epidemiology to reach this goal.

In the first design, the groups differ in their exposure to the postulated cause (e.g., smoking). The occurrence of disease (e.g., risks or incidence rates) is the compared variable. I use three different terms for this type of study design: 1) exposed *vs.* non-exposed comparisons, because this is what it consists of; 2) cohort studies, because this is their most common current term, although the name was only coined around 1960 (MacMahon et al., 1960) and can hardly be used to describe earlier experiments; cohort studies are further divided, following the terminology used in the paper on the history of cohort studies (Doll, Part II), into a) *prospective* cohort studies, when cohorts are followed as they age; and b) *retrospective* cohort studies, when a substantial part of the follow-up is performed in the past, using historical data; 3) randomized controlled trials, which are a subform of cohort studies in which exposure (usually to a treatment) has been allocated in a random manner.

In the second design, the groups are either affected or non-affected by the studied outcome (e.g., lung cancer). Past exposure to the postulated cause (e.g., cigarette smoking) is the compared variable. I use again three different terms for this type of study design: 1) affected *vs.* non-affected comparisons, because this is what the comparisons consist of; 2) case-control studies, because this is their most common current denomination, although the name was only coined around 1960 (Morris, 1964) and can hardly be used to describe earlier experiments; 3) *nested* case-control studies, which are case-control studies designed within fully enumerated populations under investigation in a cohort study.

Let us consider examples of group comparisons that illustrate the evolution of epidemiologic concepts and methods over time.

3.2. Eighteenth century

The demonstration by James Lind (1716–1794), a Scottish naval physician, that (Lind, 1753) consumption of oranges and lemons could cure scurvy is an important step in the history of epidemiologic methods (Lind, 1753). It is a very early (if not the earliest) description of a group comparison to identify the treatment of a disease.

In the 18th century, scurvy was perceived as a terrible, rapidly fatal epidemic disease, which hit seamen on long voyages, campaigning armies, besieged cities and migrant populations. Scurvy is said to have eliminated 65% of Vasco de Gama's crew in 1498. An attack of scurvy could bring down in a few days seamen and soldiers in apparently good health. Affected persons became weak and had joint pain. Black-and-blue marks appeared on the skin. At the first visible signs of scurvy, red spots around the hair follicles covered the legs, buttocks, arms and back. Gums hemorrhaged and their tissue became weak and spongy. Teeth loosened and eating became difficult and painful. Stupor and death followed rapidly.

Table 9 – Description of treatment and outcomes in James Lind's 1747 experiment on 6 pairs of seamen suffering from scurvy. Source: (Lind, 1753).

Experimental pairs	Treatment for each pair member	Qualitative outcome
1	*"a quart of cider a day"*	*"improved"*
2	*"twenty five gouts of elixir vitriol three times a-day, upon an empty stomach, using a gargle strongly acidulated with it for their mouths"*	*"mouth but not internal improvement"*
3	*"two spoonfuls of vinegar three times a-day upon an empty stomach, having their gruels and their other food well acidulated with it, as also the gargle for their mouth"*	*"no remarkable alteration (...) upon comparing their condition with others who had taken nothing but a lenitive electuary and cremor tartar ..."*
4 *"two of the worst patients, with the tendons in the ham rigid (a symptom none the rest had)"*	*"half a pint of sea water every day, and sometimes more or less as it operated, by way of gentle physic"*	*"no remarkable alteration (...) compared to those who had taken nothing but a lenitive electuary ..."*
5	*"two oranges and one lemon given them every day. These they eat with greediness at different times upon an empty stomach. They continued but six days under this course, having consumed the quantity that could be spared."*	*"the most sudden and visible good effects were perceived from the use of the oranges and lemons; one of those who had taken them being at the end of six days fit for duty. The spots were not indeed at that time quite off his body, nor his gums sound; but without any other medicine than a gargarism of elixir vitriol, he became quite healthy before we came into Plymouth, which was on the 16th of June. The other was the best recovered of any in his condition; and being now deemed pretty well was appointed nurse to the rest of the sick."*

Table 9 – (continued)

Experimental pairs	Treatment for each pair member	Qualitative outcome
6 Reference group. The electuary and the cremor tartar was meant to *"keep their belly open"* and *"for relief of their breast"*.	*"...the bigness of a nutmeg three times a-day of an electuary recommended by an hospital surgeon made of garlic, mustard seed, rad. raphan., balsam of Peru and gum myrrh; using for common drink, barley-water well acidulated with tamarinds; by a decoction of which, with the addition of cremor tartar, they were gently purged three or four times during the course."*	*"no change"*

Scurvy was an important obstacle for naval supremacy. More seamen died of disease than of shipwrecks, battles, or famine. The diet of the sailors included cheese biscuits, salt beef, dried fish, butter, peas and beans. In retrospect, lack of fresh fruits or vegetables deprived the diet of vitamin C.

In 1731, Lind became a naval surgeon. In 1747, while serving on the 50 gun, 960 ton H.M.S. Salisbury, he carried out experiments on scurvy, which he published in 1753:

> *"On the 20th May, 1747, I took twelve patients in the scurvy on board the Salisbury at sea. Their cases were as similar as I could have them. They all in general had putrid gums, the spots and lassitude, with weakness of their knees. They lay together in one place, being a proper apartment for the sick in the fore-hold; and had one diet in common to all, viz., water gruel sweetened with sugar in the morning; fresh mutton broth often times for dinner; at other times puddings, boiled biscuit with sugar etc.; and for supper barley, raisins, rice and currants, sago and wine, or the like. Two of these were ordered each a quart of cider a day. Two others took twenty-five gouts of elixir vitriol three times a day upon an empty stomach, using a gargle strongly acidulated with it for their mouths. Two others took two spoonfuls of vinegar three times a day upon an empty stomach, having their gruels and their other food well acidulated with it, as also the gargle for the mouth. Two of the worst patients, with the tendons in the ham rigid (a symptom none the rest had) were put under a course of seawater. Of this they drank half a pint every day and sometimes more or less as it operated by way of gentle physic. Two others had each two oranges and one lemon given them every day. These they*

eat with greediness at different times upon an empty stomach. They continued but six days under this course, having consumed the quantity that could be spared. The two remaining patients took the bigness of a nutmeg three times a day of an electuary recommended by an hospital surgeon made of garlic, mustard seed, rad. raphan., balsam of Peru and gum myrrh, using for common drink nearly water well acidulated with tamarinds, by a decoction of which, with the addition of cremor tartar, they were gently purged three or four times during the course." (Lind, 1753, p. 145).

"The consequence was that the most sudden and visible good effects were perceived from the use of the oranges and lemons; one of those who had taken them being at the end of six days fit for duty. The spots were not indeed at that time quite off his body, nor his gums sound; but without any other medicine than a gargarism or elixir of vitriol he became quite healthy before we came into Plymouth, which was on the 16th June. The other was the best recovered of any in his condition, and being now deemed pretty well was appointed nurse to the rest of the sick." (Lind, 1753, p. 146).

"As I shall have occasion elsewhere to take notice of the effects of other medicines in this disease, I shall here only observe that the result of all my experiments was that oranges and lemons were the most effectual remedies for this distemper at sea." (Lind, 1753, p. 128).

Table 9 summarizes the results of the Salisbury experiment as described by Lind.

Note that Lind created comparable conditions of disease presentation, setting and diet before attributing the treatments. He also deliberately did not give any putatively active treatment to one of the groups, which served as control.

In Lind's view, oranges and lemons were "*remedies*" for scurvy. He did not mention scurvy *prevention*. His experiment demonstrated the *in*activity of sulfuric acid, vinegar, etc., which were the treatments officially recommended (Carpenter, 1986, p. 54).

Lind believed that scurvy was caused by both diet and some peculiarity of the air at sea, such as its "moisture"(Carpenter, 1986, pp. 60–61). This vision makes no sense to us, but in the 18th century, the hypothesis that some inadequacy in the diet could cause scurvy would have sounded absurd. Causes had to be related to some properties of gas or acid-alkaline reaction (Carpenter, 1986, pp. 40 and 75). We know now that a deficit in vitamin C is the cause of the metabolic disorders leading to the signs and symptoms of Lind's sick seamen. Primates share with guinea pigs the misfortune of not manufacturing their own vitamin C and having to obtain it from fresh food. Ascorbic acid (vitamin C) was isolated and synthesized in 1932.

It was Gilbert Blane (1749–1834), another Scottish physician, who, 40 years after Lind's Treatise, convinced the Lords of the Admiralty to supply a quarter of an ounce of lemon juice or lime to their seamen. English sailors to this day are called "limeys",

for lime was the term used at the time for both lemons and limes. Between 1895 and 1914, the Navy consumed 7,300 tons of lime. Epidemics of scurvy disappeared.

The work of Lind raises many fascinating questions. Why did Lind decide to conduct this comparative experiment? Who or what inspired him? Why Lind ? Why in 1740? I have not found answers to these questions. Apparently, seamen were often used as experimentation subjects in these years. Let's note that the sample size was so small that Lind must have been expecting an all-or-nothing answer. Indeed, only the two sailors in the orange and lemon pair became rapidly "fit for duty".

3.3. Nineteenth century

Three hundred years ago, medicine in Europe had to deal with the consequences of the colonial expansion of most of its States. "Fever" was the cardinal symptom of many different disorders:

> *"The 18th century struggle against fever has been compared, mutatis mutandis* [from the Latin: changing what needs to be changed], *with our present day efforts against cancer and arteriosclerosis. Both are the great killers of the times."* (Troehler, 1978, p. 78).

This observation can be extended to the first half of the 19th century. There were septic fevers following amputations but also puerperal, choleric, yellow fever, slow fever, diarrheic fevers, smallpox, malaria, hepatitis, and ophthalmic infections. Fevers were everywhere and physicians did not know how to cure them.

We will review here two episodes in which group comparisons yielded the correct answer about the treatment or the etiology of fevers, one in a clinical setting and the other in public health.

3.3.1. The bloodletting controversy

In the aftermath of the French Revolution, François Joseph Victor Broussais (1772–1838), an influential Parisian physician, a Jacobin, having served in the imperial army, was convinced that he had a solution to the therapeutic nightmare of fevers. He taught that fevers were manifestations of organ inflammation and that bloodletting and leeches were efficient to treat them all. Leeches had to be applied on the surface of the body corresponding to the inflamed organ. For example, the chest of a patient suspected of having tuberculosis was covered with multitudes of leeches. At the apogee of Broussais's influence, France used tens of millions of leeches per year. In 1833 alone, France imported 42 million of these annelid worms (Ackerknecht, 1967, p. 62).

Table 10 – Age, number of bleedings, duration of illness and risk of death according to day of first bleeding in Pierre-Charles-Alexandre Louis's "Researches on the effects of bloodletting ...". Source: (Louis, 1836).

Day of first bleeding	No of subjects	Mean age (years)	Duration of disease (days)	Mortality (%)
1–4	41	41	17.8	44
5–9	36	38	20.8	25
Total	77	40	19.2	35

Pierre Charles Alexandre Louis (1787–1872), another French physician, a contemporary of Broussais, who had had some experience as a clinician in Russia before he started practicing in France, was extremely doubtful about the validity of Broussais' theory. Louis published several monographs against Broussais' views, one of which is the "*Researches on the effects of bloodletting in some inflammatory diseases*". A first version appeared as an article in the 1828 Annales de Médecine Générale (Louis, 1828). This paper, revised and expanded, became a book in 1835 (Louis, 1835). The book was translated and published in English by an American student of Louis in 1836 (Louis, 1836).

In this book, Louis reports the following experiment. He had a large collection of case descriptions, which he had accrued during years of intensive clinical activity and autopsy in the Parisian Hospital La Charité. He found in his clinical records a total of 77 patients who were comparable because they had a well-characterized form of pneumonia (Morabia and Rochat, 2001) and were in perfect health at the time of the first symptoms of the disease. Twenty-seven of them had died. For each patient he computed the duration of illness from disease onset to death or recovery.

Louis compared the duration of disease and the frequency of death according to the time during the course of the disease when the patient underwent the first bleeding (Table 10). Louis grouped those first bled during days 1 to 4 of the disease (*early bloodletting*) and those bled for the first time during days 5 to 9 after the onset of the disease (*late bloodletting*). The two groups of patients were of comparable age. Duration of disease was on average 3 days shorter in those with early bloodletting (17.8 days) than in patients with late bloodletting (20.8 days). However, risk of death was 44% in the patients bled during the first 4 days of the disease compared to 25% among those bled later. These results ruled out the strong protective effect of early bleeding claimed by Broussais.

According to Louis

"a startling and apparently absurd result" (Louis, 1836, p. 9).

Louis did not conclude that bloodletting was useless but that it was much less useful than had been commonly believed:

> *"Thus, the study of the general and local symptoms, the mortality and variations in the mean duration of pneumonitis, according to the period at which bloodletting was instituted; all establish narrow limits to the utility of this mode of treatment."* (Louis, 1836, p. 13).

In his view, the validity of the technique was limited to severe cases of pneumonia:

> *"I will add that bloodletting, notwithstanding its influence is limited, should not be neglected in inflammations which are severe and are seated in an important organ; both on account of its influence on the state of the diseased organ; and because in shortening the duration of the disease, it diminishes the chance of secondary lesions."* (Louis, 1836, p. 23).

The data reported by Louis can be revisited with modern analytical tools (Morabia, 1996). They are available on *http://www.epidemiology.ch/index3.htm*. We can compare the prevalence of early bleeding in the group of patients who died with those who survived. Or we can compare the death risk in those bled in the first four days after disease onset *vs.* those bled more than four days after disease onset. Using a survival analysis, the group bled during the first four days of disease tended to do worse, but the difference was not statistically significant ($p = 0.07$). Also, if patients bled later in the course of the disease had a better prognosis, because they had already passed the worst phase of the illness, the bias would have favored late bleeding.

Louis was a meticulous clinician convinced of the importance of population thinking in medicine. He had understood that group comparisons were required to assess, in most situations, the true effect of treatments. His real impact on the practice of medicine is, however, hard to assess. Broussais had been the leader of Paris medicine since 1816 but after 1832, his theories rapidly lost support (Ackerknecht, 1967, p. 67). It took another century of progress in medical knowledge to completely settle the bloodletting controversy.

3.3.2. The London 1854 natural experiment

John Snow and William Farr can be considered as a 19th century English duet between a physician, primarily an anesthesiologist, and a "statistician", that is, someone who collected data for the "state" (Morabia, 2001a). It is thanks to their collaboration that the two fundamental elements of epidemiology, population thinking and group comparisons, merged around 1850 to produce the core of a new scientific discipline.

Most epidemiologists are familiar with John Snow's investigations of the 1854 epidemic of cholera in London and of the now famous outbreak around the Broad Street pump. His successful study of 1854, which I will briefly recall later, was preceded by an indefatigable pursuit of all the indices that might have put Snow on the right track (Shephard, 1995; Vinten-Johansen et al., 2003). In 1849, Snow analyzed the reports of the Registrar-General (i.e., William Farr's office) from September 23, 1848 to August 25, 1849. Using the population in 1841 as denominator, mortality varied between 1.10 per thousand inhabitants in the northern district and 7.95 in the southern districts of London (ratio 7.95 ÷ 1.10 = 7.2). (Shephard, 1995, p. 169).

Note that the ratios are quite large and Snow already suspected that this higher mortality originated from the supply of water polluted by sewage. Nevertheless, Snow still did not have a case for his hypothesis that cholera was transmitted by water, linen or foods contaminated by feces of sick people. First, the mortality rates remained small (between about 1 and 10 per 1,000 inhabitants per year), and skeptical opponents could invoke many alternative reasons for which mortality could be higher south than north of the Thames (Eyler, Part IIa).

The conditions for a rigorous group comparison occurred spontaneously. In 1852, one of the major water suppliers of London, the Lambeth Water Company, in accordance with an Act of Parliament, changed its source of Thames water. Its pumps were moved from near Hungerford Bridge, where the water was certainly soiled by sewage, to a place well outside London, beyond the influence of the tide and therefore out of reach of the London sewage. In contrast, another water supplier, the Southwark and Vauxhall Company, continued to draw its water from Battersea Fields, a seriously polluted area.

> *"London was without cholera from the latter part of 1849 to August 1853. During this interval an important change had taken place in the water supply of several of the south districts of London. The Lambeth Company removed their water works, in 1852, from opposite Hungerford Market to Thames Ditton; thus obtaining a supply of water quite free from the sewage of London. The districts supplied by the Lambeth Company are, however, also supplied, to a certain extent, by the Southwark and Vauxhall Company, the pipes of both companies going down every street, in the places where the supply is mixed, as was previously stated. In consequence of this intermixing of the water supply, the effect of the alteration made by the Lambeth Company on the progress of cholera was not so evident, to a cursory observer, as it would otherwise have been. It attracted the attention however, of the Registrar-General, who published a table in the 'Weekly Return of Births and Deaths' for 26th November 1853 (...)."*
> (Snow, 1855, pp. 41–42).

William Farr had noticed that the weekly mortality from cholera in the districts partly supplied by the Lambeth Company (61 per 100,000 inhabitants) was lower

than that for those districts entirely supplied by the Southwark and Vauxhall Company (94 per 100,000 inhabitants). Note that the rate (between 0.5 to 1 per 10,000 inhabitants per week) and the ratio ([94 ÷ 61 =] 1.5) imply that these were relatively rare events with a weak association. But it was when the cholera returned to London in July 1854, that John Snow resolved to make every effort to ascertain the exact effect of the water supply on the progress of the epidemic (Snow, 1855, p. 47).

It is key to understand that Snow had to invest an enormous amount of energy to create the conditions for a clear comparison of the mortality from cholera among the clients of either of the two large water suppliers. Farr provided Snow with the addresses of all cases of cholera (Eyler, Part IIa). During the first part of the epidemic, Snow himself went to each house and collected information on the exact provider. Clients often did not know the name of the provider. Snow explained that he had

> *"to distinguish the water from the two companies with perfect certainty by a chemical test [silver nitrate]"* (Snow, 1855, p. 48).

The test may not have been as accurate as Snow pretended (Eyler, Part IIa), but it reflects Snow's concern to clearly separate the exposure to the two sources of water supply. The most cited paragraph of the second edition of "*On the Mode of Communication of Cholera*" stresses the novel idea of *group comparisons* that Snow had striven to achieve:

> *"The experiment, too, was on the grandest scale. No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentlefolks down to the very poor, were divided into two groups without their choice, and, in most cases, without their knowledge; one group being supplied with water containing the sewage of London, and, amongst it, whatever might have come from the cholera patients, the other group having water quite free from such impurity."* (Snow, 1855, pp. 46–47).

During the first seven weeks of the epidemics there were 1,361 deaths from cholera in the districts supplied by the two companies (See Table 2): 1,263 (315 per 10,000) occurred in Southwark and Vauxhall districts *vs.* 98 (37 per 10,000) in those of the Lambeth Company. The ratio of [315 ÷ 37 =] 8.5 was of a magnitude comparable with the ratio between southern and northern districts observed in 1849, but in this case the two groups were compared on a specific factor, namely water supply. Skeptics could still argue that the association was caused by a third variable, such as poverty or elevation above sea level. But the argument could be "evacuated" by Snow who also compared the mortality from cholera of the houses supplied by the *same* company in 1849 and 1854, that is, before and after the Lambeth Company had moved its pumps to cleaner areas. Mortality had remained constant for the Southwark and Vauxhall clients but was four times lower for those of the Lambeth:

> *"The table exhibits an increase of mortality in 1854 as compared with 1849, in the sub-districts supplied by the Southwark and Vauxhall Company only, whilst there is a considerable diminution of mortality in the sub-districts partly supplied by the Lambeth Company. In certain sub-districts, where I know that the supply of the Lambeth Water Company is more general than elsewhere, as Christchurch, London Road, Waterloo Road 1st, and Lambeth Church 1st, the decrease of mortality in 1854 as compared with 1849 is greatest, as might be expected."* (Snow, 1855, p. 56).

Thus, the experiment offered a double perspective on group comparisons: concurrent differences in mortality, and "before and after" changes in exposure comparisons.

To appreciate this considerable achievement of John Snow, it is important to bear in mind the state of public health at these times, and in particular the work of Farr. Farr had created a unique and innovative system of standardized procedures for the collection, classification, analysis and reporting of causes of deaths. The compiling power of Farr's administration was, in Eyler's words, "herculean" given that there were no machines to treat all the information automatically (Eyler, Part IIa).

The description of the experiments and the extent of the co-operation between Farr and Snow (Eyler, Part IIa) all indicate that Snow would not have been able to perform his epidemiologic investigations without Farr's help. Snow's genius was to recognize the conditions of a natural experiment created when the Lambeth Company moved its water inlet to a less polluted area of the Thames. But it was Farr who first noted in 1853 the potential importance of the arrangement of water supply in South London. The following year, during the epidemic of 1854, Snow carried out his investigations and communicated to Farr his first results about the relation between cholera deaths and source of water supply. Snow writes that

> *"Dr. Farr was much struck with the result and, at his suggestion, the Registrars of all the south districts of London were requested to make a return of the water supply of the house in which the attack took place, in all cases of death from cholera."* (Snow, 1855, p. 47).

The discovery of the mode of transmission of cholera appears therefore as a successful synergy between Snow and Farr, that is, medicine and public health surveillance.

3.4. Evolution of confounding

The term confounding (from the Latin *confundere*, to mix together) characterizes, in epidemiology, situations where the group comparisons cannot distinguish between the effects of multiple causes. The measured association therefore is a mix of the ef-

fects of several causes. The mixed causes beyond the one studied are the confounders or confounding variables. The presence of confounding between binary variables, that is, variables that can be categorized as 0 or 1 such as gender, treatment or disease status, can be typically assessed when the measure of the association between one cause and disease yields different results in the full population, or separately across categories of the confounding variable(s).

We will review here four episodes illustrating the progressive refinement of the concept of confounding in the 20th century.

3.4.1. The paradoxical fate of a fallacy

In 1903 the Cambridge statistician G. Udny Yule (1871–1951) first demonstrated the mechanism underlying confounding. His demonstration then almost clandestinely traveled in two major epidemiology textbooks of the twentieth century (Greenwood, 1935; Hill, 1939), and finally received widespread recognition after having been rediscovered or re-baptized as a paradox (Simpson, 1951).

Yule formally described some "fallacies that may be caused by the mixing of records" using the hypothetical example:

> *"Some given attribute might, for instance, be inherited neither in the male line nor the female line; yet a mixed record might exhibit a considerable apparent inheritance."* (Yule, 1903, p. 133).

Table 11A cross-tabulates the presence of the "attribute" (e.g., pipe smoking) in fathers and sons. Indeed, 50% of the sons have the attribute, whether or not their father has it. There is therefore no hereditary transmission in men, but note that the attribute is very common.

Table 11B indicates that 10% of the daughters have the attribute, whether or not their mother has it. There is therefore no hereditary transmission in women either, but the attribute is less common in women than in men.

Table 11A – "On the fallacies that may be caused by the mixing of distinct records." Data about fathers and sons. Source: (Yule, 1903).

Sons	**Fathers**	
	Attribute present	**Attribute absent**
Attribute present	25	25
Attribute absent	25	25
Children with attribute	[25 ÷ 50 =] 50%	[25 ÷ 50 =] 50%

Table 11B – "On the fallacies that may be caused by the mixing of distinct records." Data about mothers and daughters. Source: (Yule, 1903).

Daughters	Mothers	
	Attribute present	**Attribute absent**
Attribute present	1	9
Attribute absent	9	81
Children with attribute	[1 ÷ 10 =] 10%	[9 ÷ 90 =] 10%

However, when considering sons and daughters together (Table 11C), it appears that the attribute is more common in children when one parent has the attribute: children are more likely to have the attribute (43%) when parents have it than when they don't (24%).

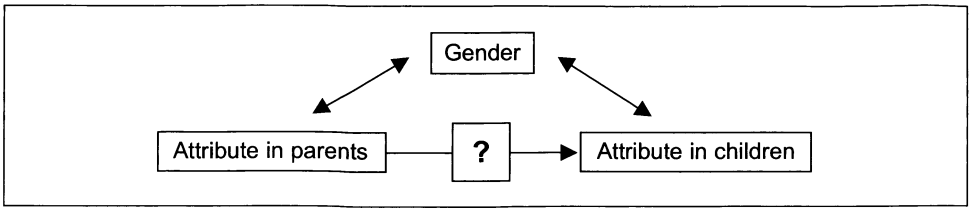
Table 11C – "On the fallacies that may be caused by the mixing of distinct records." Amalgamated data, both parents and children. Source: (Yule, 1903).*

Children	Parents	
	Attribute present	**Attribute absent**
Attribute present	26	34
Attribute absent	34	106
Children with attribute	[26 ÷ 60 =] 43%	[34 ÷ 140 =] 24%

* Yule presented in each cell the percentages of the 2 × 2 table total.

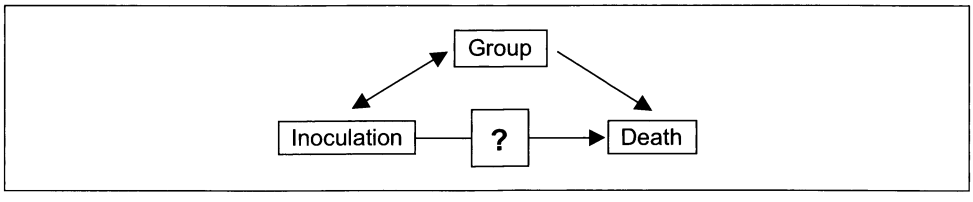
Yule gave the reason for the fallacy. The parent attribute was more common in fathers than in mothers and the children's attribute was more common in sons than in daughters. Gender was associated with having the attribute in children and with having the attribute in parents. His description announced the current definition of a binary confounder: a variable associated with both exposure and outcome. This created in the amalgamated 2 × 2 table a false association between attributes in parents and attribute in children. The situation can be represented using the arrow graphs that were introduced much later in the epidemiologic literature (Susser, 1973).

Interestingly for the fate of his fallacy, Yule suggested that a similar mixing of effect could occur in a trial yielding a fictitious association between "antitoxin" and "cure" if females exhibited a greater case fatality and the antitoxin was attributed more often to the men.



Yule had worked closely with Major Greenwood (1880–1947), eventually the first Professor of Epidemiology in the Department of Epidemiology and Vital Statistics at the London School of Hygiene and Tropical Medicine. Does this explain that 32 years after Yule's publication, an almost identical demonstration of the fallacy was made by Major Greenwood in his 1935 textbook: "*Epidemics and Crowd Diseases*" (Greenwood, 1935)? The example used by Greenwood is shown in Table 12. It refers to a potential "fallacy" that may occur when analyzing results from immunization experiments. It is found in the chapter entitled "*The artificial immunization of man*". This demonstration has remained, to my knowledge, unnoticed until recently (Zhang et al., Part II).

Table 12 shows that groups differ with respect to both prevalence of inoculation and risk of death. Group 1 has a higher mortality but group 2 has been more inoculated. Inoculation has no effect in groups 1 and 2 taken separately: 50% and 5% of the people, respectively, die. When mixing groups 1 and 2, inoculation becomes spuriously protective (9% of the inoculated *vs.* 46% of the uninoculated die). There is no reference to Yule's work. The whole citation is given as a footnote of Table 12. The relationships between "groups", "inoculation" and death can again be represented using an arrow graph. Note that this time, there is a unidirectional arrow between "group" and "death", as this relationship is likely to be causal:



Yule and Greenwood knew Bradford Hill well. Does this explain that 34 years after Yule's publication and 4 years after Greenwood's text, Hill presented an almost identical demonstration of Yule's fallacy in his 1939 textbook "*Principles of medical statistics*"? In Hill's example (Table 13), a treatment has no effect either in men or women, but appears to reduce mortality when the male and female 2 x 2 tables are collapsed. There is no reference to Yule or Greenwood's examples in Hill's text.

Hill's example corresponded exactly to Yule's suggestion: men were more treated and women died more. His explanation was:

Table 12 – Hypothetical example given by Major Greenwood of the potential fallacy resulting from mixing records. Source: "Epidemics and crowd-diseases" (Greenwood, 1935).*

	Group 1		Group 2		All	
	Inoculated	Non inoculated	Inoculated	Non inoculated	Inoculated	Non inoculated
Dead (n)	50	500	50	5	100	505
Alive (n)	50	500	950	95	1,000	595
All (n)	100	1,000	1,000	100	1,100	1,100
Percent dead	50	50	5	5	9	46

* *"One has data of the experience of inoculated and uninoculated persons collected over a wide range in space or time, and brings them together in a single statistical summary, which tells us that upon 'n' inoculated persons the attack-rate was 'a' per cent and upon 'm' uninoculated 'b' per cent. If n and m are large numbers, the kind of statistical test I have described may lead to arithmetically overwhelming odds in favour of the inoculated, yet this a priori inference might be quite wrong. It might be that in some of the experiments neither inoculated nor uninoculated ran any serious risk at all; if in these groups there were a great majority of inoculated, the final summary would show a great advantage to them. Suppose in one experiment there were 1,000 uninoculated with a death rate of 50 per cent and 100 inoculated also with a death rate of 50 per cent, while in another experiment there were 1,000 inoculated with a death rate of 5 per cent and 100 uninoculated also with a death rate of 5 per cent. Summarizing, we should find 1100 inoculated persons with 100 deaths, and 1100 uninoculated with 505 deaths, an enormous "advantage" to the inoculated group. No confidence should be placed in odds computed from such summaries."* (Greenwood, 1935, pp. 84–85).

> *"Superficially this comparison suggests that the new treatment is of some value; in fact that conclusion is wholly unjustified, for we are not comparing like with like (...). There are proportionally more females amongst the controls than in the treated group, and since females normally have a higher case fatality rate than males their presence in the control group in relatively greater numbers must lead to a comparatively high fatality rate in the total sample. Equally, their relative deficiency in the treated group leads to a comparatively low fatality rate in that total sample. No comparison is valid which does not allow for the sex differentiation of the fatality rates."* (Hill, 1939, p. 126).

A mysterious aspect of the fate of Yule's fallacy is that it is known today by most of us as "Simpson's paradox". In 1951, E. H. Simpson, for whom I have found no first name or biographical information, used an artificial example to demonstrate that,

*Table 13 – Fallacy resulting from mixing of non-comparable records *. Source: (Hill, 1939, p. 126, Table XIV).*

	Male		Female		All	
	Treatment	**No treatment**	**Treatment**	**No treatment**	**Treatment**	**No treatment**
Dead (n)	16	6	16	24	32	30
Alive (n)	64	24	24	36	88	60
All (n)	80	30	40	60	120	90
Percent dead	20	20	40	40	27	33

"Mixing of non-comparable records: (a) let us suppose that in a particular disease the fatality rate is twice as high among females as it is among males, and that amongst male patients it is 20 per cent. And amongst female patients 40 per cent. A new form of treatment is adopted and applied to 80 males and 40 females; 30 males and 60 females are observed as controls. The number of deaths observed among the 120 individuals given the new treatment is 32, giving a fatality-rate of 26.7 per cent., while the number of deaths observed amongst the 90 individuals taken as controls is 30, giving a fatality-rate of 33.3 per cent. Superficially this comparison suggests that the new treatment is of some value; in fact that conclusion is wholly unjustified, for we are not comparing like with like. The fatality-rates of the total number of individuals must be influenced by the proportions of the two sexes present in each sample; males and females, in fact, are not equally represented in the sample treated and in the sample taken as control. Tabulating the figures shows the fallacy clearly (Table XIV)." (Hill, 1939, pp.125–126).

under certain conditions, the relation of treatment to mortality could appear drastically different if it were analyzed before or after stratification by gender (Simpson, 1951). His example is shown in Table 14.

In Simpson's example, mortality is slightly lower in the treated groups, in both males and females, but this protective effect vanishes in the pooled comparison. In Yule's, Greenwood's and Hill's examples, "amalgamating" the gender-specific 2 × 2 tables fallaciously *produced* an effect. Simpson does not refer to the work of Yule, Greenwood and Hill. Epidemiologists citing Simpson's work do not usually refer to these earlier papers either (Rothman, 1986; Greenland, 1987a). Who was Simpson? Why is it that his paradox received so much visibility while that of Yule, Greenwood and Hill remained almost unnoticed?

The last episode of the saga of Yule's fallacy can be read in the text entitled "*Modern Epidemiology*". To illustrate Simpson's paradox Rothman told the example of

Table 14 – Original numerical example illustrating Simpson's paradox. Source: (Simpson, 1951).

	Male		Female		All	
	Treated	**Not treated**	**Treated**	**Not treated**	**Treated**	**Not treated**
Dead (n)	5	3	15	3	20	6
Alive (n)	8	4	12	2	20	6
All (n)	13	7	27	5	40	12
Percent dead	38	43	56	60	50	50

the delusion of a man who goes one day to buy a hat, tests some hats on two different tables and has the consistent impression that black hats fit him in general better than gray ones. The next day, however, hats from the two tables have been mixed together and now gray hats tend to fit him better. The data given by Rothman are shown in Table 15. It is funny that Rothman, who has a high appreciation of John Graunt, the British 18th century haberdasher (Rothman, 1996), invented an example about ... hats, rather than a health-related example as his predecessors.

There were two novelties in Rothman's example. First, pooling the hats from the two tables reversed the association. The best fit has moved from black to gray hat. More black hats fit on each of the tables, but more gray hats do when the content of the two tables is pooled. And second, it was followed by a mature theory of confounding:

> *"On the simplest level, confounding may be considered as a mixing of effects. Specifically, the estimate of the effect of the exposure of interest is distorted because it is mixed with the effect of an extraneous factor."* (Rothman, 1986, p. 89).

3.4.2. Early analyses of confounding

One of the first analytical adjustments for confounding was performed by Joseph Goldberger (1874–1929) and Edgar Sydenstricker (1881–1936). Goldberger and Sydenstricker provide another example of an intellectual duet comprising a physician and a statistician.

Goldberger, son of Hungarian immigrants, earned an MD at Bellevue Hospital, New York, had a private medical practice in Wilkes-Barre, Pennsylvania and then joined the United States Marine Hospital Service (later the U.S. Public Health Service) in 1899. Sydenstricker was born to missionary parents in Shanghai, received a Master's degree in sociology and economics in Virginia and, in 1907–1908, was a post-

Table 15 – Hypothetical example illustrating "Simpson's paradox". Source: (Rothman, 1986 p. 89).

	Table 1		Table 2		All	
	Black	**Gray**	**Black**	**Gray**	**Black**	**Gray**
Fit (n)	9	17	3	1	12	18
Not fit (n)	1	3	17	9	18	12
All (n)	10	20	20	10	30	30
Percent fit	90	85	15	10	40	60

graduate fellow in political economy at the University of Chicago. In 1920 he was appointed as Chief of the Office of Statistical Investigations in the U.S. Public Health Service (Wiehl, 1974).

Goldberger and Sydenstricker studied the causes of pellagra. Pellagra was first identified among Spanish peasants by Don Gaspar Casal in 1735 (Pan American Health Organization, 1988). A loathsome skin disease, it was called "mal de la rosa" and often mistaken for leprosy. In the United States, pellagra has sometimes been called the disease of the four D's – dermatitis, diarrhea, dementia, and death. By 1912, the state of South Carolina alone reported 30,000 cases and a case fatality rate of 40 percent, but the disease was hardly confined to Southern states. The US Congress asked the Surgeon General to investigate the disease. In 1914, Joseph Goldberger led that investigation.

Goldberger's theory on pellagra contradicted the medical opinion at that time. Pellagra was thought to be an infectious disease due to a still unidentified germ. The Thompson-McFadden Pellagra Commission, established under governmental auspices, had concluded in 1914 that pellagra had no relation to diet, based on an original house-to-house survey of pellagra cases in the cotton mill districts in South Carolina. The Commission's findings were interpreted as strong support for an infectious cause of pellagra (Elmore and Feinstein, 1994). But Goldberger had observed that, in mental hospitals and orphanages, the disease hit inmates but never staff. An infectious disease would not distinguish between inmates and employees.

Among all the experiments of various types that Goldberger and his coworkers carried out to study the causes of pellagra, one is especially relevant for this history of epidemiologic methods. In the spring of 1916, they began a methodologically remarkable investigation in some representative communities of South Carolina. Results have been reported in several papers published in 1920. I will focus here on one of them, "*A study of the relation of family income and other economic factors to pellagra incidence in seven cotton-mill villages of South Carolina in 1916*" (Goldberger et al., 1920).

The relation of poverty to pellagra incidence in the textile-mill communities was well established at that time. The typical sharecropper's lot was a wretched cottage, a few corn plants, and a luxurious but not edible growth of cotton (Roe, 1973). The objective of Goldberger's study was to assess whether diet could play a role, irrespective of poverty. They selected seven representative cotton-mill villages, enumerated their populations and sampled 750 households, comprising 4,160 people, exclusively Whites of anglo-saxon origin. It was

> *"an exceptionally homogenous group with respect to racial stock, occupation, and general standard of living, including dietary custom"* (Goldberger et al., 1920, p. 2678).

The homogeneity of the population did not result in being an obstacle in assessing the effect of diet on pellagra. Even villages that were similar in income were different in diet.

Pellagra incidence was assessed by a "systematic biweekly house-to-house search for cases". Cases had "clearly defined, bilaterally symmetrical dermatitis" (Goldberger et al., 1920, p. 2679). The assessment of diet and income was performed between April 16 and June 15, 1916, as this was the period immediately preceding the expected seasonal sharp rise in pellagra incidence in these villages. Food supply to the household was measured using an "accurate record for a 15-day period". The payroll records of the mills provided about 90% of the family income and statements of the housewife or other family members the other 10%. The half-month incomes so recorded were weighted by the number of "equivalent male units" of food requirement within a household. Weights were 1 for an adult male, 0.8 for an adult female, 0.5 for a child 5 to 9 years old, etc. Hence, in the paper, income is expressed as "half-month family income per adult male unit".

Table 16 shows that pellagra incidence declined rapidly as income increased. It is 16 times larger in households with less than 6 dollars (per half month and per adult male unit, adjusted rate: 41 per 1,000) compared to incomes of 14 dollars or more (adjusted rate: 2.5 per 1,000).

The footnote explaining the presence of "adjusted rates" (in the last column) is of great interest. It deals with an adjustment for age. The authors explained that they standardized the pellagra risks for age because age was associated with both income and pellagra incidence:

> *"Since a marked variation in the pellagra rate according to age and sex was found for the population studied (Goldberger, Wheeler, Sydenstricker, 1920b), and since, ordinarily, differences in the distribution of persons according to age occur in different economic groups, computation of rates adjusted to a standard population was made. The influence of differences in the sex distribution in any age group was insignificant, and practically the same incidence rates were obtained*

Table 16 – Number of definite cases of pellagra and risk per 1,000 among persons of different income classes in seven cotton-mill villages of South Carolina in 1926. Pooled men and women. Source: Adapted from Tables V and Va of (Goldberger et al., 1920, p. 2687).

Half-month family income per adult sale unit	Number of persons	Number of cases	Risk per 1,000	Adjusted risk per 1,000
Less than $6.00	1,312	56	42.7	41.0
$6.00–$7.99	1,037	27	26.0	24.8
$8.00–$9.99	784	10	12.8	14.2
$10.0–$13.99	736	3	4.1	5.2
$14 and over	291	1	3.4	2.5
All incomes	4,160	97	23.3	

after making adjustments to a standard age distribution, as is shown in the following table [last column of Table 16]." (Goldberger et al., 1920, p. 2687).

The total population (all incomes) served as standard population. They noted that the agreement between the crude and the adjusted risks ruled out the possibility that "differences in the sex and age distribution in the different income classes might give rise to" the inverse association of income and pellagra (Goldberger et al., 1920, p. 2689). In modern words, they ruled out a confounding effect of sex or age on the income-pellagra association.

Interestingly, they appropriately pooled the data of men and women: the gender-specific rates across income categories were substantially different but, if we compute the relative risks of income and pellagra incidence (which are not shown in the paper), we see that the relative associations are similar in men and women. Gender did not confound or modify the association. Hence, assessing it in the pooled sample was the correct approach.

After having demonstrated the "inverse correlation between pellagra incidence and family income", they showed that income was also strongly related to diet. Actually, both the lower income households and the pellagrous households had a diet rich in corn meal and grits and poor in milk and fresh products (Goldberger et al., 1920, Table VII p. 2692). Income was therefore, in modern terms, a possible confounder of the diet and pellagra association as it was related to both diet and pellagra incidence.

Goldberger and co-workers then proceeded to compare the diet of two villages (out of the seven studied) which showed the two most extreme incidence rates of pellagra: 0 per 1,000 in *Ny* and 64.6 per 1,000 in *In*. Both were poor villages with about half of their households having an income of 8 dollars or below. Note that this is now an affected/non-affected comparison, different from the exposed/non-exposed com-

parison shown in Table 16. They compared the fresh meat purchases in the two villages, one free from pellagra (i.e., the non-affected group) and the other severely affected. Because the two villages were similarly poor, there could be no effect of poverty in the comparison. In the non-affected village of *Ny*, 58.1% of the households had reported having purchased fresh meat twice or more in the 15 day record, whereas this proportion was only 8.5% in the affected village of *In*. This analysis demonstrated therefore that within homogeneous categories of income, dietary differences determined incidence of pellagra. The effect of diet was not confounded by income.

This study performed in 1916 and published in 1920 reveals therefore a profound understanding of the issue of confounding and familiarity with some analytical tools to adjust for confounding effects. The effect of income was first identified and clearly separated from a possible age effect by *adjustment*. Then the effect of diet was separated from that of income by *restricting* the analysis to two low-income villages with differing pellagra rates. In addition, the authors had used both exposed/unexposed and affected/unaffected comparisons. The summary and conclusions testify to the progress of epidemiologic concepts and methods in observational studies by 1920:

> *"4. In general, pellagra incidence was found to vary inversely according to family income ... 5. The inverse correlation between pellagra incidence and family income depended on the unfavorable effect of low income on the character of the diet; but family income was not the sole factor determining the character of the household diet. 6. Differences in incidence among households of the same income class are attributable (...) to the differences among household with respect to availability of food supplies from such sources as home-owned cows, poultry, gardens, etc. 7. Differences in incidence among villages whose constituent households are economically similar, are attributable to differences among them in availability of food supplies (...). 8. The most potent factors influencing pellagra incidence in the village studied were: (a) low family income, and (b) unfavorable conditions regarding the availability of food supplies, suggesting that under the conditions obtained in some of these villages in the spring of 1916 many families were without sufficient income to enable them to procure the adequate diet, and that improvement of food availability (particularly of milk and fresh meat) is urgently needed in such localities."* (Goldberger et al., 1920, p. 2711).

Goldberger was right on target. We now know that pellagra is caused by the lack in the diet of a vitamin, niacin or nicotinic acid, belonging to the B complex. Niacin can be found in yeast, organ meats, peanuts, and wheat germ. The disease is most common in areas where the diet consists mainly of corn, which, unlike other grains, lacks niacin as well as the amino acid tryptophan, which the body uses to synthesize the vitamin. Pellagra can be prevented and treated by niacin (Roe, 1973).

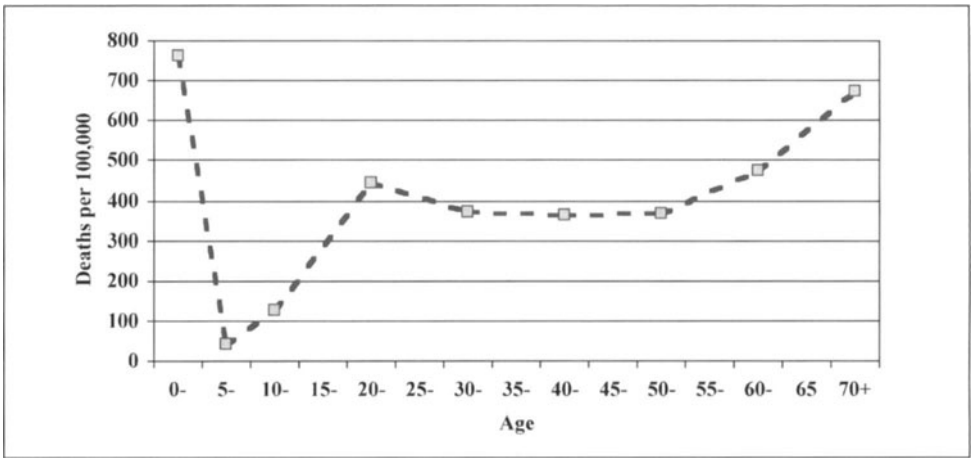


Figure 4
Massachusetts death rates from tuberculosis – all forms – by age, in 1880. Source: (Frost, 1939).

3.4.3. Cohort analysis

Around 1900, investigating the causes of tuberculosis raised new types of group comparison problems. There could be a long latency between exposure to *M. tuberculosis* and the clinical manifestations of the disease. Each population comprised a mixture of people who had been infected at various times in the past and who remained so for decades before they eventually became sick and died of the disease. Infection rates were declining. Therefore, at a given moment in time, older people were more likely to carry the bacillus than younger ones. Vital statistics showed that the mortality from tuberculosis increased with age, but it was unclear whether this was because of an age effect or because of the higher exposure during their youth of the older people.

Wade Hampton Frost (1880-1938), the first Professor and Chairman in the Department of Epidemiology and Public Health Administration at The School of Hygiene and Public Health of the Johns Hopkins University in Baltimore, addressed this question in a paper published after his death (Frost, 1939). He described the fallacy that may occur when naively interpreting cross-sectional changes in death rates with age (Comstock, Part II; Doll, Part II).

Figure 4 presents the change of mortality rates from tuberculosis across age groups in 1880 in Massachusetts. They peak at ages 0–4, are lowest at ages 5 to 9, and then rise across age groups. This apparent age effect was difficult to explain:

> *"...nothing that we know of the habits of mankind and the distribution of the bacillus would lead us to suppose that between the first and the second 5 years of*

Table 17 – Key for the interpretation of cohort versus cross-sectional mortality rates.

Age	Calendar year 1880	Calendar year 1890	Calendar year 1900
0–9	$M_{1,1}$ = 1880 mortality rates for those born in 1871 to 1880		
10–19	$M_{2,1}$ = 1880 mortality rates for those born in 1861 to 1870	$M_{2,2}$ = 1890 mortality rates for those born in 1871 to 1880	
20-29	$M_{3,1}$ = 1880 mortality rates for those born in 1851 to 1860		$M_{3,3}$ = 1900 mortality rates for those born in 1871 to 1880

$M_{i,j}$ = mortality rate for people in the i^{th} age group and the j^{th} calendar year.

> *life there is, in general, a diminution in exposure to infection which corresponds to the decline in mortality rate. And there is little, if any, better reason to suppose that the extraordinary rise in mortality from age 10 to age 20, 25 or 30 is paralleled by a corresponding increase in rate of exposure to specific infection."* (Frost, 1939).

Frost therefore proposed to study the age effect using the death rates within the same "cohort" at different ages. The term "cohort" comes from the Latin *cohors*. The antique Roman legions were composed of ten cohorts. The 480 warriors plus 6 centurions of a cohort constituted the basic fighting unit that could be traced during the battle. The term cohort has been imported to epidemiology to define a set of people who are followed or traced over a period of time. Frost showed that the cohort is the appropriate unit in which one can assess an age effect. Table 17 illustrates the rationale of Frost's reasoning. The example is based on the data used by Frost and reproduced by Comstock (Comstock, Part II).

In Table 17, the first column indicates *cross-sectional* age-specific mortality rates in 1880. Differences in mortality rates in the column can be read as the change of mortality rates with age, where each age group corresponds to a different birth cohort, that is, people born at different time points in the past. The numbers on the diagonal indicate the change in mortality rates with age for the 1871–1880 birth cohort, that is, people who were all born between 1871 and 1880. The cross-sectional (e.g., cell $M_{2,1}$) and the cohort ($M_{2,2}$) age-specific mortality rates would be similar if exposure and/or susceptibility (e.g., due to vaccination) to infection did not change between 1871 and 1900. In the absence of a cohort effect, $M_{2,1} = M_{2,2}$, and $M_{3,1} = M_{3,3}$. For a *cohort effect* to take place, mortality at a given age must change over time,

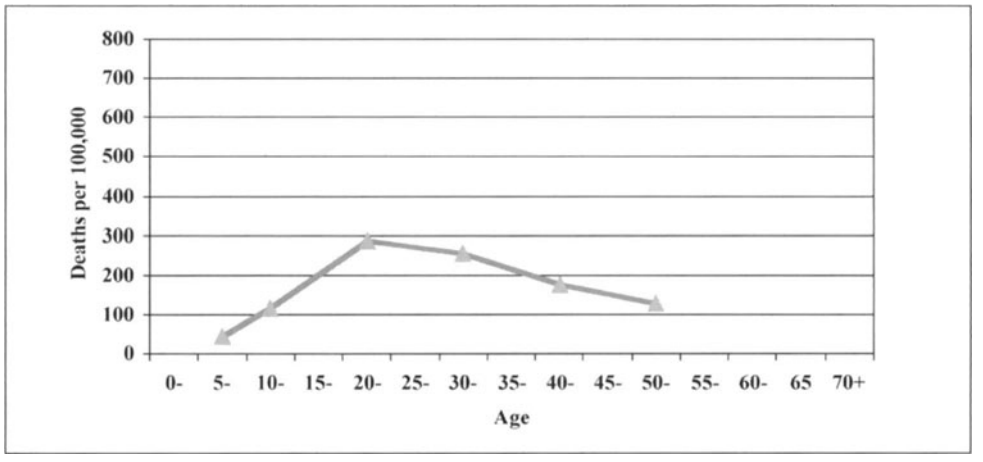


Figure 5
Massachusetts death rates from tuberculosis – all forms – by age, for the cohort of men who were born in the years 1871 to 1880. Source: (Frost, 1939).

so that the *cross-sectional* age-specific frequency is different from the cohort-and-age-specific frequency: $M_{2,1} \neq M_{2,2}$, and $M_{3,1} \neq M_{3,3}$.

Mortality rates from tuberculosis were changing rapidly across cohorts in the populations studied by Frost. The cohort analysis demonstrated that rates tended to decline with age after age 20, as shown in Figure 5 for the cohort of men born between 1871 and 1880.

Similar observations across several cohorts allowed Frost to conclude that the inexplicable variations of mortality rates with age occurring in the cross-sectional analysis were due to differences in exposure to tuberculosis across birth cohorts. The cohort analysis did not support the hypothesis that a lower exposure to *M. tuberculosis* during infancy resulted in more severe infections in adults. This would have been a major argument against vaccination. Frost concluded that:

> *"Present day 'peak' of mortality in late life does not represent postponement of maximum risk to a later period, but rather would seem to indicate that the present high rates in old ages are the residuals of higher [exposure] rates in earlier life."* (Frost, 1939).

Actually, Frost had described, without using the term, another manifestation of confounding, between age, birth cohort and mortality from tuberculosis, which could be identified by stratifying on birth cohort.

3.4.4. Alternate allocation of treatment

The contribution of Bradford Hill to the evolution of group comparisons is a major one, in particular for the implementation of a technique to prevent the confounding effects of "disturbing and extraneous factors" in therapeutic trials. It is worth quoting extensively of the section dedicated to "planning and interpretation of experiments".

> *"Thus, when the statistician's help is required, it is his task to suggest means of allowing for the disturbing causes, either in planning the experiment or in analysing the results, and not as a rule, to determine what are the relevant disturbing causes."* (Hill, 1939, p. 4).

Hill explains in simple terms the foundations of confounding:

> *"If we find that Group A differs from Group B in some characteristic, say, its mortality-rate, can we be certain that difference is due to the fact that Group A was inoculated (for example) and Group B was uninoculated? Are we certain that Group A does not differ from Group B in some other character relevant to the issues as well as in the presence or absence of inoculation? For instance, in a particular case, inoculated persons might, on the average, belong to a higher social class than the uninoculated and therefore live in surroundings in which the risk of infection was less."* (Hill, 1939, pp. 4–5).

Hill then explains the role of alternate allocations of treatment:

> *"The reason why in experiments in the treatment of disease the allocation of alternate cases to the treated and untreated groups is often satisfactory, is because no conscious or unconscious bias can enter in, as it may in any selection of cases, and because in the long run we can fairly rely upon this random allotment of the patients to equalize in the two groups the distribution of other characteristics that may be important. Between the individuals within each group there will often be wide differences in characteristics, for instance, in body-weight and state of health, but with large numbers we can be reasonably sure that the numbers of each type will be equally, or nearly equally, represented in both groups."* (Hill, 1939, pp. 5–6).

Hill mentions a form of treatment allocation that would be blocked on specific characteristics (e.g., sex) to increase the likelihood of getting comparable groups:

> *"If it be known that certain characteristics will have an influence upon the results of treatment and on account of relatively small numbers the distribution of these characteristics may not be equalized in the final groups, it is advisable to extend*

Table 18 – Effect of serum treatment on case fatality in patients with type I or type II lobar pneumonia. Aberdeen, London, Edinburgh and Glasgow, United Kingdom, 1930–1933. Source: (Table III Therapeutic Trial Committee of the Medical Research Council, 1934).

Pneumonia	Age (years)	Conventional treatment			Serum treatment			
		N	Deaths	% case fatality [A]	N [B]	Deaths	% case fatality	Expected deaths [A × B]*
Type I	20–40	224	25	11.2	140	8	5.7	16
	40–60	77	20	26.0	44	10	22.7	11
						18		27
Type II	20–40	194	44	22.7	111	14	12.6	25
	40–60	111	38	34.2	53	19	35.8	18
						33		43

* The "expected deaths" are those which would have been recorded if the serum-treated groups had died at the same percentage rates as the corresponding controls.

> *this method of allocation. For instance, alternate persons will not be treated but a division will be made by sex, so that the first male is treated and the second male untreated, the first female is treated and the second female is untreated. Similarly age may be equalized by treating alternate males and alternate females at each age, or in each broad age-group if individuals whose ages are within a few years of one another may in the particular case be regarded as equivalent."*
> (Hill, 1939, pp. 5–6).

Hill cites as example the report of the Therapeutic Trials Committee of the Medical Research Council on the serum treatment of an infection located in lobes of the lung and called lobar pneumonia. Hill had been an investigator in the trial (Therapeutic Trial Committee of the Medical Research Council, 1934). The trial was conducted almost simultaneously in London, Edinburgh, Aberdeen and Glasgow. Patients admitted in Aberdeen, London, Edinburgh for a pneumonia received, alternatively according to the order of admission, either a serum treatment or the conventional treatment, which served as control. In Glasgow, the control group was selected from another hospital without alternate allocation of treatment. The treatment with serum was begun within some hours after admission, but some patients must have been excluded after treatment was allocated as "all patients dying within 24 hours of admission to hospital were taken out of the series". Results of the trial are shown in Table 18.

The mortality risk in the conventional treatment (i.e., controls) was applied to the numbers of people treated with serum to compute the expected number of deaths under the assumption of no serum treatment effect. The analysis was stratified by age. The expected number of deaths was then compared to the number of deaths observed in those treated with serum. Serum appeared to reduce the number of deaths more for type I (18 observed death *vs.* 27 expected) than for type II pneumonia (33 observed *vs.* 43 expected). The effect was stronger in younger subjects. Results were actually very similar when using only the data from the centers, which had used the alternate allocation (London, Edinburgh and Aberdeen).

The example reviewed here paved the way for the development of the modern form of the randomized controlled trial. The James Lind Library provides more elements on this episode and on his place in the history of randomized allocation of treatments (Chalmers, 2004).

3.4.5. Logic of confounding

The logic of confounding received a further boost after the publication, in 1950, of case-control studies showing that smoking was a likely cause of lung cancer. Among those who were skeptical about the link of tobacco to lung cancer stood Ronald A. Fisher (1890–1962), from Cambridge, perhaps the most original statistician of the 20th century (Fisher, 1959; Stolley, 1991). Even though Fisher articulated many criticisms against a causal implication of smoking (Stolley, 1991), his essential argument was that both smoking and lung cancer resulted from genetic predispositions. To support his thesis he presented data shown in Table 19.

Data suggested that some genetic predisposition could explain the habit of smoking. Homozygotic twins (whose genetic constitution is almost identical) were more alike in their smoking behavior than heterozygotic twins (who have only half of their genes in common). Of 51 homozygotic twin pairs, 39 (76%) had similar smoking habits, as opposed to less than half (15 ÷ 31) among heterozygotic twins.

Table 19 – Smoking habits of heterozygotic and homozygotic twins. Source: (Fisher, 1959, p. 40).

Smoking habits	Heterozygotic twins		Homozygotic twins	
	Pairs	Percent	Pairs	Percent
Different	16	52*	12	24
Alike or somewhat alike	15	48	39	76
Total	31	100	51	100

* Fisher's original is 51 but should be 52.

Fisher's idea was that the association between smoking and lung cancer was spurious. Smoking was related to some genetic predisposition, which caused lung cancer. Fisher did not use the word confounding, but his point was that the relation of smoking to cancer was "confounded" by genetic predisposition. The historical interest of this example is that Fisher had articulated the suspicion of confounding in a form that would become typical in epidemiology.

The reason for Fisher's open antagonism is not established. It may well have been another manifestation of the historical dispute between Fisher and his fellow statistician but declared enemy, Karl Pearson (1857– 1936), Galton Professor of Eugenics at Cambridge University. Fisher must have felt that he had to criticize the work of all disciples of Karl Pearson. Fisher died in 1962, before the publication of the US Surgeon General report on Smoking and Health.

3.5. Case-control studies

The principle of the case-control study is the comparison of past exposures between groups of affected and non-affected subjects. The case-control study may be looked upon as a natural extension of the practice of physicians to take case histories as an aid to diagnosis. (Mantel and Haenszel, 1959; Paneth et al., Part II). To get clues about the etiology of the disease, clinicians may have begun the tradition to compare patients suffering from a specific disease with other patients who were free of that disease.

We usually refer to this technique today as a case-control study, but it was first called a "retrospective study". The term "retrospective" meant that the researcher went back from the disease to its potential causes in the past, just as a physician obtains the personal history from a patient.

In 1950, two studies of that type conducted in the United States to assess the relation of smoking to lung cancer were published in the Journal of the American Medical Association (Wynder and Graham, 1950; Levin et al., 1950). Both indicated that exposure to tobacco smoke was more common in lung cancer cases than in controls.

The first study had been led by Morton L. Levin (1904–1995), a student of Frost in the mid-1930s, then hired as cancer epidemiologist at Roswell Park Memorial Institute, Buffalo, New York. Around 1948, Levin and his colleagues identified 1,507 men admitted to Roswell Park Memorial Institute, between 1938 and 1948. The proportion of subjects who had smoked for more than 25 years was 54.1% in lung cancer cases, 34.9% in other cancer controls, 36.9% in lung non tumors, and 29.8% in non-cancer controls (Levin et al., 1950). Levin et al concluded that:

> *"The data suggest, although they do not establish, a causal relation between cigarette and pipe smoking and cancer of the lung and lip, respectively. The statisti-*

> *cal association may, of course, be due to some other unidentified factor between these types of smoking and lung and lip cancer."* (Levin et al., 1950 p. 474).

Ernst L. Wynder (1922–1999) was still a pre-med student in Saint Louis when he began studying cancer. In 1948 and 1949, supported by Professor Evarts A. Graham (1883–1957), Chairman of the Department of Surgery at Washington University School of Medicine, Wynder (not alone) interviewed 605 men and 25 women with lung cancer (other than adenocarcinoma) from several hospitals and private practices around the United States (Wynder, 1997). Controls were 780 men and 552 women without cancer of the lung admitted to general hospitals. The conclusion of Wynder and Graham's paper was that

> *"excessive and prolonged use of tobacco, especially cigarets [sic], seems to be an important factor in the induction of bronchiogenic carcinoma"* (Wynder and Graham, 1950, p. 336).

The studies by Levin et al. and Wynder et al. had a number of strengths (Paneth et al., Part II), but methodologically they were not different from the studies conducted in the pre-war era. It is once again a study involving Bradford Hill that indicates a methodological watershed. In September 1950, Hill and Richard Doll, at that time a research assistant whom Hill had invited to investigate the causes of lung cancer, published the preliminary report of a study commissioned by the Medical Research Council (Doll and Hill, 1950). This study is now viewed as a model case-control investigation because, for the first time, it had been conceived and designed as such to solve a specific question and generate valid results (Paneth et al., Part II). Doll and Hill were therefore able to answer the question of the relation of smoking to lung cancer more thorougly and more convincingly.

Twenty hospitals of London informed Doll and Hill of their diagnosed cancer cases (lung, colon, stomach, rectum). Cancers of the colon, stomach and rectum served as "contrasting groups". Research almoners (i.e., social workers) interviewed the cases as well as a patient of the same sex, within the same five-year age group, and in the same hospital at or about the same time, who did not have lung cancer. Attention was paid to the duration of smoking, to histories of starting and stopping smoking, and to the amount smoked. Contrasts were made between cases of lung cancer and matched controls in overall smoking, amount smoked most recently, maximum ever smoked, age of onset of smoking, type of tobacco and duration of smoking. Stratified analyses were used to deal with potential confounders, including urban/rural residence.

Table 20 presents the simplest analysis of the case-control study results. The conclusion that cigarette smoking could cause lung cancer was based on refined analysis of the data, presented in many tables and figures. I like, however, the presentation of the data as in Table 20 because we may not perceive today that 95.8% of the controls had smoked. Even though almost all cases were smokers of some sort, there was

Table 20 – Percentage of ever smokers in cases of lung cancer and hospital controls. Source: (Doll and Hill, 1950).

Smokers	%Cases (n = 649)	%Controls (n = 649)
Yes	99.7	95.8
No	0.3	4.2

space for skepticism. Indeed, it is with skepticism that these results were received in the medical and political community.

The authors firmly concluded that cigarette smoking was

> *"a factor, and an important factor, in the production of carcinoma of the lung."* (Doll and Hill, 1950).

The British Medical Journal wrote a favorable review (Editorial, 1950). But a retrospective account of the events surrounding the publications of these three case-control study articles shows that things had not been easy (Armenian and Szklo, 1996; Wynder, 1997; Terris, 1997; Doll, 1998). Doll has explained how the publication of their study was delayed:

> *"By the end of 1949 the position was so clear that we had written a paper based on our findings in 709 pairs of lung cancer cases and control patients drawing the conclusion that (I quote) 'smoking is a factor, and an important factor, in the production of carcinoma of the lung'. When, however, we showed the paper to Sir Harold Himsworth, who had by then succeeded Sir Edward Mellanby as Secretary of the Medical Research Council, he wisely advised us to postpone publication until we had checked that similar results would be reproduced outside London. We consequently withheld publication and started to interview similar groups of patients in some of the principal hospitals in and around Bristol, Cambridge, Leeds and Newcastle. Before we had obtained much more data, however, Wynder and Graham (1950), reported very similar findings in their study of patients in the US, and we consequently published ours a few months later (in September 1950) without waiting for the results of the extended study. The latter was published in 1952, relating to 1,465 pairs of patients and controls and showed essentially identical results in all centers – except that heavy smoking by women had not spread outside London."* (Doll, 1998, p. 134).

Both the Royal College of Physician's 1962 report entitled *Smoking and Health*, and the US Surgeon General's Report of the same title, published in 1964, relied heavily on case-

control studies in their assessment of the evidence. The Royal College of Physicians Committee cited 23 case-control studies, all of which showed a relationship of smoking to lung cancer, and the Surgeon General's Report cited 29 such studies, all but one of which (a study in women) confirmed the association. The powerful consistency of these case-control studies, and the replication of their findings in cohort studies promoted the general acceptance of the case-control study as a scientific research tool.

3.6. Cohort studies

The conclusion that cigarette smoking was an important cause of lung cancer was accepted by

> *"very few other scientists at the time, who were unaccustomed to the idea that firm conclusions about causation could be drawn from case-control studies, and it was clear that if the conclusion was to be widely accepted the conclusions would have to be checked by some other method of enquiry"* (Doll, Part II).

The case-control design was deemed susceptible to all sorts of biases, based either on inaccurate recall or on selection. It was believed that it led more often to erroneous conclusions than to correct ones (White, 1990) and that it was inherently biased (Doll, 1984). In response to the criticisms expressed towards case-control studies, Doll and Hill designed a new type of study based on very different premises. In their 1954 paper on "*The mortality of doctors in relation to their smoking habits*" (Doll and Hill, 1954), they noted that a number of studies had been made of the smoking habits of patients with and without lung cancer and that further studies of the same kind were unlikely to shed new light upon the nature of the association. An entirely new approach was needed, which would be free of the potential flaws of case-control studies. They proposed to call the new approach "prospective", which the Oxford English Dictionary defined as "characterized by looking forward into the future". They sent a short questionnaire eliciting smoking habits to 59,600 British Doctors. The history of the British Doctor Study is told by Richard Doll in this book (Doll, Part II).

In January 1, 1952, E. Cuyler Hammond (1912–1986) and Daniel Horn (1916–1992), respectively Director and Assistant Director of statistical research at the American Cancer Society, launched the U.S. counterpart of the British Doctor Study, but with an almost four times larger sample size. They designed and pretested a questionnaire on smoking habits, trained 22,000 American Cancer Society volunteers and asked each of them

> *"to have the questionnaire filled out by about 10 white men between the ages of 50 and 69 whom they knew well and would be able to trace"* (Hammond and Horn, 1958).

They received 204,547 completed questionnaires from California, Illinois, Iowa, Michigan, Minnesota, New Jersey, New York, Pennsylvania and Wisconsin. The health status of the participants was checked each year and death certificates obtained for men recorded as dead. In 1958, Hammond and Horn published in the Journal of the American Medical Association an analysis of the death rates in relation to the smoking habits of 187,783 men, traced from 1952 through 1955, for an average 44 months and representing 667,753 man-years (Hammond and Horn, 1958). The analysis consisted in comparing the observed number of deaths to

> *"...the number of deaths which would have occurred among men in each smoking category if their age-specific death rates had been exactly the same as those for men who never smoked. This will be referred to as the 'expected' number of deaths."* (Hammond and Horn, 1958, p. 1160).

Hammond and Horn used the observed and expected number of deaths to compute both the mortality ratio (observed divided by expected) and the excess deaths (observed minus expected). Their study was so large that they were able, after a relatively short follow-up, to ascertain the potential associations of tobacco smoke with many causes of death beyond lung cancer. Table 21 presents some of the results of the American Cancer Society cohort study.

A notable aspect of the paper was that it presented both the mortality ratio and the excess deaths. These two measures of effect combined revealed important features of the health effects of smoking. Mortality *ratios* indicated that the strongest association was with lung cancer: smokers had 10.73 times the risk of dying from lung cancer compared to never smokers. The association was weaker with coronary artery disease (mortality ratio = 1.70). The excess deaths indicated, however, that coronary artery disease was, by far, a more common cause of excess deaths, since it accounted for 52.1% of all excess deaths in smokers compared to non-smokers. Both relative (mortality ratio) and absolute causality (excess deaths) were needed to fully understand the effects of smoking on health:

> *"the relative importance of the association is dependent on the number of deaths attributed to each disease, as well as on their degrees of association with cigarette smoking"* (Hammond and Horn, 1958, p. 1308).

A group of American statisticians and epidemiologists, including Hammond (Cornfield et al., 1959), would re-express one year after the publication of Hammond and Horn's study the need for both absolute and relative measures of association in epidemiology:

> *"Relatively, cigarettes have a much larger effect on lung cancer than on cardiovascular disease, while the reverse is true if an absolute measure is used. Both the*

Table 21 – Mortality ratio and excess deaths of various causes among men with a history of regular cigarette smoking. American Cancer Society, 1952 cohort study. Source: (Hammond and Horn, 1958).

Cause of death	Observed deaths	Mortality ratio (observed ÷ expected)	Excess deaths (observed-expected)	Percentage of all excess deaths (%)
Coronary artery disease	3,361	1.70	1,388	52.1
Lung cancer	397	10.73	360	13.5
Other cancer	1,063	1.50	359	13.5
Other heart and circulation disorders	676	1.30	154	5.8
Pulmonary (except cancer)	231	2.85	150	5.6
Cerebral vascular	556	1.30	128	4.8
Gastric and duodenal ulcers	100	4.00	75	2.8
Cirrhosis and liver	83	1.93	40	1.5
All other	849	1.01	11	0.4
Total	7,316		2,665	

> *absolute [attributable or excess risk, risk difference] and the relative measure [relative risk, odds ratio] serve a purpose. The relative measure is helpful in (...) appraising the importance of an agent with respect to other possible agents inducing the same effect (...). The absolute measure would be important in appraising the public health importance of an effect known to be causal."* (Cornfield et al., 1959, p. 194).

The importance of smoking as a cause of coronary artery disease may have been misinterpreted if the association with smoking had only been reported as a mortality ratio, or more generally, as a relative risk. The study stressed the importance of looking at the data under two different perspectives, easily derived from cohort studies. One type is related to the interpretation of relative risks. It is a very intuitive concept; e.g., the risk in the exposed is twofold, threefold, etc. greater (or smaller) relative to the risk in the unexposed. Relative risks are useful to identify risk factors, even when the disease is extremely rare. They do not require population data on prevalence or incidence and, therefore, can be estimated from a case-control study. However, a relative risk of 10 can be obtained from a ratio of 10/1, 100/10, or 10,000/1,000. It does not reflect the public health or clinical importance of the association. In contrast, the various forms of attributable risks (synonymously defined as risk difference, excess

risk, absolute risk, or excess deaths in Hammond and Horn's paper) corresponding to the previous relative risks of 10 are, respectively, 9, 90 and 9,000 cases, say, per million over a given time period. They have a straightforward public health or clinical interpretation: the exposure causes an absolute number of cases in excess over a given time period. I have proposed elsewhere to call these two types of perspectives relative and absolute causality (Morabia, 2001b).

3.7. Selection bias

Another opponent of the smoking-lung cancer association was Joseph Berkson, (1899–1982) who graduated (MD and ScD) from The Johns Hopkins University and later became Head of the Biometry and Medical Statistics Division at the Mayo Clinic in Minnesota. It is in that position that he developed a theoretical mechanism of bias, known today as "Berkson's bias" or "Berkson's fallacy", that could plague hospital-based case-control studies (such as those of smoking and lung cancer) and therefore invalidate their findings (Berkson, 1946).

Berkson's argument was that case-control studies comparing hospital patients with different diagnoses – note that the three influential smoking and lung cancer studies known to Berkson were hospital-based – could yield false associations only as a result of a selective process of hospitalization. Conceptually, if a larger fraction of all exposed cases was likely to be hospitalized than that of exposed controls, then a comparison of hospitalized cases and controls would find an association between smoking and lung cancer even if no such association existed in their population. Berkson's paper demonstrated that this was possible mathematically and in doing so probably represented the "first algebraic analysis of an epidemiologic selection bias" (Greenland, 1987b, p. 86).

Berkson's paper starts with explaining the essential difference between a case-control study, which he refers to as the "practical statistics", and the laboratory experiment. The laboratory experiment compares groups of exposed and unexposed animals. The outcome is a true random variable, while in the case-control study, we search for an association between exposure and disease after disease has already occurred:

> *"all the effects are already produced before the investigation starts"*
> (Berkson, 1946).

The paper uses the example of a hospital-based case-control study of the association of diabetes (cases) and cholecystitis (exposure). Controls are ophthalmology patients who came to the clinic to get glasses because of refractive errors. Berkson demonstrates that, under specific assumptions, the case-control study may spuriously observe an excess of cholecystitis of 2.32% (± 0.5%) in patients with diabetes than

Table 22 – Population frequency, referral rates, and hospital frequency for exposed and unexposed cases and controls. Source: Table 5.2. in (Schlesselman, 1982, p.129).

Exposure	Disease	Population frequency	Proportion referred	Hospital frequency
Yes	Case	A	s_1	s_1A
	Control	B	s_2	s_2B
No	Case	C	s_3	s_3C
	Control	D	s_4	s_4D

Population odds ratio: $\Psi = AD \div BC$
Hospital odds ratio: $\Psi' = [(s_1s_4) \div (s_2s_3)]$ Ψ = bias $\times$ Ψ

among controls while there is no such association in the whole population from which cases and controls originate (Berkson, 1946).

The mechanism underlying Berkson's bias was elegantly explicated by the epidemiologist James J. Schlesselman in his 1982 "Case-control Studies" (Schlesselman, 1982). To facilitate a modern interpretation of these data, Schlesselman shows the impact of the bias on the odds ratio, that is, the ratio of the odds of exposure in the cases over the odds of exposure in the controls:

> *"Differential referral patterns are another source of potential bias in hospital or clinic-based case-control studies. Table 5.2* [Table 22 above] *shows that differential rates of hospitalization for exposed and unexposed cases and controls can distort the odds ratio- determined in the hospital from that in the population. Whereas the population odds ratio is $\Psi = AD \div BC$, the odds ratio in hospital is $\Psi' = b\ \Psi$ where the bias term $b = (s_1s_4) \div (s_2s_3)$ depends on the (usually unknown) differential referral rates s_1, s_2, s_3, and s_4 defined in Table 5.2* [Table 22 above]." (Schlesselman, 1982, p. 128).

Table 23 presents Berkson's data from a hypothetical hospital-based case-control study of the association of diabetes with cholecystitis, in which cases suffer from diabetes and controls from ocular refractive errors requiring glasses, using Schlesselman's notation defined in Table 22.

Applying Schlesselman's factorization of the cross-product ratio of the sampling fractions to Berkson's data, we get:

Population $\Psi = AD \div BC = (3{,}000 \times 960{,}300) \div (29{,}700 \times 97{,}000) = 1$
Hospital $\Psi' = [(s_1s_4) \div (s_2s_3)] \times \Psi = [(0.2087 \times 0.20) \div (0.32 \times 0.069)] \times 1 = 1.89 \times 1 = 1.89$

Table 23 – Example of a hypothetical hospital-based case-control study of the association of diabetes with cholecystitis, in which cases suffer from diabetes and controls from ocular refractive errors requiring glasses. The proportions referred are: 0.05 for diabetes, 0.2 for refractive errors and 0.15 for cholecystitis. All forces of hospitalization are independent of each other. Source: (Berkson, 1946).

Exposure	Disease	Population frequency	Proportion referred*	Hospital frequency
Yes	Case	A = 3,000	$s_1 = 0.2087$	626
	Control	B = 29,700	$s_2 = 0.32$	9,504
No	Case	C = 97,000	$s_3 = 0.069$	6,693
	Control	D = 960,300	$s_4 = 0.20$	192,060
Odds ratio**		$\Psi = AD \div BC =$ $(3{,}000 \times 960{,}300) \div$ $(29{,}700 \times 97{,}000)$ $= 1$		$\Psi' = [(s_1 s_4) \div (s_2 s_3)] \times \Psi$ $= [(0.2087 \times 0.20) \div$ $(0.32 \times 0.069)] \times 1$ $= 1.89 \times 1 = 1.89$

* The reader should refer to Berkson's paper to understand how these probabilities were computed.

** Computed using the equations of Table 22.

Thus, diabetes is not associated with cholecystitis in the population (odds ratio = 1), but it is in the hospital-based study because of the differential sampling fractions (or forces of hospitalization) of the different categories of cases and controls (odds ratio = 1.89).

Berkson's argument was not based on real data and has never been clearly demonstrated empirically. Some investigators got close though (Vineis, Part IIa; Roberts et al., 1978). There was, however, an element in Berkson's example that was peculiar: the "exposure" was a disease (i.e., cholecystitis), which, alone, could contribute to hospitalization. In case-control studies, exposures are usually risk factors, which are not sufficient motives of hospitalization. e.g., being a smoker does not lead to hospitalization independently of the diseases that smoking causes. Therefore, it was argued that in Berkson's example the bias occurred only because cholecystitis *contributed independently* to hospitalization (Kraus, 1954). Otherwise, there would have been no selection bias (i.e., using Schlesselman's notation, s1 = s3 and s2 = s4). The counterargument therefore was to Berkson's criticism that epidemiologic studies of smoking and cancer could not have been threatened by Berkson's bias.

Berkson's criticism did not end up hurting epidemiology. On the contrary, it stimulated the development of a formal theory of selection (or response) biases, of which Berkson's bias has since become a classic example. It was shown that there were

many conditions under which the imbalanced selection of cases and controls would not lead to selection bias:

> *"Physicians or self referral are two of many selective factors operating to produce the final case-control series in any hospital-based study. In general, if one regards the terms s1 to s4 as the sampling proportions for the four cells of the 2 × 2 table with population frequencies A, B, C, and D in Table 5.2 [Table 22 above], then a general condition for the absence of bias in the estimation of the odds ratio is that* $b = 1$, *implying that* $s_1 s_4 = s_2 s_3$. *For example, if among cases one is k times more likely to choose an exposed individual, and if among controls one is also k times more likely to choose an exposed individual, then* $s_1 = ks_3$ *and* $s_2 = ks_4$, *resulting in* $b = 1$. *Thus, in principle, a biased selection of cases can be compensated by a biased selection of controls. However, one usually strives to choose both cases and controls in a manner that assures that exposed and unexposed individuals have equal probabilities of selection, that is,* $s_1 = s_3$ *and* $s_2 = s_4$."
> (Schlesselman, 1982, p. 128).

The theory of selection bias in case-control studies was later expanded to losses of follow-up in cohort studies and further generalized. About 40 years after Berkson's paper, epidemiologists had arrived at a well-formalized theory of selection and response bias.

3.8. Interaction

The concept of interaction is used in epidemiology to define a situation in which an association differs in subgroups of the population. It implicates at least three elements: an exposure, an outcome and another factor, which is sometimes referred to as the "effect modifier". The interaction between fava bean consumption, hemolytic anemia and the glucose-6-phosphate dehydrogenase genetic deficiency is an extreme example, in which the association between fava beans and hemolytic anemia can be observed when the genetic variant is present but not when it is absent. More commonly, an effect modifier modulates the effect of the studied exposure. There is *synergy* when the effect modifier amplifies the effect of exposure. There is *antagonism* when the effect modifier reduces the effect of exposure.

According to Major Greenwood, the Roman physician Galen (AD129-AD210) considered that ill-health depended on the interaction between *temperament*, *procatarctic* factors and *constitutions*, which correspond grossly to our current genetic, behavioral and environmental risk factors:

> *"Let us imagine, for instance, that the atmosphere is carrying diverse seeds of pestilence, and that, of the bodies exposed to it, some are choked with excremen-*

titious matters apt in themselves to putrefy, that others are void of excrement and pure. Let us further suppose an obstruction of orifices and resultant plethora in the former, likewise a life of luxury, much junketing, drinking, sexual excess and the crudities which must attend on such traits; in the latter let us suppose cleanliness, freedom from excrementitious matters, orifices unobstructed and uncompressed, desirable conditions, as we may say, free transpiration, moderate exercise, temperance in diet. All this being supposed, judge thou, which class of body is likelier to be injured by the inspiration of putrid air."
(Greenwood, 1935, p. 27).

Thus, for Galen, an environmental risk factor is more likely to affect a debauched than an ascetic person. If the logical content of the epidemiologic concept of interaction can be found in antiquity, the theory of interaction is absent from the epidemiologic literature until the 1960s. The relation between exposure to asbestos, smoking and lung cancer has become the classical example of interaction between several causes.

Asbestos was used on a very large scale during the 20th century, most particularly for construction and public infrastructures (see Stellman, Part II). Its carcinogenic effect was first noted as an occupational disease. During the 1950s, workers exposed to asbestos were more likely to develop lung cancer than the general population (Doll, 1955), but they, as most of the male population, usually also smoked. Whether asbestos had an *independent* contribution to the risk of lung cancer remained to be demonstrated. The group led by Irving Selikoff (1915–1992), Director of the Division of Environmental and Occupational Medicine at Mount Sinaï Hospital, New York, assembled a cohort of the members of a union, *The International Association of Heat and Frost Insulators and Asbestos Workers*, that is, workers exposed to asbestos and working in the United States or Canada.

The cohort comprised 17,800 workers who had filled out a questionnaire in 1966. The causes of death were systematically registered after that. In 1976, 397 cases of lung cancer occurred among the 12,051 workers who had been exposed at least 20 years to asbestos. They contributed 77,391 person-years of follow-up.

As, by definition, members of the *International Association* were all exposed to some degree to asbestos, Selikoff searched for an external cohort of workers who would be comparable in terms of work conditions but essentially unexposed to asbestos. With Cuyler Hammond, they identified a subgroup of the American Cancer Society cohort study described above (Hammond and Horn, 1958), comprising 73,763 men who had blue-collar jobs in environments rich in dust, fumes and vapors but not asbestos. These blue-collar workers had been followed up between 1967 and 1972.

The results of this study comparing a cohort of asbestos workers to a cohort of blue-collar workers are shown in Table 24 (Hammond et al., 1979).

Table 24 – Age-standardized lung cancer death rates (per 100,000 per year) for cigarette smoking and/or occupational exposure to asbestos dust compared with no smoking and no occupational exposure to asbestos dust. Source: (Hammond et al., 1979, p. 487).

Group	Exposure to asbestos	History cigarette smoking	Death rate	Mortality difference*	Mortality ratio**
Control	No	No	11.3	0.0***	1.00***
Asbestos workers	Yes	No	58.4	+47.1	5.17
Control	No	Yes	122.6	+111.3	10.85
Asbestos workers	Yes	Yes	601.6	+590.3	53.24

* Attributable risk.
** Relative risk.
*** Reference group.

The authors interpreted the data in Table 24 as follows:

"The mortality differences shown here were calculated by subtracting the death rate of the "no, no" group from the death rate of each of the four groups. The mortality ratios were calculated by dividing the death rate of each group by the death rate of the "no, no" group.
The mortality ratios are 1.00 for "no, no" (asbestos, no; cigarette smoking, no); 5.17 for "yes, no" (asbestos, yes; cigarette smoking, no); 10.85 for "no, yes" (asbestos, no; cigarette smoking, yes) and 53.24 for "yes, yes" (asbestos, yes, cigarette smoking, yes).
Now, suppose that occupational exposure to asbestos dust and cigarette smoking acted independently in respect to the production of lung cancer. In that event, the lung cancer death rate of asbestos workers with a history of cigarette smoking should be very close to the sum of the following three numbers: 11.3 (the rate for the "no, no" group), 47.1 (the mortality difference for the "yes, no" group), and 111.3 (the mortality difference for the "no, yes" group). The sum comes to 169.7 lung cancer deaths per 100,000 man-years which is a reasonable estimate of what the lung cancer death rate of the asbestos workers with a history of cigarette smoking would have been if there had been no synergistic effect of the combined exposure. In contrast, the observed lung cancer death rate of the "yes, yes" group was 601.6 per 100,000 man-years. The difference (601.6 – 169.7) = 431.9 lung cancer deaths per 100,000 man-years, was presumably due to a synergistic effect in men with both of the two types of exposure (asbestos dust and cigarette smoking)." (Hammond et al., 1979).

The mortality difference of asbestos workers who smoked (relative to blue-collar workers unexposed to asbestos and non-smokers) was expected to be 158.4 /100,000/yr, which corresponds to the sum of the individual effects of smoking and asbestos. But it was 590.3, that is, much larger than expected. The authors therefore concluded that there was synergy because smoking amplified the mortality difference (i.e., absolute causality) due to asbestos. But they would have come to a very different conclusion if their reasoning had been based on the mortality ratio (i.e., relative causality). The mortality ratios are [58.4 ÷ 11.3 =] 5.17 for asbestos alone, [122.6 ÷ 11.3 =] 10.85 for smoking alone. In the absence of interaction between asbestos and smoking we would expect the mortality ratio for those exposed to both asbestos and smoking to be [5.17 × 10.85 =] 56.09, which is very similar to the observed mortality ratio (53.24). What was then the correct interpretation?

Table 24, or similar findings observed in other studies, provoked a controversy about the definition and interpretation of interaction that took place, mostly in the American Journal of Epidemiology between 1976 and 1980. Three of the contenders coauthored a paper which they hoped would "lay to rest" the concept of interaction:

> *"We believe that the controversy surrounding the concept of interaction can be laid to rest with specification of the context in which the interaction is being evaluated. Four broad contexts can be distinguished: statistical, biological, public health, and individual decision-making. Each has different implications for the evaluation of interaction."* (Rothman et al., 1980).

They distinguished four types of interactions according to the purpose or the context: statistical, biological, public health and individual decision-making.

1) The evaluation of *statistical* interaction depended on the model chosen, whether additive (modeling risk differences) or multiplicative (modeling relative risks). It served to describe the relation between the two factors and the outcome irrespetive of the nature of their biological links.

Checking in Table 24 for smoking, asbestos and lung cancer, there is *additive interaction* as the observed attributable risk (AR, referred to in the table as mortality difference) is greater than the AR expected if the AR for smoking and the AR for asbestos were independent:

> *Observed AR for smoking & asbestos = 590.3*
> *Expected AR for smoking & asbestos if no interaction = 47.1 + 111.3 = 158.4*

In contrast, Table 24 does not suggest *multiplicative interaction* as the observed relative risk (RR, referred to in the table as mortality ratio) is not substantially different from the RR expected if the RR for smoking and the RR for asbestos were independent:

Observed RR for smoking & asbestos = 53.24
Expected RR for smoking & asbestos if no interaction = 5.17 × 10.85 = 56.09

2) In *biological* interaction, the choice of the additive or multiplicative model was based on some speculation about the underlying biologic model, whether the two factors were believed to act additively or multiplicatively.

3) *Public health* interaction had to be evaluated using *additive* models, as an additive interaction implied that the preventive yield of a public health intervention would differ according to the target population. Let's turn again to the example of smoking, asbestos and lung cancer. The AR associated with preventing smoking among asbestos-exposed workers is:

AR (smoking & asbestos) – AR (asbestos alone) = 590.3 – 47.1 = 543.2

The corresponding absolute risk reduction of lung cancer associated with removing asbestos exposure among smoking workers is:

AR (smoking & asbestos) – AR (smoking alone) = 590.3 – 111.3 = 479.0

The attributable risk for removing smoking among asbestos workers appears therefore greater (543.2 per 100,000 per year) than that of removing asbestos among smoking workers (479.0 per 100,000 per year). Hammond et al. (1979) had chosen the right model. Public health interaction can lead to key strategic choices, even if in practice things are not that simple and the absolute number of cases prevented by each of the interventions would require considering the prevalence of smoking and of asbestos exposure.

4) The *individual decision-making* interaction is similar to the public health interaction but in the context of medical practice. The presence of additive interaction between a specific drug (e.g., oral contraceptive), a risk factor (e.g., hypertension) and a disease (e.g., stroke) may imply that the drug is contraindicated (or particularly beneficial) in subgroups of patients.

Interaction is the most recent of the epidemiologic concepts. It has not been "laid to rest" yet.

3.9. Causal inference

Causal inference is another long-lasting conceptual development fostered by the smoking-lung cancer controversy. The criteria used by epidemiologists today to establish a plausible causal connection are primarily associated with the name of, once again, Bradford Hill (Hill, 1965). Hill's work summarized a generation of thought by several eminent epidemiologists including Jacob Yerushalmy (1904– 1973), Carroll

E. Palmer (1909–1969), Abraham Lilienfeld (1920–1985), Philip Sartwell (1908–1999) and Mervyn Susser.

Hill's 1965 paper entitled: *"Environment and disease: Association or causation"* has been so influential that it is worth citing some parts at length. Hill starts by stating the question underlying causal inference:

> *"In what circumstances can we pass from this observed association to a verdict of causation? Upon what basis should we proceed to do so? (...)The decisive question is whether the frequency of the undesirable event B will be influenced by a change in the environmental feature."* (Hill, 1965, p. 295).

Causal inference comes after we have ruled out the role of chance or bias in the interpretation of the data:

> *"Our observations reveal an association between two variables, perfectly clear-cut and beyond what we would care to attribute to the play of chance. What aspects of that association should we especially consider before deciding that the most likely interpretation of it is causation?"* (Hill, 1965, p. 295).

Then comes the list of nine aspects that tend to characterize causal relations, in the order given by Hill and, when relevant, accompanied by an example.

> 1. Strength: *"...prospective inquiries into smoking have shown that the death rate from cancer of the lung in cigarette smokers is nine to ten times the rate in non-smokers and the rate in heavy cigarette smokers is twenty to thirty times as great. But to explain the pronounced excess in cancer of the lung in any other environmental terms requires some feature of life so intimately linked with cigarette smoking and with the amount of smoking that such a feature should be easily detectable". [However] "We must not be too ready to dismiss a cause-and-hypothesis merely on the grounds that the observed association appears to be slight."* (Hill, 1965, pp. 295–296).

> 2. Consistency: *"...the consistency of the observed association. Has it been repeatedly observed by different persons, in different places, circumstances and times ?(...) The Advisory Committee to the Surgeon-General of the United States Public Health Service found the association of smoking with cancer of the lung in 29 retrospective and 7 prospective inquiries (US Department of Health, Education & Welfare 1964). The lesson here is that broadly the same answer has been reached in quite a wide variety of situations and techniques. In other words we can justifiably infer that the association is not due to some constant error or fallacy that permeates every inquiry. (...)I would myself put a good deal of weight upon similar results reached in quite different ways, e.g. prospectively and retrospectively."* (Hill, 1965, pp. 296–297).

3. Specificity: *"If, as here, the association is limited to specific workers and to particular sites and types of disease and there is no association between the work and other modes of dying, then clearly that is a strong argument in favor of causation. (...) If other causes of death are raised 10, 20 or even 50% in smokers whereas cancer of the lung is raised 900–1,000% we have specificity - a specificity in the magnitude of the association."* (Hill, 1965, p. 297).

4. Temporality: *"...which is the cart and which the horse? Does a particular diet lead to disease or do the early stages of the disease lead to those peculiar dietetic habits?"* (Hill, 1965, pp. 297–298).

5. Biological gradient: *"For instance, the fact that the death rate from cancer of the lung rises linearly, with the number of cigarettes smoked daily, adds a very great deal to the simpler evidence that cigarettes smokers have a higher death rate than non-smokers. (...)The clear dose-response curve admits of a simple explanation and obviously puts the case in a clearer light."* (Hill, 1965, p. 298).

6. Plausibility: *"It will be helpful if the causation we suspect is biologically plausible. But this is a feature I am convinced we cannot demand. What is biologically plausible depends upon the biological knowledge of the day."* (Hill, 1965, p. 298).

7. Coherence: *"On the other hand the cause-and-effect interpretation of our data should not seriously conflict with the generally known facts of the natural history and biology of the disease - in the expression of the Advisory Committee to the Surgeon-General it should have coherence. Thus in the discussion of lung cancer the Committee finds its association with cigarette smoking coherent with the temporal rise that has taken place in the two variables over the last generation and with the sex difference in mortality - features that might well apply in an occupational problem. The known urban/rural ratio of lung cancer mortality does not detract from coherence, nor the restriction of the effect to the lung."* (Hill, 1965, p. 298).

8. Experiment: *"Occasionally it is possible to appeal to experimental, or semi-experimental, evidence. For example, because of an observed association some preventive actions are taken."* (Hill, 1965, pp. 298–299).

9. Analogy: *"In some circumstances it would be fair to judge by analogy. With the effects of thalidomide and rubella before us we would surely be ready to accept slighter but similar evidence with another drug or another viral disease in pregnancy."* (Hill, 1965, p. 299).

Then comes the crucial caveat that the aspects of causal relations should not be summed as a causality score:

"Here then are nine different viewpoints from all of which we should study association before we cry causation. What I do not believe – and this has been suggested – is that we can usefully lay down some hard-and-fast rules of evidence that must be obeyed before we accept cause and effect. None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a sine qua non. What they can do, with greater or less strength, is to help us to make up our minds on the fundamental question – is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect?" (Hill, 1965, p. 299).

Followed by the often forgotten reminder that there is no statistical test for causal inference:

"No formal tests of significance can answer those questions. Such tests can, and should, remind us of the effects that the play of chance can create, and they will instruct us in the likely magnitude of those effects. Beyond that they contribute nothing to the "proof" of our hypothesis. (....) Fortunately I believe we have not yet gone so far as our friends in the USA where, I am told, some editors of journals will return an article because tests of significance have not been applied. Yet there are innumerable situations in which they are totally unnecessary – because the difference is grotesquely obvious, because it is negligible, or because, whether it be formally significant or not, it is too small to be of any practical importance." (Hill, 1965, p. 299).

The elaboration of a structured approach to causal inference accompanied the preparation of the historical 1964 Surgeon General report stating that cigarette smoking caused lung cancer (US Department of Health, 1964). The unrelenting rise in cigarette sales in the US and in Europe from the 1920s was finally curbed in the early 1980s, at least among men.

3.10. The rare disease assumption

Superficially, cohort and case-control studies may appear to be two designs with opposite logic. Jerome Cornfield (1912–1979) has been, among many other prestigious positions, Chairman of the Department of Biostatistics at The Johns Hopkins University and Director of Biostatistics Center at The George Washington University. It is of historical interest to note that Cornfield was President successively of the American Epidemiologic Society (in 1972) and of the American Statistical Association (in 1974). Cornfield played a decisive role in demonstrating the close link between cohort and case-control study designs and therefore creating the basis for the modern understanding of case-control studies.

Table 25 – Results of the case-control study of cigarette smoking and lung cancer among white males of aged 40–49 used as an example by Cornfield (Cornfield, 1951). Source: (Schrek et al., 1950).

Cigarette per day	Lung cancer cases	Controls with tumors of other sites
10 or more	$p_1 = 77\%$	$p_2 = 58\%$
Else	$1\text{-}p_1 = 23\%$	$1\text{-}p_2 = 42\%$
N	35	171

P = proportion of cigarette smoking.
[Odds Ratio = $[p_1 \div (1\text{-}p_1)] \div [p_2 \div (1\text{-}p_2)] = (0.77 \div 0.23) \div (0.58 \div 0.42) = 2.42$]*.
* Cornfield does not use the term odds ratio.

In 1951, Cornfield showed that it was possible to estimate a relative risk (in principle, only computable in a cohort study) from case-control study data. The procedure assumed that the cases and the controls were representative of the same groups in the general population. Cornfield's procedure is explained in Tables 25 and 26. Table 25 gives the case-control data and notation that Cornfield used as an example. They came from a paper by Schrek et al. (Schrek et al., 1950; Paneth et al., Part II).

The primary result of the study was that smoking 10 cigarettes or more per day was more common in cases ($p_1 = 77\%$) than in controls ($p_2 = 58\%$). Intuitively, it is logical to expect that if more cases of lung cancer smoked, the risk of lung cancer was greater in smokers. But, mathematically, there was apparently no relationship between the proportions of exposed in cases and controls and the smokers/non-smokers risks of lung cancer.

Going beyond simple proportions, p1 and p2 could be re-expressed as the odds of smoking in cases [$(p_1 \div 1\text{-}p_1) = (0.77 \div 0.23)$] and the odds of smoking in controls [$(p_2 \div 1\text{-}p_2) = (0.58 \div 0.42)$]. Dividing the odds of smoking in cases by the odds of smoking in controls yields the *odds ratio*, that is, 2.42. This odds ratio means that cases have 2.42 times the odds of smoking 10 or more cigarettes per day than controls. Still there is no apparent connection with risks of lung cancer in smokers and non-smokers.

Table 26 gives the formula proposed by Cornfield to compute the incidence rates based on the proportions of smokers, p_1 and p_2. For the purpose of the demonstration, Cornfield needed an additional piece of external information, that is, an "annual prevalence rate" of lung cancer in the population, which he estimated was 15.5 per 100,000 people per year. Cornfield was alluding to new cases of lung cancer diagnosed over a year and he must have meant annual *incidence* rate.

Table 26 shows that by some conjuring trick Cornfield had been able to transform proportions of smokers into incidence rates. This transformation allowed him to compute a relative risk from a case-control study. The trick depended on a simple

Table 26 – Formula used by Cornfield to transform smoking proportions (p1 and p2, see Table 25) into incidence rates. APR = "annual prevalence rate" of lung cancer in the population of = 15.5 /100,000. Source: (Cornfield, 1951).

Cigarette per day	Formula	Computations	Incidence rates (/100,000/yr)
10 or more	(p1 × APR) ÷ [p2 + APR × (p1 p2)]	(0.77 × 0.000155) ÷ [0.58 + 0.000155 × (0.77 0,58)]	20.6*
Else	(1-p1) × APR ÷ [(1-p2) – APR × (p1 p2)]	(0.23 × 0.000155) ÷ [0.42 – 0.000155 × (0.77 0.58)]	8.5**

Relative Risk = [incidence rate in '10 or more' ÷ incidence rate in 'else'] = [20.6 ÷ 8.5] = 2.40.
* 20.5 in Cornfield's paper.
** 8.6 in Cornfield's paper.

condition. If the proportion of the general population developing cancer of the lung, the "annual prevalence rate", is small relative to both p_2 and $1\text{-}p_2$, the contribution of the term APR × $(p_1\text{-}p_2)$ is trivial and can be neglected. In Table 26, this term is equal to [0.000155 × (0.77–0.58) =] 0.00003.

Table 26 also shows that once the APR × $(p_1\text{-}p_2)$ term is deleted, we are left with a formula for the relative risk, which is:

$$RR \cong [p_1 \times (APR \div p_2)] \div [(1\text{-}p_1) \times (APR \div (1\text{-}p_2))]$$

The equality is not exact because the term APR × $(p_1\text{-}p_2)$ was added to p_2 and subtracted from $1\text{-}p_2$. But we need four digits to show the inequality. The relative risk = 2.4240 before simplification and 2.4243 afterwards. We can again simplify APR from the numerator and the denominator of the new equation. We are left with the relative risk being almost equal to the odds ratio:

$$RR \cong [p_1 \div (1\text{-}p_1)] \div [p_2 \div (1\text{-}p_2)] = odds\ ratio = 2.42$$

The great news was therefore that knowledge of the population incidence rate (i.e., Cornfield's APR) was not needed to approximate the relative risk by the odds ratios. The connection between the odds ratio and the relative risk was now obvious and confirmed the intuition:

> *"...whenever a greater proportion of the diseased than of the control group possess a characteristic, the incidence of disease is always higher among those possessing the characteristic. This is the intuition on which the procedures used in*

such clinical studies [i.e., case-control studies] is based. Although it has frequently been questioned, it can now be seen as correct." (Cornfield, 1951, p. 1270).

In 1960 Cornfield and William M. Haenszel (1910–1998), biostatistician at the National Institutes of Health (1960, pp. 525–526) re-expressed the derivation of what we now call the approximation of the relative risk by the odds ratio under the rare disease assumption. The relation between the odds ratio and the relative risk had become the relation between cohort and case-control studies, which they still referred to with the old terminology of prospective (= cohort) and retrospective (= case-control) studies:

"Studies which start with populations grouped initially into subclasses, for each of which one counts the number of new cases of a disease which develop during some subsequent period of time, are ordinarily referred to as "prospective" or "population-based" studies. The annual incidence of most diseases is sufficiently small, so that prospective studies designed to supply estimates of the incidence rate for different classes of the population, or of their ratios, must cover large numbers of persons. Thus, in a prospective study of lung cancer in a population of 100,000 males over age 40, one might at the end of 1 year of study expect to find 50 to 75 new cases. This is a small return for a large effort. The "retrospective" or "case-control" study provides a more economical way of estimating the relative risk than the prospective method because it does not require devotion of a large part of the study resources to those who did not develop the disease. In such a study one identifies all, or a well-defined sample, of the new cases of a disease as they occur during some period of time, and only after the occurrence of the disease does one classify them by the presence or absence of the characteristic (hence the name "retrospective"). The remainder of the population, i.e., those who did not develop the disease during the period, is also sampled and similarly classified by presence or absence of the characteristic. Thus, a retrospective study of lung cancer of the same population of 100,000 males over age 40 would (in principle) uncover exactly the same 50 to 75 newly developed cases but would be free to study the characteristics of only a fraction of the remaining 99,925 to 99,950 males who did not develop lung cancer. Retrospective studies might on the surface appear to supply only estimates of the proportion of persons with and without the disease who possess the characteristic and not to estimate relative risk. Such an estimate can easily be derived, however." (Cornfield and Haenszel, 1960).

3.11. Refinements of the theory of case-control studies

Two factors have stimulated the refinement of the theory of case-control studies. First, before the 1980s, the computational problems associated with the analysis of

(moderately, e.g., n = 4,000) large cohort studies required computational alternatives that facilitated the sound treatment of the data. Nathan Mantel, statistician at the National Cancer Institute, proposed in 1973 that cohort studies could be analyzed as case-control studies without loss of validity in estimating the odds ratio:

> *"The prospective study can be converted into a synthetic retrospective study by selecting a random sample of the cases and a random sample of the non-cases, the sampling proportion being small for the non-cases, but essentially unity for the cases."* (Mantel, 1973).

Mantel called this new design "synthetic retrospective studies". We know it today as "nested case-control studies" (Doll, Part II). The computational burden is reduced by sampling a small fraction of the non-cases.

The second stimulus stemmed from the work of Cornfield (section 3.10) following which all case-control studies were now viewed as variants of cohort studies in which *cases* were a sample of all cases, and *controls* a sample of all the subjects who did not develop the disease during follow-up. In this context, did the way controls were sampled matter? In Cornfield's paper (Cornfield, 1951), controls were non-cases , that is, sampled among people who had not developed the disease at the end of the follow-up period. Miettinen (Miettinen, 1976a) had noted that waiting until all the cases had been recruited to sample the controls was an uncommon way of performing case-control studies of chronic diseases. Usually, the sampling of controls ran parallel to that of cases. Controls were free of disease at the time of their recruitment, but the investigator could not always rule out that they did not develop the disease later within the same risk period. There could therefore exist two different schemes of sampling controls. This distinction proved to be very fruitful for the evolution of the theory relating cohort to case-control study designs as well as odds ratios to relative risks or relative incidence rates.

3.11.1. Sampling schemes of controls

A series of papers in the seventies and eighties, including (Miettinen, 1976a; Greenland and Thomas, 1982; Hogue et al., 1983; Smith et al., 1984; Greenland et al., 1986) led to the conceptualization of three types of case-control studies according to the mode in which controls were sampled within the underlying cohorts.

To understand the theoretical reasoning we have to *imagine* that the cases and controls are sampled within fully enumerated cohorts of exposed and unexposed subjects, as if we were conducting *nested* case-control studies. If we define the "risk period" as the time interval during which cases are ascertained in the exposed and unexposed cohorts, controls could be sampled either: a) at the end of the risk period; b) from the population at risk during the risk period; or c) from the base. Building on Miettinen's work, Sander Greenland, from the Division of Epidemiology, UCLA School of Public Health and Duncan C. Thomas, then at the Department of Epi-

demiology and Health, McGill University in Canada, referred to these three sampling schemes as a) traditional b) incidence-density and c) case-base sampling (Greenland and Thomas, 1982; Greenland et al., 1986).

Figure 6 is an attempt to present graphically the differences between these three types of sampling schemes and modes of calculating odds ratios (Morabia et al., 1995).

a) In the *traditional* case-control study, controls are sampled among subjects remaining at risk at the end of the risk period. Thus, none of the controls has had the disease at some point during the risk period: controls are "non-cases" as in Cornfield's example. This is "cumulative incidence" sampling of controls (Greenland and Thomas, 1982). The traditional odds ratio is computed as:

$$Odds\ ratio_{traditional} = [cases_{exposed} \div cases_{unexposed}] \times [non\text{-}cases_{unexposed} \div non\text{-}cases_{exposed}]$$

Where "cases" and "non-cases" stand, respectively, for the number of cases and controls.

b) In the *incidence-density* case-control study, subjects in the population-at-risk are eligible as controls at multiple points in time within the risk period given that they are disease-free at the time of selection. However, they may also be sampled later as *cases* if they develop the disease. Controls are selected from all subjects still free of disease at the time of occurrence of the "index case", that is, the particular new case occurring at that time. Thus, the number of available controls for each index case is a function of the duration of follow-up. It is obtained by multiplying the number of subjects at risk times the average duration of follow-up (T), which is equivalent to computing person-times. Thus, it is as if controls were counted as person-times rather than individuals. This is "incidence density" sampling (Greenland and Thomas, 1982). The incidence density odds ratio is computed as:

$$Odds\ ratio_{incidence\ density} = (cases_{exposed} \div cases_{unexposed}) \times (person\text{-}times_{unexposed} \div person\text{-}times_{exposed})$$

The "person-times" are the total number of person-times accrued in, respectively, the exposed and the non-exposed subset of the cohorts included in the case-control study. The $OR_{incidence\ density}$ is, strictly speaking, the ratio of two incidence rates. The formula above can be re-written as:

$$Odds\ ratio_{incidence\ density} = (cases_{exposed} \div person\text{-}times_{exposed}) \div (cases_{unexposed} \div person\text{-}times_{unexposed})$$

$$= Incidence\ Rate_{exposed} \div Incidence\ Rate_{unexposed}$$

c) In the *case-base* (or case-cohort) study, controls are sampled from the baseline cohorts (i.e., the base), regardless of whether they become cases or not during the sub-

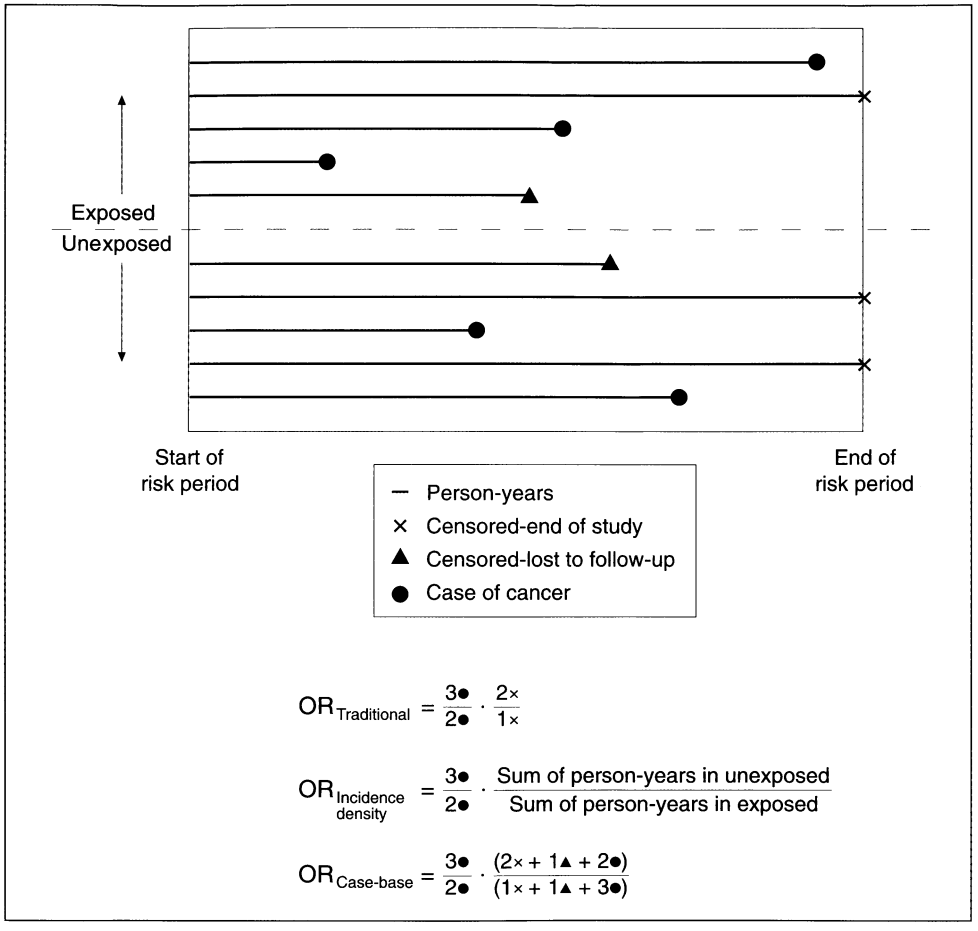


Figure 6
Methods for calculating traditional, incidence density and case-base odds ratios (adapted from Morabia et al., 1995).

sequent follow-up (Greenland et al., 1986). Thus, some individuals count both as cases and controls. The "case-base" odds ratio is computed as follows:

$$Odds\ ratio_{case\text{-}base} = (cases_{exposed} \div cases_{unexposed}) \times (cohort_{unexposed} \div cohort_{exposed})$$

The "cohorts" are the total number of people in the exposed and the non-exposed subset of the baseline cohorts, respectively, who were included in the case-control study. The $OR_{case\text{-}base}$ is strictly speaking a ratio of two risks:

$$Odds\ ratio_{case\text{-}base} = (cases_{exposed} \div cohort_{exposed}) \div (cases_{unexposed} \div cohort_{unexposed})$$
$$= Risk_{exposed} \div Risk_{unexposed}$$

The important consequence of this theory was that the relation of the odds ratio to measures of relative risks, either as ratio of incidence rates, or of risks, was *independent* of whether the studied disease was rare in the population. For example, no "rare-disease assumption" was needed to interpret the incidence-density odds ratio as the ratio of two incidence rates. The complete theory is more complex than its simplified version presented above and takes into account the stability of incidence and exposure during the risk period (Miettinen, 1976a; Greenland and Thomas, 1982; Greenland et al., 1986). I also chose one terminology for the measures described in this section, but the latter varies according to the authors (Rothman and Greenland, 1998).

At the end of this theoretical big bang, Cornfield's "rare disease assumption" had become a special case of a case-control study design. Nevertheless, when the disease is "rare" – that is, when the risk of disease is lower than about 10% over the risk period – the values of all risk ratios and odds ratios are very similar. In these situations, Jerome Cornfield's contribution is valid and the theory of control sampling schemes has less practical relevance.

3.11.2. Sampling controls independent of exposure

Kenneth Rothman has summarized the new concept of the case-control study resulting from the evolution described above in his 1986 textbook "*Modern Epidemiology*" (Rothman, 1986). Imagine a case-control study in which h cases are individuals who became ill during an average duration of time t, some being exposed (a) and other unexposed (b) to the studied cause. The controls are exposed (c) and unexposed (d) individuals, representing a proportion, k, of the combined exposed and unexposed cohorts that gave rise to the cases. The total number of exposed in the underlying cohorts is, therefore, c ÷ k for the exposed and d ÷ k for the unexposed. The total person-times in the underlying cohorts is (c ÷ k) × t or (c × t) ÷ k for the exposed and (d × t) ÷ k for the unexposed. Thus, the cohort incidence rates among exposed and unexposed could be estimated as

$$I_{exposed} = k\frac{a}{c \cdot t} \quad and \quad I_{unexposed} = k\frac{b}{d \cdot t}$$

The relative risk is the ratio of the incidence rates in the exposed over the incidence rate in the unexposed. In a case-control study, k, the sampling fraction for controls, is usually not known. We cannot therefore estimate the disease incidence rates based on the known a, b, c, d and t. However, both k and t are in principle similar for exposed and unexposed and can be cancelled when we compute the *ratio* of the incidence rates. The relative risk can therefore be obtained as:

$$RR = \frac{I_{exposed}}{I_{unexposed}} = \frac{a \cdot d}{b \cdot c}$$

In Rothman's words:

> *"Since the sampling fraction, k, is identical for both exposed and unexposed, it divides out, as does t. The resulting quantity, ad ÷ bc, is the exposure odds ratio (ratio of exposure odds among cases to exposure odds among controls), often referred to simply as the odds ratio. This cancellation of the sampling fraction for controls in the odds ratio thus provides an unbiased estimate of the incidence rate ratio from case-control data (Sheehe, 1962; Miettinen, 1976). The central condition for conducting valid case-control studies is that controls be selected independently of exposure status to guarantee that the sampling fraction can be removed from the status ratio calculation.*
> *The case-control design can be conceptualized as a follow-up design [follow-up = cohort] in which the person-time experience of the denominators of the incidence rates is sampled rather than measured outright. The sampling must be independent of exposure; by revealing the relative size of the person-time denominators for the exposed and unexposed incidence rates, the sampling process allows the calculation of the relative magnitude of incidence rates. Viewed in this way, the case-control study design can be considered a more efficient form of the follow-up study, in which the cases are the same as those that would be included in a follow-up study and the controls provide a fast and inexpensive means of inferring the distribution of person-time experience according to exposure in the population that gave rise to the cases."* (Rothman, 1986, pp. 63–64).

3.12. Evolution of group comparisons in epidemiology

When do we find the first group comparison? Is it already present in Graunt? Graunt compared the mortality of plague across calendar years. In a mysterious sentence, he explains that the proportion of plague deaths in 1625, that is, 68.4% ([35,417 ÷ 51,758], or about 7 to 10) was three times larger than the corresponding proportion in 1592, that is, 44.4% ([11,503 ÷ 325,886], or about 2 to 5):

> *"In the year 1625, we find the Plague to bear unto the whole in proportion as 35 to 51. or 7 to 10. that is almost the triplicate of the former proportion* [2 to 5 or 40% in 1592], *for the Cube of 7.being 343. and the Cube of 10. being 1000. the said 343. is not 2/5 of 1000."* (Graunt, 1662, p. 34).

Why does Graunt conclude that 70% can be the "triplicate" of 40%? The exact relative mortality is [68.4% ÷ 44.4% =] 1.6. The puzzling aspect is that the odds of deaths from plague in the year of 1625 ([0.684 ÷ 0.316 =] 2.16) is about three times larger than the odds of death in 1592 ([0.444 ÷ 0.556 =] 0.80): odds ratio = [2.16 ÷ 0.80 =] 2.7. Graunt's conclusion seems meaningless except if we use our modern

odds ratio. This important variation of plague mortality across times suggested to Graunt that the plague was more related to environmental than to human constitutional factors.

Around 1720, both John Arbuthnot (1665–1735), London physician, and James Jurin (1684–1750), physician and natural philosopher of Cambridge, tried to establish the mortality from natural smallpox and compare it with the mortality due to inoculation. They used the London Bills of mortality, other evidence when available, and a good deal of reasoning: Arbuthnot found 1 death out of 10 exposed to smallpox, and 1 death out of 100 inoculated. Jurin got 1 of 7 or 8, and 1 of 91, respectively (Rusnock, 2002). A fascinating example of population thinking and group comparison in the early 18th century.

3.12.1. Comparing like with like

The principle of comparing like with like already guides group comparisons in the 18th century. In his investigation of treatments for scurvy, the Scottish naval physician, James Lind (section 3.2) was very cautious to compare six pairs of seamen under similar conditions. He laid them together in one place and fed them with the same diet. One pair served as non-treated controls. Other examples from the 18th century can be found in the James Lind Library and in (Troehler, 2004).

About a century later, in his book about the effects of bleeding as a treatment of pneumonia and other illnesses, Pierre Louis (section 3.3.1) described the principle of valid comparisons:

> *"...what was to be done in order to know whether bloodletting had any favorable influence on pneumonitis, and the extent of that influence? Evidently to ascertain whether, other things being equal, the patients who were bled on the first, second, third or fourth day, recovered more readily than those bled at a later period. In the same manner it was necessary to estimate the influence of age, or any other circumstance, on the appreciable effects of bloodletting."* (Louis, 1836, p. 55).

For Louis, the experiment had to compare recoveries in groups of patients bled at different times. Louis was expecting that one group would recover "more readily" than the other. This was not an all-or-none response to treatment. The comparison had to be done "other things being equal". The potential influence of things that were not equally distributed had to be evaluated. In another section of the book, Louis mentioned diet before bleedings, age, severity of symptoms at the beginning of the disease and treatments other than bloodletting as

> *"causes which, independently of the period of the first bleeding, must have affected some difference in the mean duration of the disease"* (Louis, 1836, p. 6).

John Snow insisted in his description of the 1854 natural experiment that the clients of the different water companies were alike in many aspects (section 3.3.2). He was responding to the criteria that William Farr had expressed six months earlier, on November 19, 1853, in relation to Snow's hypothesis:

> *"To measure the effect of bad or good water supply, it is requisite to find two classes of inhabitants living at the same level [elevation], moving in equal space, enjoying an equal share of the means of subsistence, engaged in the same pursuits, but differing in this respect, – that one drinks water from Battersea [supposedly polluted water], the other from Kew .... But of such experimenta crucis the circumstances of London do not admit ...."* (cited by Vinten-Johansen et al., 2003, p. 260).

Basically, the proof required a study design that would minimize the confounding effects of those factors that were viewed as causes of cholera under different theories.

3.12.2. Fallacies resulting from group incomparability

The first half of the 20th century saw the first theories on potential fallacies that may have resulted from comparing incomparable groups. In 1903, Yule had published his "*Notes on the theory of association of attributes in statistics*" (Yule, 1903), which put in evidence, using a hypothetical example, "fallacies that may be caused by the mixing of records" (section 3.4.1). Yule's fallacy has been transmitted to us as Simpson's paradox and described the fundamental mechanism underlying what we now term "confounding".

In their 1916 investigation of the causes of pellagra, Goldberger and Sydenstricker used different forms of stratification and restriction in the data analysis to separate the effects of diet from those of income, age or gender and presented age-standardized risks (section 3.4.2).

In 1939, Wade Hampton Frost described a fallacy resulting from comparing the mortality from tuberculosis between people of different ages but born at different times (section 3.4.3). The mortality at different ages may in reality reflect different life exposures to the tuberculosis bacillus. We would say today that the effect of age on tuberculosis mortality was confounded by differences in exposure across cohorts.

3.12.3. Treatment allocation

The concern of comparing like with like rapidly led to the idea that allocating treatment could help. Louis had already written that:

> *"In any epidemics, for instance, let us suppose five hundred of the sick, taken indiscriminately, to be subjected to one kind of treatment, and five hundred others,*

> *taken in the same manner, to be treated in a different mode; if the mortality is greater among the first than among the second, must we not conclude that the treatment was less appropriate, or less efficacious in the first class than in the second? It is unavoidable; for among so large a collection, similarities of conditions will necessarily be met with, and all things being equal, the conclusion will be rigorous."* (Louis, 1836, p. 59).

The notion of taking the patients "indiscriminately", taking the two groups between which the treatment is compared "in the same manner" and the large sample size (1,000 patients being "such a large collection") indicate that Louis had a theory of group comparisons, and even of random allocation of treatment, whereby all other factors would have been distributed equally between the compared groups.

There is plenty of evidence that the use of alternate allocation to constitute comparable groups was a common idea by the end of the 19th century. In his now classic public controversy with the British bacteriologist Almroth Wright (1861–1947) about the efficacy of anti-typhoid inoculation, the statistician Karl Pearson proposed:

> *"only to inoculate every second volunteer. In this way spurious effect really resulting from a correlation between immunity and caution [to avoid exposure] would be got rid of"* (Pearson, 1904).

In 1898, the Danish Nobel laureate, Johannes Fibiger (1867–1928), published the apparently first clinical trial with alternate allocation of treatment (Hrobjartsson et al., 1998; Lilienfeld, 1982). In 1930, serotherapy was alternatively allocated to the patients of some the centers participating in the British Medical Research Council trial (section 3.4.4).

3.12.4. The name of the game

However, before 1945, it would be an anachronism to baptize the methods and concepts that were used with the names we use today. Take Louis's and Snow's analyses. Clearly, none of them consciously chose one study design or the other because its properties were more adapted to the questions they wanted to address. They could not rely on any existing theory of study designs. There were no epidemiology textbooks to which they could refer. Louis and Snow had to *invent* their way through the group comparisons. Indeed, there is little consensus among contemporary epidemiologists about how to categorize *a posteriori* Snow's 1854 "natural experiment". It has been viewed as a cohort study (Rothman, 1986; Sartwell, 1965), a survey (Doll, Part II), and a combination of ecologic and retrospective cohort studies (Winkelstein, Jr., 1995).

The same is true for Goldberger and Sydenstricker who performed exposed/unexposed (e.g., comparing incidence rates of pellagra in subgroups differing by diet or

income) and affected/unaffected (e.g., comparing dietary habits in subgroups differing by pellagra) comparisons (section 3.4.2). Their methodological contribution is mentioned both in histories of "cross sectional field surveys" (Susser, 1985), cohort studies (Liddell, 1988) and of case-control studies (Paneth et al., Part II).

It would be decades before "investigations" or "analyses" would become "studies" and the logic of exposed/non-exposed or affected/unaffected comparisons would become formal study designs, with their measures of effect, biases, and ability to disentangle the effects of multiple causes. We can only find the unspoken premises of these methods and concepts in these early works.

3.12.5. Case-control and cohort studies

The study of chronic traits, such as lung cancer and cardiovascular diseases, would have to address complex problems: risk factors (e.g., cigarette smoking) could not be randomly allocated, diseases lasted long, had multiple causes, and those exposed to one of the causes tended to be exposed to many others. Typically, smokers were more likely to drink, eat more meat and less fruits and vegetables, and engage less in physical activity. When studying the effect of any of these factors, it was important to treat the effects of the other factors appropriately. In this context, a theory of observational study designs comparing exposed/non exposed (i.e., cohort study) or affected/non affected (i.e., case-control study) became indispensable (sections 3.5 and 3.6).

The elements of theory accrued before 1945 were eventually fused into a theory of study designs after World War II. This process was driven by the quest for the causes of a huge epidemic of lung cancer among Western men that became recognized around 1950. The data showing that exposure to tobacco was the cause triggered an enormous controversy, which contributed importantly to the refinement and formalization of case-control studies, prospective studies and concepts such as confounding, interaction and bias.

The story begins more or less in 1940 (White, 1990). According to Ernst L. Wynder, the medical profession in the forties and fifties did not seriously think about smoking as a potential cause of major diseases (Wynder, 1997). In contrast, physicians interested in public health were astonished when, after World War II, vital statistics were showing a dramatic increase of lung cancer mortality in men. Around 1900, lung cancer was extremely rare (White, 1990, p. 30). Its incidence seemed to grow at a fast pace but the evidence did not convince everyone. It was argued that better diagnosis and aging of the population could explain the trends. An editorial in the British Medical Journal in 1942 stated:

> *"It is doubtful whether the higher incidence of cancer of the lung observed in recent years is real or only apparent."* (Editorial, 1942).

The Medical Research Council of Great Britain in 1950 still used the same expression as the British Medical Journal, that is,

> *"the increase [in lung cancer incidence] may, of course, be only apparent"* (cited by White, 1990).

The opinion had clearly changed in 1952, when the other major British medical journal, The Lancet, wrote:

> *"Few trends are more dramatic than the rise during the last 30 years in the notified death rates from cancer of the lung. There is little doubt that the increase is both real and numerically important."* (Editorial, 1952).

The population-based registries in Denmark and Connecticut reported marked increases in incidence in the forties and fifties. The Connecticut annual, age-adjusted incidence rates per hundred thousand were 9.7, 13, 20.6, 31.1 for 1935–39, 1940–44, 1945–49 and 1950–54 (White, 1990).

The smoking-lung cancer controversy epitomizes this new phase of methodological development, but the causes of many complex traits were discovered. In this process the theory of case-control studies (section 3.5), of cohort studies (section 3.6), concepts of confounding (section 3.4.5), bias (section 3.7), interaction (section 3.8) and causal inference (section 3.9) were further formalized. The relation of case-control to cohort studies was understood (sections 3.10 and 3.11).

Finally, the demonstration that a case-control study could be viewed as a way of sampling subjects within cohorts has unified concepts across study designs (section 3.10). It was also understood that the most usual way of sampling controls, that is, concurrently to case occurrence, yielded the relative incidence rates, without the rare-disease assumption (section 3.11.1). This led to the confinement of the need for the rare disease assumption to relatively uncommon ways of sampling controls (section 3.11.2).

There is a meaningful aspect of the smoking-lung cancer controversy for the history of epidemiologic methods and concepts: the arguments that were used to oppose the smoking-lung cancer connection finally contributed to strengthening epidemiologic methods. In trying to demonstrate that lung cancer was *not* related to smoking but to some genetic factor, the statistician of the University of Cambridge, Fisher contributed to the formalization of the concept of confounding (section 3.4.5). In his criticism of hospital-based case-control studies, another statistician of the Mayo Clinic, Joseph Berkson, laid the foundations for a theory of selection bias (section 3.7). Instead of derailing epidemiologists, these criticisms proved useful and were further elaborated and integrated into the emerging discipline by several epidemiologists.

4. Epistemology

4.1. Tribute to Piaget

The work of the epistemologist Jean Piaget (1896–1980) has inspired me to present the genesis of epidemiology as an evolving process from very intuitive to more theoretical and abstract concepts (Piaget, 1970). During the last phase of his career (1940 to 1971), Piaget was Professor of experimental psychology at the University of Geneva, Switzerland. He had created in 1955 and directed the International Centre of Genetic Epistemology. His description of the genesis of scientific disciplines offered an attractive model for explaining the development of epidemiologic methods, a model that fitted well my perception of the evolution of epidemiologic principles, population thinking, and group comparisons reviewed in this essay.

Before I present *my* understanding of Piaget's genetic epistemology, I want to stress that I do not pretend to be a Piagetian. I do not know if this essay reflects his views, first because he never dealt with epidemiology but, even more importantly, because of his intellectual style. Piaget was a fascinating thinker. He wrote thousands of pages of epistemology, which read as a continuous flow of ideas. His thinking was in perpetual construction (Piaget, 1967). Piaget constantly polished ideas, and debated against other schools of thought. But unlike manuals, his books neither really start nor end. Indeed, one can hardly find a synthesis in Piaget's writing. Syntheses written by his students and scholars are often less accessible than Piaget's original contributions. Of course, these scholars probably deeply disagree with what I just wrote.

Thus, I am hesitant to relate this essay to Piaget's ideas. But at the same time I want to acknowledge that his writings inspired me. In Piaget's terms:

> *"Genetic epistemology attempts to explain knowledge, and in particular scientific knowledge, on the basis of its history, its sociogenesis, and especially the psychological origins of the notions and operations upon which it is based."* (Piaget, 1970, p. 1).

There are in my view two key elements in Piaget's epistemology. The first is that humans actively gather knowledge. Human knowledge is derived from actions:

> *"I think that human knowledge is essentially active. To know is to assimilate reality into systems of transformations. To know is to transform reality in order to understand how a certain state is brought about."* (Piaget, 1970, p. 15).

This may seem self-evident to most readers of this book, but Piaget was among the first to express it in a qualified way. The world does not reveal its truth passively. It resists and we must therefore act upon it and learn from these actions. We assimilate

reality by developing systems that transform it in order to reveal how certain states are produced. We can learn *by acting* on a physical object using *simple* actions, such as throwing, pushing, touching. For example, we can lift different objects and realize that they have different weights. This is how sciences like physics accumulate knowledge, but what about mathematics? In abstract sciences, knowledge is derived from *coordinated* actions and it is the coordination of actions rather than the transformation of reality that generates knowledge. For example, I can count ten lined pebbles from left to right and then from right to left and find out that their sum is independent of their order. This concept is known in mathematics as commutativity. It was not acquired by changes in the pebbles but by the action of counting them in different orders. Actions can be concrete or abstract.

For Piaget, thoughts being invariably related to actions, they need to evolve. Knowledge consists of established causal relations, which he refers to as laws, "modes of production", explanations. Identified causal relations open the way to more action, and therefore more and increasingly elaborated knowledge.

This leads to the second key element in Piaget's idea: scientific knowledge and discipline are in perpetual evolution. Science is a process. It is in continual construction and organization. Other epistemologists may have defended similar ideas, but Piaget's thinking is characterized by the importance of psychological and sociological factors in this construction process. Piaget postulates that there is a parallelism between the progress made in the logical and rational organization of scientific knowledge and the development of human psychology during an individual's life.

> *"The fundamental hypothesis of genetic epistemology is that there is a parallelism between the progress made in the logical and rational organization of knowledge and the corresponding formative psychological processes. Well, now, if that is our hypothesis, what will be our field of study? Of course, the most fruitful, most obvious field of study would be reconstituting human history – the history of human thinking in prehistoric man. Unfortunately, we are not very well informed about the psychology of Neanderthal man or about the psychology of Homo siniensis of Teilhard de Chardin* [1881–1955, paleontologist]. *Since this field of biogenesis is not available to us, we shall do as biologists do and turn to ontogenesis. Nothing could be more accessible to study than the ontogenesis of these notions. There are children all around us. It is with children that we have the best chance of studying the development of logical knowledge, mathematical knowledge, physical knowledge, and so forth."* (Piaget, 1970, pp. 13–14).

As a psychologist, Piaget studied the development of intelligence in children and observed that it is a progressive but structured process that starts with the acquisition of very simple skills, which become in turn the bases for acquisition of more complex ones. Steps are added to the ladder and each step up offers a wider perspective on the

world. In Piaget's view, the genesis of a scientific discipline follows an analogous process. Scientific disciplines evolve from very intuitive concepts based on primitive notions to always more abstract and formalized concepts, which are intellectual constructions, made possible by the previous steps. Each level of formalization is a precondition for reaching higher levels because simple theories become tools that allow us to construct theories that are more complex. Without the simple theories, we cannot achieve the more complex and abstract ones. In the process of transforming intuitive or naive notions into universal concepts and theories, scientific disciplines progressively became more abstract, theoretical and mathematical.

4.2. Evolution of physics

Before applying this interpretation scheme to epidemiology, let us see how it applies to physics, as this is the science on which common epistemological models are based. Physics is a very old science that can be viewed as an extension of the sensory and muscular systems of the human body (vision, muscle power, audition, touch, temperature, etc). Physics is related to the essential activities of social life: creation of utensils (flint) or weapons, control of fire, wind (navigation) and water, etc. (Bernal, 1972 chapter 2). Every craftsman develops an intuitive (empirical) knowledge. Seamen have their own theory about sailing, the pitcher about throwing the ball, the cook about the use of the fireplace. Their mode of acquiring knowledge is from a psychological perspective a very primitive one, based on action and reaction, trials and errors. Knowledge comes from repeating the same action and eventually modifying it until one gets what is expected or discovers something new. The first physical experiments were very intuitive. The law of buoyancy of the Greek mathematician Archimedes (BC 287-BC 212) is taught in high school. It is quite intuitive to understand that

> *1. A completely submerged body displaces a volume of liquid equal to its own volume. The buoyant force equals the weight of the fluid displaced.*
> *2. When an object weighs less than the total volume of fluid it can displace, it will settle down until the buoyant force equals the weight and it floats partially submerged.*

The law is easy to remember and the anecdote of Archimedes exclaiming Eureka! (which means: "I found" in Greek) while immersing himself in his bathtub belongs to popular wisdom. But intuition is insufficient to understand more subtle physical phenomena. The Greek philosopher Aristotle (BC 384-BC 322), founder of the Lyceum of Athens, stated, for example, a law of motion according to which

> *"the moving body comes to a standstill when the force which pushes it along can no longer so act as to push it"* (Einstein and Imfeld, 1966, p. 6).

This "law" reflects our intuitive perception of movement (in the presence of friction and air resistance) but it is false. You can compare it with the Newtonian definition of a force given above (section 2.4.1). Eventually, repeated and organized actions on nature reveal invariant phenomena, observations and laws. In the 16th century, mathematics would become the indispensable complement of experiments in physics. The Italian scientist Galileo Galilei (1564– 1642) claimed that he repeated his experiments a hundred times and always observed the *exact* same results (Bernal, 1972, p. 27). Galileo apparently ignored measurement errors but his experiments allowed him to state a law according to which falling bodies of different weights and sizes took the same time to reach the ground. Galileo's discovery of gravity required a more elaborated mode of reasoning to obtain a result that was, and still is, counter-intuitive (intuitively, most of us expect bodies of different weight to accelerate differently in their fall). After Galileo, the process of formalization went on. Newton built on Galileo's work and predicted the behavior of even less intuitive phenomena (planets), and needed advanced algebra to understand the nature of forces.

Physics finally moved to a form of population thinking, but in terms of particles. Classical physics defined the position and velocity of a single particle (or of one planet). The world of mechanical physics was three-dimensional. It was made of particles whose interactions were governed by a specific law depending on distance and fields. But this mechanical view did not explain why matter appeared to have a granular structure. Hence,

> *"quantum physics formulates laws governing crowds and not individuals ..."* (Einstein and Imfeld, 1966, p. 297).

Quantum physics defined the probability of finding one particle of a certain velocity and a certain position, based on many observations. Thus, physics became statistical in the 20th century. Ultimately, from the 20th century on, physics became so abstract and theoretical that it went beyond intuition and only "geniuses" such as the mathematicians and physicists Albert Einstein (1879–1955) or Niels Bohr (1885–1962) could carry it on to new levels of knowledge.

Epidemiology, at least when defined as a set of methods and concepts, is a much younger discipline than physics. Physics has existed for more than 2500 years considering that its first levels of formalization occurred in ancient Greece. Or 400 years considering, as Einstein did, that scientific reasoning in physics began with Galileo. When did epidemiology first appear?

4.3. Was Hippocrates an epidemiologist?

In this review of the scientific work that has been referred to at one moment or another as "epidemiologic", I systematically searched for the simultaneous presence of

population thinking and group comparisons. I started, of course, with the texts of Hippocrates (BC 460-BC 377), who is described in many epidemiology texts as the "father of epidemiology" (MacMahon et al., 1960; Lilienfeld and Lilienfeld, 1980; Pan American Health Organization, 1988, p. 3). These texts do not mention, however, who the "mother" was!

Hippocrates, we believe, was an independent and ambulatory physician, born on the Island of Cos, between current Greece and Turkey. He and others after him described their activity and thinking in medical texts that occupy an undoubtedly important place in medicine. At the time when most medicine, treatment, and cures relied on magical or divine phenomena, the Hippocratic texts used rational thinking, attributing diseases to environmental or other natural causes and proposing empirical treatments such as surgery, diet, herbal remedies, etc. They did not consider divine or magical causes in the etiology or treatment of diseases. Causes were to be found in nature. The quality of the description of diseases and symptoms in Hippocratic texts may explain their influence in the centuries that followed. They expressed 2,500 years ago the kind of materialism that still drives Western medicine today. Hippocratic theories are often easier to understand for a modern reader than the theoretical constructs of physicians who followed him. For example, the theory of reproduction based on the mixing/blending of male and female seminal fluids remain a perfectly satisfactory explanation for what most people observe with their own eyes, much more so than the homunculus theory of Aristotle (Sykes, 2002, p. 41).

A remarkable feature of Hippocratic thinking, which struck those defending the cause of public health in the 19th century and later, is its appraisal of environmental and lifestyle factors as health determinants. In his book "*On Airs, Waters and Places*" (Hippocrates, 400a BCE), we read that the traveling physician arriving to a foreign place had to examine its geographical position, winds, sun, quality of water and yearlong climatic variation:

> "*Whoever wishes to investigate medicine properly, should proceed thus: in the first place to consider the seasons of the year, and what effects each of them produces for they are not at all alike, but differ much from themselves in regard to their changes. Then the winds, the hot and the cold, especially such as are common to all countries, and then such as are peculiar to each locality. We must also consider the qualities of the waters, for as they differ from one another in taste and weight, so also do they differ much in their qualities. In the same manner, when one comes into a city to which he is a stranger, he ought to consider its situation, how it lies as to the winds and the rising of the sun; for its influence is not the same whether it lies to the north or the south, to the rising or to the setting sun.*" (Hippocrates, 400a BCE).

The second part of "*On Airs, Water and Places*" is less often quoted but reveals the speculative side of Hippocrates's thinking:

> *"The other races in Europe differ from one another, both as to stature and shape, owing to the changes of the seasons, which are very great and frequent, and because the heat is strong, the winters severe, and there are frequent rains, and again protracted droughts, and winds, from which many and diversified changes are induced."* (Hippocrates, 400a BCE, Part 23).

Hippocratic texts indicate that there was a time when physicians included environmental factors in their diagnostic approach. The role of the environment may have been downplayed later until rediscovered in the 19th century. Hence the fascination of public health practitioners and early epidemiologists towards this major figure of antiquity who seemed to have shared their vision of the role of environment in disease causation. However, the gap between the Hippocratic treatises and modern preventive medicine has lasted so many centuries that it is not clear to me whether Hippocrates can be viewed as a pioneer of modern medicine.

But was Hippocrates an epidemiologist in the sense that he combined population thinking and group comparisons? Can we really trace the roots of epidemiology in antiquity? Here is Major Greenwood's opinion:

> *"Although Hippocrates was before all else a clinician, he was also a student of preventive medicine and epidemiology, of the doctrine of disease as a mass phenomenon, the units not individuals but groups."* (Greenwood, 1935, p. 18).

Is it true that the Hippocratic texts considered diseases as mass phenomena? In "*On Airs, Waters and Places*" we find the distinction between "endemic" diseases, that are always present in a population and "epidemic" diseases, which can become excessively frequent and then disappear. Moreover, the following description of an epidemic (probably of mumps) shows that there is some qualitative description of the frequency of symptoms (e.g., "in many", "in all cases") and their distributions in the population (e.g., in children and adults but seldom attacked women):

> *"1. In Thasus, about the autumn equinox, and under the Pleiades, the rains were abundant, constant, and soft, with southerly winds; the winter southerly, the northerly winds faint, droughts; on the whole, the winter having the character of spring. The spring was southerly, cool, rains small in quantity. Summer, for the most part, cloudy, no rain, the Etesian winds, rare and small, blew in an irregular manner. The whole constitution of the season being thus inclined to the southerly, and with droughts early in the spring, from the preceding opposite and northerly state, ardent fevers occurred in a few instances, and these very mild, being rarely attended with hemorrhage, and never proving fatal. Swellings appeared about the ears, in many on either side, and in the greatest number on both sides, being unaccompanied by fever so as not to confine the patient to bed; in all cases they disappeared without giving trouble, neither did any of them come to suppuration, as is common in swellings*

from other causes. They were of a lax, large, diffused character, without inflammation or pain, and they went away without any critical sign. They seized children, adults, and mostly those who were engaged in the exercises of the palestra and gymnasium, but seldom attacked women. Many had dry coughs without expectoration, and accompanied with hoarseness of voice. In some instances earlier, and in others later, inflammations with pain seized sometimes one of the testicles, and sometimes both; some of these cases were accompanied with fever and some not; the greater part of these were attended with much suffering. In other respects they were free of disease, so as not to require medical assistance." (Hippocrates, 400c BCE).

But the purpose of these concepts was to help clinicians better understand the reasons why individuals (their patients or clients) in some populations were more likely to be affected by some diseases than others:

"It appears to me a most excellent thing for the physician to cultivate Prognosis; for by foreseeing and foretelling, in the presence of the sick, the present, the past, and the future, and explaining the omissions which patients have been guilty of, he will be the more readily believed to be acquainted with the circumstances of the sick; so that men will have confidence to entrust themselves to such a physician. (...) Thus a man will be the more esteemed to be a good physician, for he will be the better able to treat those aright who can be saved, having long anticipated everything; and by seeing and announcing beforehand those who will live and those who will die, he will thus escape censure." (Hippocrates, 400b, BCE).

To the best of my knowledge, Hippocratic texts do not use the group as a unit of thinking. They describe patients one at a time and do not derive knowledge from looking at aggregated cases. There is no formal attempt to group the symptoms under the same disease entity or suggest that they occur in a well-defined combination.

The Hippocratic central preoccupation is to predict what will happen to an individual patient, a question that lies at the heart of medicine. And clearly, the environment was, in Hippocratic texts, an important predictor. But there are no traces of formal population thinking and practically no simple, controlled observations in Hippocrates's treatises. It is mostly, from the 17th century on, when population thinking became philosophically and mathematically founded, that disease entities started to be defined by a set of common symptoms in a population of patients. Before that, there could be no epidemiology.

4.4. Traces of epidemiology in the Bible?

This last statement seems to be contradicted by an example of a supposed group comparison reported in the Book of Daniel, which belongs to the Old Testament of

the Bible. The Book of Daniel probably reflects attitudes from the 2nd – 1st centuries before our era (Weingarten, 2004).The episode belongs to the attempt of the King of Babylon to familiarize a group of noble Israelite prisoners, including Daniel, captured after the fall of Jerusalem with the customs of the Chaldeans. The text reports the following episode:

> *"Then Daniel said to the guard whom the master of the eunuchs had put in charge of Hananiah, Mishael and Azariah and himself 'Submit us to this test for ten days. Give us only vegetables to eat and water to drink; then compare our looks with those of the young men who have lived on the food assigned by the king and be guided in your treatment of us by what you see.' The guard listened to what they said and tested them for ten days. At the end of ten days they looked healthier and were better nourished than all the young men who had lived on the food assigned them by the king. So the guard took away the assignment of food and the wine they were to drink and gave them only the vegetables."* (Weingarten, 2004).

Apparently, this story suggests that a controlled experiment took place. Hananiah, Mishael, Azariah and Daniel received a vegetarian diet and water and their looks were compared after 10 days to those of "all the young men" who ate the meat and wine assigned by the king. This example is exceptional in many aspects. The principle of comparing two groups to assess the benefit of some diet appears as a crafty tactic of the prisoners to keep their dietary practice. The comparison must have appealed to the king's sense of logic, which may not have been culture-specific. The results of the experiment were absolutely convincing, almost miraculous: the *four* Jewish men looked healthier and were better nourished than *all* the young men who had lived on the food assigned by the king. But here stops the analogy with group comparisons as we mean it today. It did not come to anybody's mind that the better look of Daniel and his friends could be attributable to something else than their diet. Daniel did not expect his friends and him to look *on average* healthier and better nourished than the other young men. There is no population thinking in Daniel's ruse: each of the Jewish men looked healthier than each of the king's men. The tale is therefore not eligible as a first epidemiologic trial. Epidemiologic group comparisons go along with population thinking, and population thinking did not exist before the 17th century. Such an innovation could not have skipped centuries.

4.5. The impossible comparison

The propensity to trust non-controlled observations is a striking feature of human populations. Consider a patient complaining about flu-like symptoms, who is given antibiotics and feels rapidly better thereafter. The improvement will be attributed to the antibiotics. Similarly, the person who drinks herbal tea after each meal and does

not catch the flu for the whole winter will tend to causally relate the herbal tea and his/her resistance to the flu. These are examples of "*post hoc, ergo propter hoc*" ("after it, therefore because of it") reasoning. Conclusions may have been radically different if controlled observations were available, that is, had there been an instance to compare what happened with the antibiotics or with the herbal tea to what would have happened without them.

It seems impossible that brilliant thinkers and clinicians such as Hippocrates, Galen, Thomas Sydenham (1624–1689), also called the "English Hippocrates", Jean-Nicholas Corvisart (1722–1809), etc., in addition to generations of shamans and other primitive therapists had not reflected about this issue. There must be some deep, essential reason for which no therapist integrated controls in their approaches. A simple explanation is that a controlled observation with oneself is *logically impossible*. Once the antibiotic has been prescribed and eventually taken, the situation of the patient has irreversibly changed. We cannot go back in time to the situation where the patient was suffering from flu-like symptoms and had not been treated yet. There is a logical impossibility to get the "counter fact" that would be needed to perform a perfect controlled observation.

The principle of a controlled experiment is therefore logically impossible when we are dealing with an individual human, and more generally with an individual living, complex organisms. Once an action has occurred, we cannot go back to square one and act as if that action had never occurred. Both the subject (e.g., the clinician) and the object (e.g., the disease of the patient) of the action have been modified by the action itself.

To make a cautious step in the direction of Piaget, we can ask whether this logical impossibility of the counterfactual action explains why children do not develop an intuition for controlled experiment. Children develop their psychology by experimenting with the world around them. They compare all the times. They compare what they expect to what they observe. They do this repeatedly. They learn by repeated trials and errors, but they *cannot* compare what happened after their specific action to what would have happened if they had not acted like that.

We are therefore facing a dilemma: there is no scientific knowledge without comparison or controlled experiment, but comparing or controlling medical intervention on a specific patient is impossible. It is actually more than a dilemma. Physicians develop the art of predicting outcomes in individuals and are reluctant to see any relationship between their art and the techniques of mass or crowd prediction.

4.6. Why did epidemiology appear so late in human history?

The logical impossibility of experiments in which the same individual serves *simultaneously* as her own control can only be overcome if the problem is posed at the population level, in probabilistic terms. While individuals are unique, unpredictable and incomparable, the average behavior of groups is predictable and comparable.

Often paradoxes have a solution only if we radically change our perspective on the problem.

Consider one of Zeno's paradoxes. Anyone who wants to move from one point to another (say, 100 meters) must first reach half the distance (i.e., 50 m), and thereafter half of half distance (i.e., 25 m), etc. Since space is infinitely divisible, one has to reach an infinite number of mid-distances in a finite time. This being impossible, we cannot go anywhere and motion is illusory. This seems logically correct but intuitively absurd. To formally perceive the logical error, we have to change perspective. We stop viewing ourselves as being unable to make a first step across the first mid-distance of our journey. We consider each mid-distance as belonging to a geometric series (e.g., 1 + 1/2 + 1/4 + 1/8 + 1/16 + 1/32 + ...), which luckily for us does converge when the multiplicative factor (in our example, 1/2) is less than one. Thus, 1 + 1/2 + 1/4 + 1/8 + 1/16 + 1/32 + ... is equal to 2. Here we go.

Similarly, the controlled observation has no solution at the individual level but, paradoxically, it has one at the population level. As Sherlock Holmes told Dr Watson who was once more amazed by the sagacity of his friend:

> *"Winwood Reade [novelist, William Winwood Reade, 1838–1875] is good upon the subject," said Holmes. "He remarks that, while the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant. So says the statistician."* (Doyle, 1890a, chapter 10).

Population thinking implies that one can establish what would happen on average in the presence (or the absence) of the cause, and use this as the *best guess* to make predictions at the *individual* level. A controlled experiment is possible, but only at the population level.

Glimmerings of population thinking in epidemiology appear in the work of William Farr 175 years ago. In *On Prognosis*, which is reproduced in this volume (Farr, Part II), Farr begins by discussing the Greek etymology of the word "prognosis":

> *"Fore-telling presupposes fore-knowledge; and prognosis is employed, in medicine, to designate the art of fore-seeing and foretelling the course and issue of diseases."* (Farr, Part II).

He then explains how the probabilities have different interpretations when applied to populations or to predict the occurrence of an event in an individual "case":

> *"In prognosis patients may be considered in two lights: in collective masses, when general results can be predicted with certainty; or separately, when the question be-*

comes one of probability. If 7,000 of 10,000 cases of fever terminate fatally, it may be predicted that the same proportion will die in another series of cases; and experience has proved that the prediction will be verified, or so nearly verified as to leave no room for cavil or skepticism. The recovery or death of one of the cases is a mere matter of probability. (...). The rate of mortality determined for 10,000 cases applies, as a general standard, to each patient; and the probability of death is 0.07, of recovery is 0.93; the probability that the fever patient will recover is 93 to 7, raised or lowered by particular circumstances." (Farr, Part II).

The concepts expressed in this paragraph are radically different from those found in Hippocrates's treatises. Farr was much more clearly a population thinker. Population thinking allows Farr to make an observation which would have certainly fascinated Hippocrates, and which is probably valid for medicine at large:

"It is, nevertheless, rare that the physician has to perform this mournful function, and to prophesy death. There is almost always a chance, and generally a strong probability of recovery. Nine times in ten he is the messenger of glad tiding; and it is seldom that he cannot point out some dawn of hope – some streak of light – when the horizon is darkest." (Farr, Part II).

In other words, around 90% of the patients will live regardless of the physician's intervention. Physicians should first avoid aggravating the death risk by their intervention.

A radical change occurred between antiquity and the 19th century, between Hippocrates and Farr. Population thinking emerged as a mode of conceptualization, observing, and approaching problems. It made the development of group comparisons as a methodological tool possible. Group comparisons could from then on belong to a formal scientific activity, because probabilistic statements and probabilities had become part of "high sciences" thinking. Epidemiology came late in human history because it had to wait for the emergence of probability.

4.7. Emergence of probability

According to the Professor of History of Philosophy Ian Hacking,

"the decade around 1660 is the birth time of probability" (Hacking, 1975, p. 11).

The reasons for the late emergence of probabilistic thinking in philosophy and mathematics are not clear. A vulgar version of probability had been present for hundreds of years in "low sciences" such as alchemy, astrology, geology, or in games. The first textbook of probability was written by Huygens in 1657 (Hacking, 1975). It is of major interest for epidemiology that the application of probabilities to human health-re-

lated issues also occurs during this decade, around 1660. Graunt's opus "*Natural and Political Observations upon the Bills of Mortality*" appeared in 1662.

Hacking notes that there are immediately two usages of probability: a) for producing frequencies that have "law-like" regularity on the basis of statistical data, b) for assessing reasonable degrees of belief in propositions, even if they are guesses not based on statistical evidence (Hacking, 1975, p. 44). The second point would represent a major revolution in science because it opened the way to the existence of scientific knowledge that was not necessarily "demonstrated" by irrefutable and reproducible experiments. Before probability, and more specifically according to Hacking, before the publication in 1739 of the English philosopher David Hume (1711–1776) "*A Treatise of Human Nature*" (Hume, 1739), knowledge was the privilege of the sciences such as mechanics or optics, which could demonstrate the existence of natural laws. Sciences that could not achieve demonstrations could not produce knowledge either, only opinions.

Hume explained that past observations did not necessarily determine what would happen in the future. The fact that by custom and habit we come to associate two qualities does not legitimize the belief that the association will hold in the future. Bread nourishes me but this is no demonstration that the next piece of bread will also prove nourishing. Nevertheless, most of us would bet that it would still be nourishing. Why? Because we do believe that we can generalize on the basis of what we have repeatedly observed in the past and make reasonable predictions. The work of Hume has lent support to the view that in addition to the knowledge of *what has been demonstrated* there is a knowledge of *what is probable*, based on sound generalizations (Hacking, 1975, p. 176 and 183).

Both aspects of probability were to be used in epidemiology. Population thinking in epidemiology may correspond to the statistical usage of probability described by Hacking. We have seen that the word statistics itself means the systematic collection of data about the state. The City of London began in 1603, a bad year of plague, to keep a weekly tally of births and deaths. This activity provided Graunt with "statistical", population data, which he used as evidence to compute the frequencies of different causes of deaths.

4.8. A theory of group comparison

John Stuart Mill (1806–1873) is one of the most famous British philosophers of the 19th century. He wrote extensively on epistemology. In "*A System of Logic*", first published in 1843, he described four methods (which he calls "canons") of experimental enquiry: the canons of agreement, difference, residues, and concomitant variations, which are all based on the principle of comparison.

The method of difference, Mill's second canon, states the fundamental principle of group comparisons:

> *"In an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon"*. (Mill, 1950, pp. 215–216).

This is typically the rationale for epidemiologic designs, either cohort or case-control studies, aiming to compare like with like. In epidemiology, however, the method has to be reformulated in probabilistic terms.

The method of agreement, Mill's first canon, is the counterpart of the method of difference. It searches for "the only one circumstance in common". The method of residues has become Holmes's maxim:

> *"Eliminate all other factors, and the one which remains must be the truth."* (Doyle, 1890b).

The method of concomitant variation is needed when the objects of the experiment cannot be manipulated in full, such as the moon or the earth. We cannot remove the moon, but we can correlate the position of the moon with water levels.

There is an aspect in Mill's canons that is particularly appealing to epidemiologists: the causes that the methods of difference, agreement, etc., contribute to discover are single invariable antecedents of the studied phenomenon. This is the type of cause that epidemiologists usually seek: a single preventable risk factor, such as smoking, polluted water, alcohol, saturated fat, physical activity, etc. (Vineis, Part IIb). It is therefore not surprising that Mill's canons of causality are often referenced in the writing of epidemiologists (Susser, 1973; MacMahon et al., 1960). Mervyn Susser, former director of the Department of epidemiology at Columbia University School of Public Health (Susser, 1973) has cited and illustrated the canons with modern examples. Rothman's theory of sufficient causes comprising multiple component causes (Section 2.6) evokes Mill's concept of a cause (Rothman, 1976):

> *"The cause, then, philosophically speaking, is the sum total of the conditions, positive and negative taken together; the whole of the contingencies of every description, which being realized, the consequent invariably follows..."* (Mill, 1950, pp. 197–198).

In summarizing the modes of coordinating actions that can contribute to the discovery of causes, Mill has established a theory of comparison. The examples he gives indicate that these methods apply primarily to experimental sciences: "the planetary path" as an example of the method of agreement, the law of "falling bodies" as an example of both the methods of agreement and difference, the "cosmical motions" as an example of the methods of agreement and of concomitant variations (Mill,

1950, p. 236). But for Mill these methods could be successfully applied in the social sciences, where by definition population thinking is the rule. Thus, Mill can also be viewed as having formalized a theory of *group* comparisons. From Mill on, group comparisons combined with population thinking became a philosophically valid principle of knowledge acquisition.

4.9. Causal inference

The second aspect of probability identified by Hacking, that is, its usage to generate *degrees of belief* associated with specific statements not (directly) based on data were to permeate epidemiology much later. This apparently occurred as part of the tobacco-lung cancer controversy, when the question of the causal nature of the association was posed. A theory of causal inference would then be built in successive steps.

In 1959, Yerushalmy and Palmer, from the Division of Biostatistics of the University of California at Berkeley and the Division of Special Health Services in Washington, published a paper entitled "*On the methodology of investigation of etiologic factors in chronic diseases*" (Yerushalmy and Palmer, 1959). They first summarized the criteria of causal inference proposed in textbooks of bacteriology, putatively attributed to the German bacteriologist Robert Koch (1843–1910) and known as "Koch's postulates". Causality required

> *"A. The simultaneous presence of organism and disease and their appearance in the correct sequence, and*
> *B. The specificity of effect of the organism on the development of the disease."* (Yerushalmy and Palmer, 1959, p. 31).

These two types of evidence were incompatible with the multiplicity of causes for chronic diseases. Yerushalmy and Palmer therefore restated Koch's postulates in terms of population thinking and group comparisons:

> *"1. The suspected characteristic must be found more frequently [population thinking] in persons with the disease in question than in persons without the disease [group comparisons], or*
> *2. Persons possessing the characteristic must develop the disease more frequently [population thinking] than do persons not possessing the characteristic [group comparisons]."* (Yerushalmy and Palmer, 1959, p. 32).

But that did not suffice. For example, concluding that smoking was a cause of lung cancer could not be based strictly on data. Yerushalmy and Palmer (Yerushalmy and Palmer, 1959) contributed, along with many epidemiologists, to a lively debate. Austin Bradford Hill, the British epidemiologist and perhaps the most creative

methodologist of the 20th century, authored the consensual paper. The question addressed by Hill was:

> *"In what circumstances can we pass from this observed association to a verdict of causation? Upon what basis should we proceed to do so?"* (Hill, 1965, p. 295).

Epidemiologic studies were what they were and no more. Their results could be indicative of statistical associations but it became clear that causal statements had to be based on a synthesis of all available information, epidemiologic, biological, toxicological, etc. Hill described nine "aspects" or "viewpoints" that could help the causal inference process by precisely increasing or decreasing our degree of belief in the causal statement (Section 3.14).

Hill, in contrast to Yerushalmy and Palmer, did not mention some philosophical or scientific origin to the causal viewpoints. There are, however, striking similarities between them and the "rules by which to judge of causes and effects" (Hume, 1739, pp. 173–6) given by David Hume in his 1739 "*A Treatise of Human Nature*" (Morabia, 1991).

Table 27 matches Hill's and Hume's aspects of causality as they are expressed in their two publications (Hume, 1739; Hill, 1965). Because Hume's treatise was published 225 years before Hill's report, it is obviously impossible to get a perfect match. For example, Hume presented his rules as universal statements, whereas Hill's viewpoints are worded specifically for preventive medicine. The comparison suggests nevertheless a potential philosophical kinship between Hume and Hill.

The identity is striking for the aspects of causal relations Hill has identified as "temporality," "biologic gradient" and "consistency". Strength of the association, as a measure of relative effect, does not have an exact complement in Hume's rules. Nevertheless, Hume's constant-conjunction formula is, just like the relative risk, a measure of association. There is some resemblance between Hill's concept of analogy and Hume's fifth rule. It is possible to find in Hume's writing formulas that correspond to Hill's concepts of specificity and coherence. For historical reasons, Hume could not have expressed two aspects of causality mentioned by Hill. The concept of "biological plausibility", as biology is a 19th century science, and that of experimental or semi-experimental evidence. Again, the key element is the similarity of the intellectual approaches rather than the exact formulations.

I do not know whether Yerushalmy or Hill were familiar with Hume's rules. However, independently of whether Hume's Rules were known to Hill or Hill's predecessors, Hume had a concept of causality assessment that was very similar to that of most contemporary epidemiologists. Hume's rules sound reasonable to us, and most likely Hill's ideas would have sounded reasonable to Hume. Actually, both Hume and Hill say essentially the same thing: when deterministic demonstration is not available, it is imperative to screen the causal statement for illogicalities or gross contradictions between what has been found and what we think we know. Hume and Hill's com-

Table 27 – Hill's criteria and corresponding Hume's "Rules by which to judge of causes and effects".

Hill's criteria (see section 3.9 for more details)	**Hume's rules**
1. Temporality: "The temporal relationship of the association – which is the cart and which is the horse?"	1. "The cause must be prior to the effect" (Rules 1 and 2)
2. Dose-response: "If the association is one which can reveal a biological gradient, or dose-response curve, then we should look most carefully for such evidence. (...) The clear dose-response curve admits of a simple explanation and obviously puts the case in a clearer light."	2. "When any object encreases or diminishes with the encrease or diminution of its cause, 'tis to be regarded as a compounded effect, deriv'd from the union of the several different effects, which arise from the several different parts of the cause. The absence or presence of one part of the cause is here suppos'd to be always attended with the absence or presence of a proportionable part of the effect. This constant conjunction sufficiently proves, that the one part is the cause of the other" (Rule 7).
3. Consistency: "Has it [the association] been repeatedly observed by different persons, in different places, circumstances and time?"	3. "...multiplicity of resembling instances, therefore, constitutes the very essence of power or connexion" (not a specific rule but in the premises of the catalog, p. 163)
4. Strength of association	4. "There must be a constant union betwitxt the cause and effect" (Rule 3)
5. Analogy: "In some circumstances it would be fair to judge by analogy. With the effects of thalidomide and rubella before us we should surely be ready to accept slighter but similar evidence with another drug or another viral disease in pregnancy."	5. "...where several different objects produce the same effect, it must be by means of some quality, which we discover to be common amongst them" (Rule 5). "Like effects imply like causes" (Rule 5).
6. Specificity	6. "The same cause always produces the same effect, and the same effect never arises but from the same cause" (Rule 4)
7. Biological possibility	7. Not applicable
8. Experiment	8. Not applicable
9. Coherence	9. Not a rule

plicity may thus have a historical ground. Hume provided the philosophical bases of the 17th century "probabilistic" revolution, which gave birth to the two fundamental epidemiologic principles, population thinking and group comparisons.

Thus, the logic of causal inference described by Hill and generally regarded as the appropriate approach by today's epidemiologists finds its origin in the intellectual changes that occurred in Western philosophy in the 17th and 18th centuries.

4.10. Principles of knowledge acquisition in epidemiology

According to Piaget, the genesis of a scientific discipline is based on a principle of knowledge acquisition. Direct action on an object has been traditionally the principle of knowledge acquisition in physics. What is then the corresponding knowledge-generating tool in epidemiology? Group comparisons combined with population thinking appear as good candidates. Indeed, group comparisons consist of coordinated actions. The process is very similar to that of the mathematical example described previously in which we were able to discover the law of commutativity by counting the same pebbles in different orders. In group comparisons, people are sampled from a population and rearranged (e.g., grouped into exposed and unexposed categories and simultaneously followed) in such a way that the perspective offered by the reorganization of the population in groups differing by exposure or affection can reveal potential causal links. The mode of knowledge acquisition in epidemiology is therefore closer to logic and mathematics than to physics.

Group comparisons and population thinking started to contributing to knowledge as soon as they merged, in the 18th century. The physician "experimenting" with treatments from one scurvy patient to the other was not able to derive universal knowledge that would still be valid today. But when Lind showed that, *other things being equal*, the sailors who ate the oranges and lemons were cured from scurvy while those in the other groups were not, some knowledge had been acquired that is still valid today. Lind and his successors did not know why citrus fruits could treat scurvy. But it did not really matter. They had a cause, on which they could act to modify a health outcome. Lind succeeded with an n = 12 experiment where centuries of trials and errors by physicians had failed. This shows how powerful the group comparison approach was. It is likely that physicians had had many opportunities to observe the beneficial effect of citrus fruits, but it is only when the observation was made within a given experimental design with coordinated actions that it generated knowledge, exportable to other places and valid for other populations than the sailors of the Salisbury.

In the 18th and 19th centuries, it suddenly became obvious that group comparison was the only strategy available if the outcome was to be observed only in a fraction of subjects exposed to the postulated cause. Consider Snow's London experiment in the summer of 1854. There had been an average of about 9 cholera deaths per thou-

sand households. There was no way the role of the water supply could be put in evidence without grouping the households based on a clear definition of exposure. Snow and Farr knew that a controlled experiment was needed to demonstrate the effect of polluted water but such an experiment obviously could not be conducted. This may explain why Snow immediately recognized the "Grand Experiment" in the data produced by Farr's administration when the 1854 epidemic took place. And indeed, after grouping the households by water providers, there were about 30 cholera deaths per 1,000 households supplied by the Southwark and Vauxhall Company *vs.* 4 per 1,000 households supplied by the Lambeth Company and 6 per 1,000 households in the other districts of London.

Human thinking, philosophy, and mathematics became mature enough to embrace group comparison in the 19th century. Once the principles of knowledge acquisition existed, the evolution of epidemiology could be traced as the progressive refinement or enrichment of these principles by methods and concepts. At the beginning, these were very intuitive forms of counts and comparisons of like with like. With time, they became more abstract and formalized. We saw how ratios led to proportions, and then risks and rates, and how intuitive group comparisons paved the way towards a theory of epidemiologic study designs. The work of Lind, Louis or Snow consisted of simple forms of comparisons and frequency measures. When Einstein discovered relativity, there was not a single methodological textbook of epidemiology. As theory developed, methods became less intuitive and served for designing experiments suitable for solving complex problems. Still, the understanding of epidemiologic methods did not require any mathematical skills. In its latest phase, the methods and concepts have become much more abstract and are virtually out of reach for people who do not have some mathematical background.

5. Phases of epidemiology

On the basis of the evidence available today to a non-historian, it is reasonable to conclude that before the 18th century there was no research based on population thinking or group comparisons and that there could therefore be no epidemiology as the discipline we know today. In the 18th century, group comparisons and population thinking merged in the activities of physicians such as James Lind or the English proponents of the "medical arithmetik", that is, the usage of mass observations collected on patients as an additional source of knowledge for medical practice beyond the teaching of the great clinicians (Troehler, 2000). Since then, epidemiology has emerged as a set of research methods, which have contributed to elucidating important questions related to human health. Over about 150 years, epidemiologists have developed and refined the designs of cohort and case-control studies, the concepts of confounding and interaction, the categorizations of types of bias, and the process of causal inference.

In this continuous genesis of epidemiologic methods and concepts, I propose to distinguish four phases, characterized by qualitative changes in the level of formalization and abstraction of the concepts and methods: preformal, early, classic and modern epidemiology. "Preformal" means that none of the concepts and methods had been *formally* defined.

The point of this section is to show that this categorization in four phases is meaningful and that each phase had unique features of its own, mostly using material that has been already presented in the previous sections. An exhaustive historical review has still to be written.

5.1. Preformal epidemiology

Until the end of the 19th century, there was no specific theory of population thinking and group comparisons backing the activity of epidemiologists. The mathematical and philosophical bases existed but no formal theory. Let us call this first phase, *Preformal epidemiology*, during which scientists used population thinking and group comparisons, spontaneously, without referring to some theory. People such as Lind, Snow or Farr *invented* their way into epidemiologic research and therefore set the bases for the future development and formalization of methods and concepts.

5.1.1. Preformal epidemiologists

Preformal epidemiologists were mainly physicians but with diversified interests. For example, Farr was a physician, a public health professional and a statistician. Snow was an anesthesiologist, a clinician and a public health scientist. These people had different objectives. Some searched for ways to act on the environment to improve public health, other assessed the efficacy of treatments to improve patient care, and probably all aimed to develop human knowledge with respect to the determinants of health and disease. But their common denominator is the fact that they strived for their objectives using the same two principles: population thinking and group comparisons. Eventually, the use of these two principles was to characterize epidemiology, and differentiate it from medicine, statistics, economics, etc. with its own conceptual and methodological corpus.

5.1.2. Population thinking and group comparisons

We have seen that the use of different measures of disease occurrence, such as risks and rates, can be traced back to Graunt. William Farr established a clear conceptual difference between risks and rates. This first theory of risks and rates shows that the distinction of phases in the evolution of epidemiologic methods and concepts is somewhat arbitrary, and that the evolution has really been a continuous process. It is also

true however that people like Farr were exceptional. Population thinking and group comparisons found a lot of resistance, especially in medicine, hampering their use by medical doctors.

For physicians, population thinking appeared to conflict with the fundamental principles of medicine. How can we generate information from a group of patients that is relevant for the single patient? Isn't the patient unique? How can medical knowledge rely on probabilities? The controversy that surrounded Louis's numerical method illustrates the types of criticism that were expressed by physicians.

A first group of physicians rejected Louis' numerical method because they believed medicine was an art of individual prediction and could not rely on group-based probabilities. A professor of pathology and general therapy from Montpellier, in the south of France, Benigno Juan Isidoro Risueño d'Amador (1802–1849) represented the category of opponents for whom medicine was the art of healing individual patients. He requested an audience at the April 25, 1837 session of the French Royal Academy of Medicine. His point was that the role of the physician was care for individual patients, and that no statistics could predict what would happen to a specific patient. If on average 10% of the patients died from a given intervention, the physician could not forecast which patients these would be. The information was therefore useless to the physician whose primary concern is to determine which individual would become sick or die. Thus, the uniqueness of each patient made it impossible to generalize from past patients to future patients, and made the calculus of probabilities "completely useless in medicine" (cited by Matthews, 1995a, p. 27).

Claude Bernard (1813–1878), one of the most esteemed and influential medical physiologists of his century, is emblematic of another category of opponents to Louis's methods. Bernard agreed that group comparisons were needed to evaluate therapies. But he also professed that medical knowledge could not be based on probalility. For Bernard, averages did not exist in nature. Physiology, in contrast to statistics, described medical phenomena as they were repeatedly and constantly observed across experiments. Physiology discovered facts and laws. In the presence of variation across experiments, the physiologist would search for the determinants of such variation and certainly not hide it by making average descriptions of experiments.

Joseph Lister (1827–1912), the English founder of modern antiseptic surgery, expressed ideas similar to those of Bernard. Lister had actually compared the mortality related to surgical procedures some years before (1864 & 1866) the introduction of antiseptic methods and during the three following years (1867– 1869). Mortality had been cut by three, from 1 death every 2.17 cases to 1 every 6.66 cases (Lister, 1870a). For some reason Lister did not include the data for 1865. He and many of his contemporary colleagues interpreted these results as strong evidence in favor of anti-sepsis. But much fewer were those who recognized that the effect of antiseptics was due to their capacity to kill germs. Antiseptics prevented infections, which were the real causes of death, but still physicians were inclined to attribute their striking effect to "some specific virtue" of the antiseptic:

> *"... the striking results which were recorded were too often attributed to some specific virtue of the agent. The antiseptic system does not owe its efficacy to any such cause, nor can it be taught by any rule of thumb. One rule, indeed, there is of universal application – namely this: whatever be the antiseptic means employed (and they may be very various), use them so as to render impossible the existence of a living septic organism in the part concerned."* (Lister, 1870b, p. 288).

Thus, for Lister, group comparisons were too superficial. They could not reveal that the true scientific foundation of the antiseptic effect was the presence of germs responsible of infection, that is, the "germ theory of putrefaction" (Lister, 1870b, p. 288).

Finally Auguste Comte (1798–1857), the leader of the French school of positivism, relied some biological arguments to oppose Louis's principles, arguing that comparing the statistical effects of two treatments was "impossible" because a sick human organism reacted differently than a healthy one.

These episodes indicate how isolated population thinkers, and therefore epidemiologists, were in the 19th century scientific, and especially medical, environments.

It is of note that preformal epidemiologists were at ease with exposed/non-exposed or affected/non-affected group comparisons. John Snow's 1854 Grand Experiment compared households exposed to polluted water and households that were not. But in another investigation performed during the same epidemic around the Broad Street pump, frequencies of exposure to the water pump were compared in people affected *vs.* non-affected by cholera (Paneth et al., Part II). Pierre Louis describes his work on the use of bleeding in the treatment of pneumonia in terms of exposed/unexposed comparison, but he also used affected/unaffected comparisons in other circumstances, as for example, to assess the potential hereditary origin of emphysema, a chronic lung disease leading to respiratory insufficiency:

> *"Of 28 patients with emphysema, 18 had their mother or father affected by that same disease", while "of 50 individuals free of emphysema, only three had affected relatives."* (Louis, 1837, p. 255).

The conjunction of population thinking and group comparisons was necessary for the emergence of the new discipline of epidemiology. There was no progress in the understanding of the causes of infectious diseases when public health data were not ordered and analyzed according to the principles of group comparisons. A recent reanalysis of a report of the City Council of Ferrara, Italy, on the cholera epidemic of 1855 illustrates the limitations of public health without epidemiology (Scapoli et al., 2003; Morabia, 2003; Vandenbroucke, 2003). The cholera epidemic in Ferrara occurred a year after the London cholera epidemics during which Snow's Grand Experiment took place. Why is it that the determinants of the Ferrara epidemic are barely understandable, even using modern statistical techniques to analyze them,

whereas in London, John Snow was able to successfully demonstrate that the cause of cholera was related to polluted water? The dominant model in public health was that air pestilence, poverty, overcrowding, and lack of hygiene were responsible for the epidemic of cholera. The data from Ferrara showed that more people tended to be diagnosed and die from cholera when they lived in dirtier streets, smaller and less hygienic houses, etc. This corroborated the model, even though the data also indicated higher case fatality rates in large and more hygienic houses, which did not fit the poverty model too well. The Ferrara City Council may not have been pursuing the correct hypothesis, but, more important, it was not using the right methodological approach either. For an 1855 observer in Ferrara, ecological correlations indicated that socio-demographic and urbanistic factors had weak and sometimes paradoxical effects on mortality from cholera. Its records did not lend themselves to non-ambiguous group comparisons. What was missing was the conceptual leap that gave birth to the corpus of epidemiologic methods and concepts, that is, collecting data in such a way that *comparing groups* on specific exposure and outcome could shed light on potential causal associations.

5.1.3. More examples

In the first part of this essay, we have glimpsed the work of Lind, Snow, Farr and Louis. These were not the sole pre-formal epidemiologists who combined population thinking and group comparisons. In his investigation of the epidemic of measles on the Faroe Islands in 1846, the Danish physiologist Peter Ludwig Panum (1820–1885) used an early form of relative risk to compare the age-specific number of deaths during the first 8 months of 1846 with the average number of annual deaths from 1835 to 1845. For example, there had been 50 deaths under age 1 in 1846 *vs.* "18 1/11th" in 1835–45, yielding a relative risk of about 2.8. Panum wrote that:

> *"Number of times mortality in first two-thirds of 1846 was greater than the usual in an ordinary whole year: about 2 9/11."* (Pan American Health Organization, 1988, p. 38).

Ignaz Philipp Semmelweis (1818–1865), a Hungarian physician teaching medicine in Vienna, observed that the mortality from puerperal fever was two to four times higher among women delivered by physicians compared to women delivered by midwives. In 1846, mortality had been about 11.4% in medical deliveries ("First clinic") *versus* 2.7% in midwife deliveries ("Second clinic") (Carter, 1983). Semmelweis speculated that these differences were caused by the fact that examining physicians went from pathological dissections and contact with dead bodies to deliveries without thorough cleansing between the two activities. At the end of May, 1847, Semmelweis introduced the practice of washing the hands with a solution of chloride of lime before the examination of lying-in women. Subsequently, the mor-

tality from puerperal fever stabilized around 2% or less for both midwives and physicians (Carter, 1983).

In his report on the mortality of Cornwall miners, 1860–1862, William Farr presented annual mortality rates by ten-year age groups, which were, for metal miners, 3.77, 4.15, 7.89, 19.75, 43.29 and 45.04, and, for "males exclusive of metal miners", 3.30, 3.83, 4.24, 4.34, 5.19 and 10.48. He used these rates to compute relative risks of mortality from pulmonary disease comparing metal miners to males who were not miners:

> *"...assuming as before that the rate of mortality among the males exclusive of miners is represented at each period of life by 100, then that among the miners would be represented by 114 between the ages of 15 and 25 years, by 108 between 25 and 35, by 186 between 35 and 45, by 455 between 45 and 55, by 834 between 55 and 65, and by 430 between 65 and 75 years. It is therefore evident that pulmonary diseases are the chief cause of the excess mortality among the Cornish miners."* (Pan American Health Organization, 1988, pp. 68–69).

William Augustus Guy (1810–1885), Professor of Forensic medicine and Hygiene at King's College Hospital in London, compared the occurrence of "pulmonary consumption" across a variety of occupations. Guy used odds, that is, the ratio of the number of cases with pulmonary consumption to the number of other diseases, as a measure of risk (Lilienfeld and Lilienfeld, 1979). Guy, in 1843, had also considered (and ruled out) the possibility that the relation of job and health could reflect the self-selection of jobs by workers according to their health status rather than to the effect of the job on health (Vineis, Part II).

Christiaan Eijkman (1858–1930), a Dutch physician, received the Novel Prize for having established that beriberi was a nutritional disease. Beriberi was a fatigue disease, involving weight loss, muscle weakness, loss of feeling and eventually death in up to 80% of the cases. In the local idiom, the word beri means weak, and doubling it intensifies its meaning. The contribution of epidemiology to this discovery came from Adolphe Vordermann (1857–1902). Between May and September of 1896, this supervisor of the Civil Health Department of Java compared the occurrence of beriberi among the 280,000 inmates of 100 Java prisons. According to local customs, prisoners were fed either polished rice, half-polished rice, or a mixture of both. Beriberi was found in 2.7% of the prisons feeding half-polished rice (corresponding to 1 in 10,000 prisoners), in 46.1% of the prisons preparing a mixture of polished and half-polished rice (1 in 416 prisoners), and in 70.6% of the prisons serving exclusively polished rice (1 in 39 prisoners) (Allchin, 2000; Carpenter, 2000). On the other hand, beriberi was not associated with hygienic conditions of the prisons such as the age of the building, the permeability of the floor, ventilation, or population density. It was later established that it was the polished rice deficiency in thiamine (vitamin B1) that was causing beriberi.

5.1.4. Definition of epidemiology

Overall, the balance between successes and failure is positive for epidemiology during this preformal phase. There was no discipline called epidemiology, and defined as such, but the fight against infectious disease was a domain of activity that acquired a name. The first scientific society of epidemiology, the London Epidemiology Society, was created in 1850. Some of its members were epidemiologists, but none had an academic appointment and extremely few (e.g., Farr) wrote theoretical/methodological work. The situation changed dramatically in the 20th century.

5.2. Early epidemiology

Let us call *early epidemiology* the development phase in which some epidemiologic concepts and methods were assembled for the first time into a theory of population thinking and group comparisons.

5.2.1. Early epidemiologists

Before 1880, epidemiologists were essentially amateurs (general practitioners like Snow, Semmelweis, military and naval physicians and surgeons). After 1880, public health professionals were hired in England to practice "epidemiology" (e.g., John Simon, William Frederick Barry, Theodore Thompson, H. Timbrell Bulstrode, Edward Ballard, William G. Savage) (Hardy, Part II).

A salient trait of this second phase is the creation of university positions of professors of epidemiology and the publication of the first textbooks. Almost simultaneously in the US and the UK, epidemiology became an academic field. After World War I, Major Greenwood was appointed lecturer and in 1930 professor of epidemiology in the Department of Epidemiology and Vital Statistics created in 1927 at the London School of Hygiene and Tropical Medicine, where he remained until he retired in 1945 (Hardy and Magnello, Part II). In the United States, Frost was appointed in 1922 as Professor and Chairman in the Department of Epidemiology and Public Health Administration at The School of Hygiene and Public Health of the Johns Hopkins University in Baltimore (Comstock, Part II).

The line of demarcation between epidemiology and statistics remained fuzzy. Major Greenwood considered himself a "professed statistician" (Greenwood, 1935, p. 21) and wrote one of the first textbooks of epidemiology (Winkelstein, Jr., 2002; Winkelstein, Jr., 2003; Lilienfeld, 2003; Bracken, 2003). Greenwood and Bradford Hill had strong connections with statistics, and were disciples of the Cambridge statistician Karl Pearson. Bradford Hill entitled his textbook "*Principles of medical statistics*", but the text contained very little mathematics and could perfectly have been called "*Principles of clinical epidemiology*". Even though the title of Hill's book re-

ferred to medical statistics, it had a lot to do with epidemiology and group comparisons. For Hill

> *"The essence of the statistical method lies in the elucidation of the effects of these multiple causes."* (Hill, 1939, p. 3).

And by statistical method he understood:

> *"methods specially adapted to the elucidation of quantitative data affected by a multiplicity of causes"* (Hill, 1939, p. 3).

We may consider the scientific duet between Snow and Farr as a preformal collaboration between epidemiology and statistics. Other duets of this type existed in this early phase. Edgar Sydenstricker was the "first national public health statistician" (Wiehl, 1974). He played a key role in the methodological developments of the early phase of epidemiology. He worked closely with Goldberger on the pellagra investigations and developed a life-long collaboration with Frost, whom he provided with the Massachusetts data used for the cohort analysis paper (Section 3.4.3).

5.2.2. Population thinking

Preformal epidemiology had paved the way for population thinking by early epidemiologists. The latter further refined the description of disease occurrence in population by separating prevalence from incidence. In a lecture given on December 15, 1931 at the Johns Hopkins University School of Hygiene and Public Health in Baltimore, Sydenstricker distinguished the "prevalence of illness" based on surveys, which are affected by "cases of long duration and of chronic type" from the "incidence of illness", based on continuous recording of an illness in a population. According to Sydenstricker, incidence of illness was first measured on a large scale in the studies he had conducted with Goldberger on the causes of pellagra (Pan American Health Organization, 1988, pp. 168–169).

Somewhat related to the distinction between prevalence and incidence, the study of chronic diseases also called for new methods of surveillance and of group comparisons. In 1935, the Connecticut State Legislature authorized a population-based cancer registry. In Denmark, a cancer registry covering the whole population was set up in 1942. These registries played an important role in revealing the rising trends of lung cancer incidence, which motivated the following generation of epidemiologists (Terracini and Zanetti, Part II).

5.2.3. Group comparisons

In 1927, Frost published an article entitled "*Epidemiology*" in the Nelson Loose Leaf Encyclopedia (Frost, 1941), which, according to his successor as Chair of epidemiology at The Johns Hopkins School of Public Health, Abraham M. Lilienfeld, was "the first systematic exposition of epidemiology as a scientific discipline" (Lilienfeld, 1983). The paper can be viewed as the module of a textbook. Two citations from "*Epidemiology*" show that for Frost, epidemiology consisted in the conjunction of population thinking and group comparisons. Population thinking:

> *"For the clinical description of a disease the unit is an individual, and the phenomena of the clinical reaction may be described in terms of the character and distribution of the anatomic lesions and the nature and sequence of symptoms. For epidemiologic description the unit is the aggregation of individuals making up a population, and description of mass-phenomena of a disease consists of a statement of its types and frequency of occurrence in the population as a whole and in its different component groups."* (Frost, 1941, p. 494).

And group comparisons:

> *"In every epidemiologic investigation, whether its immediate purpose be to explain a localized epidemic or to elucidate the general spread of an obscure disease, the first step is to investigate the association between the occurrence of the disease and some special condition or set of conditions. This is primarily a statistical process of ascertaining the frequency of the disease in two or more populations separated with respect to the particular condition."* (Frost, 1941, p. 540).

Frost referred to the work of Snow as being a model of group comparisons. It is of note that Frost also gave one of the earliest descriptions of the cohort study design. Frost wrote that:

> *"The simplest and most direct method of determining whether or not such an association exists* [that is, *"that the occurrence of the disease is in some way associated with the use of sewage polluted drinking water"*], *is the method used by Dr. Snow, namely, that of ascertaining what different water supplies are used within the area of investigation, and how those supplies differ with respect to sewage pollution; then classifying (1) the persons who have died from cholera, and (2) the entire population, according to their sources of water supply. It remains to ascertain the frequency of deaths from cholera in each of the two groups of the population, which differ with respect to the sources and sewage pollution of their water supplies, that is, to ascertain the ratio of deaths to total of persons in each group. If the difference of incidence in the two groups is found, as in this instance, to be entirely outside the range of such differences as may be expected in two*

groups of such size drawn at random from the population, it may reasonably be inferred that in this area the use of sewage polluted water is positively associated with the liability to death from cholera. It is further found, by still another independent inquiry, that the two groups are, so far as can be ascertained, quite similar in all other conditions of composition and environment, hence the association of cholera mortality with character of water supply is a rather direct one. It is, of course, equally necessary to show that the two water supplies actually differ materially with respect to the degree of sewage pollution." (Frost, 1941, pp. 537–538).

The oral history of epidemiology says that Frost taught the "techniques of prospective and retrospective studies" already in 1933–34 (Susser, 1985, p. 152). Thus, early epidemiologists built on the experience accrued in the 19th century to strengthen the foundations of group comparisons. Frost "made John Snow a hero" (Vandenbroucke et al., 1991) because, in retrospect, Snow was *the* historical example of the successful combination of group comparisons and population thinking.

Distinct improvements also occurred for the affected/non-affected comparisons. Clinicians started to compare groups of patients suffering from disease believed to have different etiology on a larger scale than ever before. In 1926, the British physician and former Dean of the London School of Medicine for Women Janet Lane-Claypon (1877–1967) compared 500 hospitalized breast cancer cases and 500 controls with non-cancerous illnesses from both inpatient and outpatient settings in London and Glasgow (Lane-Claypon, 1926). This early case-control study, which had Major Greenwood as statistician, indicated that cases were more likely to be single or to have lower fertility when married. In 1928, a New England Journal of Medicine's paper (Lombard and Doering, 1928) compared the habits, characteristics and environment of individuals with and without cancer in Massachusetts and showed that cancer patients had smoked more pipes and cigarettes than non-cancer controls. These were first experiences with a new type of study design, eventually termed the case-control study.

In the first half of the 20th century, epidemiologists also contributed to improving the design, analysis and interpretation of therapeutic trials. The James Lind Library (http://www.jameslindlibrary.org) has already assembled the documentation of a substantial collection of therapeutic trials performed during that time.

5.2.4. Concepts

During this second phase, epidemiologic methods and concepts acquired some theoretical foundations. These were somewhat less intuitive than in the previous phase but remained quite basic. A major theoretical contribution of this phase consisted in identifying sources of fallacious interpretations of group comparisons, and in proposing solutions to minimize them.

The idea that an observed association may in reality be indirect or spurious because the compared groups differ in some important way has always been present in the epidemiologic thinking. Preformal epidemiologists intuitively understood the concept of "confounding". Lind, Louis, Snow were always preoccupied with comparing like with like. The examples that we reviewed in this essay speak for themselves of the evolution of the concept of confounding during the following phase.

In 1904, Yule gave the first formal description of the mechanism of confounding. He referred to it as a fallacy associated with the mixing of records. The mechanism of confounding described by Yule was also reported in their textbooks by Greenwood in 1935 and by Hill in 1937. Hill did not cite Greenwood, who did not cite Yule. They must have thought that this was an exercise of simple logic and not a meaningful discovery. Indeed, Greenwood mentions that this type of fallacy has "vitiated many published reports" (Greenwood, 1935).

In 1920, Goldberger and Sydenstricker performed stratified analysis and computed standardized rates to separate the effects of diet, age, gender and income, which were correlated causes of pellagra. In 1930 Hill applied the alternate allocation of treatment in the British Medical Research Council therapeutic trial as a form of study design that could increase the comparability of groups and allow the researcher to compare "like with like". In a posthumous paper published in 1939, Frost explained the use of cohort analysis as a way to prevent fallacious interpretations of cross-sectional data, especially when looking at diseases with a prolonged survival such as tuberculosis. The formalization of the mechanism and modes of control of confounding had clearly progressed.

5.2.5. Definitions of epidemiology

It is also during this period that epidemiology got its first definitions as a discipline. The evolution of the definitions of epidemiology reflects its process of differentiation from other scientific disciplines. "Epidemiologists" themselves were still unclear about what epidemiology was. In 1919, Frost defined epidemiology as the study of the determinants of infectious diseases (Comstock, Part II). This definition implied that people studying non-infectious diseases were not epidemiologists. The case of Goldberger around the time of Frost's first definition is an interesting one. He studied the causes of pellagra, a disease people believed to be infectious, but which he believed was produced by diet and poverty.

The epoch of early epidemiology was characterized by the transition from the dominance of acute infectious diseases to that of chronic diseases in the "global" burden of disease. Epidemiologists became increasingly involved with the study of chronic conditions and the definitions of epidemiology changed accordingly. In 1927, Frost expanded his definition to include some but not all non-infectious diseases. In 1935 Greenwood defined epidemiology as

> *"the study of disease, any disease, as a mass phenomenon"* (Greenwood, 1935, p. 15).

or as

> *"a science of group etiology"* (Greenwood, 1935, p. 21).

In 1937 Frost finally generalized his definition to all aspects of human health (Comstock, 2001).

5.3. Classic epidemiology

The years after 1945 were particularly fruitful for the development of epidemiology. The discipline of epidemiology, in contrast to all other human and social sciences, has been uniquely able to perform vast community-based studies to investigate the causes of heart disease, cancer and other chronic conditions, which characteristically have a long incubation and require long-term follow-up. Millions of people have been involved in epidemiologic studies. As a result, new epidemiologic methods were developed and older ones were refined, in particular in the context of the controversy about the health effects of tobacco smoke. The process occurred almost in parallel in the US and in Great Britain.

5.3.1. Classic epidemiologists

Most classic epidemiologists who authored textbooks were medical doctors (e.g., Jerry Morris, Brian MacMahon, Mervyn Susser, Abraham Lilienfeld). Few had formal training in epidemiology or statistics. The close collaboration with statisticians persisted in this phase. The case of Jerome Cornfield is an interesting one. Cornfield, who played a decisive role in creating the bases for the modern understanding of case-control studies, graduated from New York University in 1933 with a B.A. in history but later became President successively of the American Epidemiologic Society (in 1972) and of the American Statistical Association (in 1974).

5.3.2. Population thinking

We have reviewed in previous sections the considerable development of population thinking during this phase, related to a better understanding of the relation of prevalence to risk. On the one hand, the prevalence of a *disease* could be viewed as the product of its incidence and of its duration (i.e., $P = I \times D$). On the other hand, large fractions of disease cases in a population could be produced by small risks applied to large fractions of the population with low levels of exposure (i.e., Rose's prevention paradox).

A special note must be made here about ecologic correlations, a form of population thinking which had been used in the past but which received a strong impetus in classic epidemiology. Ecologic correlations consist of relating exposure and outcome data, which are only available as group averages and not as individual observations. Typically, the 1964 Surgeon General Report included a graph showing that countries with higher per capita cigarette consumption in 1932 tended also to have higher mortality rates from lung cancer in 1970 (US Public Health Service, ed, 1964, p. 176). Ecologic correlations were common in other fields, such as sociology, in which it had been established that they could not be used as substitutes for individual correlations (Robinson, 1950).

The very influential article by Doll and Richard Peto, also epidemiologist at Oxford University, entitled "*The causes of cancer*" (Doll and Peto, 1981) was essentially an ecologic study. The Office of Technology Assessment of the US Congress had commissioned it:

> *"If the foregoing is accepted as justifying that much human cancer is avoidable, then a crude estimate of the proportion of cases that might be avoided in any one community can be obtained by comparing for each separate type of cancer the incidence in that community with the lowest reliable incidence recorded elsewhere."* (Doll and Peto, 1981, p. 1205).

That is, the lowest incidence observed was considered inevitable, while the entire excess incidence beyond the lowest level could be attributed to external factors and, in theory, be prevented. Doll and Peto concluded that:

> *"75 or 80% of the cases of cancer in both sexes might have been avoidable."* (Doll and Peto, 1981, p. 1205).

In particular, they estimated that 30% of U.S. cancer deaths were due to tobacco, 35% to diet, and 4% by occupational exposures (Doll and Peto, 1981, pp. 1256). The section of the article entitled "*4.3. Use of epidemiological information*" (pp. 1217–1219) reflects the level of self-assurance reached by classic epidemiology:

> *"... to make estimate of the proportion of today's cancers that are attributable to avoidable causes (...) epidemiology, influenced by laboratory investigation, is by far superior to the latter alone. Epidemiology has at present an undeservedly low reputation among people who have first artificially limited themselves to wondering which environmental pollutants to restrict and who then find that almost none of the few thousand chemicals they are worried about have been adequately studied by epidemiologists. This is, however, to condemn epidemiology for failing to achieve ends that it does not have."* (Doll and Peto, 1981, p. 1219).

In his 1973 textbook (Susser, 1973), Susser stressed that individuals and the group represented different "levels of organization". This second feature of ecologic correlations provided qualitatively distinct insights into exposure-disease associations. Elucidating the source of discrepancy between individual and ecological correlations could illuminate causal links. The debate on the interpretation of ecologic correlations and their role in the armamentarium of epidemiologists is still ongoing (Schwartz, 1994).

5.3.3. Study designs

Large cohort studies are one of the major features of classic epidemiology. In October 1951, Doll and Hill launched the British Doctors prospective study (Doll and Hill, 1954; Doll, Part II).

> *"Bradford Hill suggested that doctors would make a suitable population to study as they might be more interested in responding to a questionnaire about smoking habits than most other people, that having had a scientific training they might be more accurate in the description of their smoking habits, and, most importantly, that they would be relatively easy to follow up, because of the need to keep their names on the Medical Register for legal reasons."* (Doll, Part II).

In October 1951, Doll and Hill sent a short questionnaire (seven questions) to 59,600 members of the medical profession in the United Kingdom eliciting their smoking habits. Of the 41,024 replies, 40,564 were sufficiently complete to be utilized. Initially, the Office of the Registrar General of births and deaths (the national bureau of vital statistics which Farr had directed 100 years before) provided the death certificates from all doctors. Hammond and Horn began the first American Cancer Society Cancer Prevention Study sending questionnaires to 188,000 white males, aged 50 to 59 (Hammond and Horn, 1958). The causal nature of the association of smoking and lung cancer was finally fully recognized in both countries in the early sixties.

Several cohort studies were performed on the basis of historical records, that is, "retrospective cohort studies", especially to study occupational exposures (Stellman, Part II). In this volume, Doll describes the Framingham Heart Study, whose results have had an enormous impact on clinical medicine (Doll, Part II).

During this phase, case-control studies acquired a theoretical basis as a study design. It was three case-control studies, two in the United States (Levin et al., 1950; Wynder and Graham, 1950) and one in Great Britain (Doll and Hill, 1950) that firmly launched the debate on the association between smoking and lung cancer. Cornfield established that the odds ratio computed in the case-control study was, under certain assumptions, a close approximation to the relative risk. From then on, case-control studies became the most common study design in epidemiologic research (Paneth et al., Part II).

The theory of study designs made its appearance progressively in the literature (Cornfield et al., 1959; Cornfield, 1951; Cornfield and Haenszel, 1960; Dorn, 1959), but especially in textbooks, which flourished during this phase (Zhang et al., Part II).

5.3.4. Concepts

Classic epidemiology brought the concepts of confounding, bias and interaction to a higher level of generalization. In previous sections we have reviewed the developments that occurred during this phase around the concepts of interaction and causal inference.

In this essay, I only discussed the evolution of selection bias (section 3.7). The history of bias is of course much richer. The concept evolved from a list of dozens of biases to types of based on their mechanism, the two main ones being information (i.e., misclassification) and selection bias (Vineis, Part IIa). The history of misclassification bias is relatively recent as it is closely related to that of screening. The new availability of simple diagnostic devices (e.g., blood or urinary sugar concentration, Papanicolaou smear tests) created the conditions for the development of population-wide screening, especially for diabetes, cervical and breast cancer (Morabia and Zhang, 2004). The screening tests were not dangerous but had imperfect validity. A whole theory of test interpretation, involving the concepts of sensitivity, specificity and predictive values, was developed (Morabia and Zhang, 2004). Applied to group comparisons, these new concepts contributed to the development of a theory of misclassification bias (Newell, 1962; MacMahon et al., 1960).

5.3.5. Definition of epidemiology

The "*Dictionary of Epidemiology*" sponsored by the International Epidemiology Association provides the following definition of epidemiology:

> *"The study of the distribution and determinants of health-related states of events in specified populations, and the application of this study to control of health problems."* (Last, 2001).

This "classic" definition of epidemiology integrates population thinking ("study of distributions") and group comparisons ("study of health determinants"). The second part of the definition specifies that the usage of these principles is oriented towards the improvement of the public health. Indeed, in classic epidemiology, the theory cannot be separated from its medical and social applications. "Classic" textbooks read like essays on the determinants of human health and the methods to assess them (Zhang et al., Part II).

Not only did classic epidemiology develop epidemiologic theory, but it had some distinct achievements, the most salient being the establishment of a causal link be-

tween exposure to tobacco smoke and risk of lung cancer synthesized in the classic "*US Surgeon General Report*" of 1964. Classic epidemiology created the foundations for further theoretical developments.

5.4. Modern epidemiology

Let us call this latest phase of the genesis of epidemiology *modern epidemiology*, in reference to the most influential textbook presenting these new theoretical developments (Rothman, 1986; Rothman and Greenland, 1998).

5.4.1. Modern epidemiologists

There is a strong contrast in the professional backgrounds and profiles between the generation of epidemiologists who contributed to this new phase and the classic epidemiologists. Many have PhDs but not MDs. This is the case for many authors of the textbooks of this new phase (Rothman and Greenland, 1998; Kelsey et al., 1986; Kleinbaum et al., 1982). Most if not all have a strong background in mathematics or statistics. This generation of epidemiologists went further in the formalization of methods and concepts. As a result the discipline became much more mathematical. Where classic epidemiology expressed concepts that had no necessary mathematical translations, almost all concepts (e.g., bias, confounding, interaction, etc.) in modern epidemiology can be written either in words or equations.

The methodological and conceptual core of modern epidemiology can be found in a textbook entitled "*Theoretical Epidemiology*" (Miettinen, 1985). Its author, Olli Miettinen, expressed the watershed between classic and modern epidemiology by saying that epidemiology was previously

> *"widely regarded as commonsense activity, a line of research that any physician – even one without statistical education – is prepared to engage in"* (Miettinen, 1985, p. VIII).

Modern epidemiology went beyond common sense. The novelty of the approach proposed was first only understood by a small circle of students, who re-expressed the new concepts and made them accessible to a wider audience.

Let us make a short digression here and consider again the analogy with physics. There is a point at which theoretical progress becomes irrelevant for our everyday life. As far as physics is concerned, we can comprehend most of the phenomena in our daily life if we don't go beyond the Newtonian, mechanical vision of the world. Few people are versed in relativity, even less in quantum physics.

It is the same in epidemiology. At a given moment, the theoretical developments become irrelevant for the bread and butter activity of the epidemiologist. A

Cornfieldian vision of epidemiology suffices. For example, the algebraic relationship between cumulative incidence and incidence density, the impact of control sampling schemes on effect estimation, are usually pointless when the phenomenon studied is rare. This is why classic epidemiology remains for many active epidemiologists the phase of reference. But the developments of modern epidemiology are crucial for our understanding of what we do, for the identification of the exceptional situations in which the choices of study design and of measure of disease occurrence matter, and for our ability to carry the discipline forward. They are progressively becoming part of the intermediate epidemiology curriculum.

5.4.2. Population thinking

We have seen in the earlier sections that the concepts of risks and rates underwent a profound transformation during this phase, with new names (e.g., cumulative incidence and incidence density), and new formal, mathematical links (e.g., cumulative incidence can be viewed as a function of incidence densities).

5.4.3. Study designs

Classic epidemiology had established a theory of cohort and case-control studies, and discovered that the relative risk could be estimated from both designs. Still, there remained many doubts about the validity of case-control studies. In an influential paper published in 1959 by the New England Journal of Medicine (Dorn, 1959), Harold Dorn (?–1963), chief of the Biometrics Branch at the Division of Research Services of the National Institutes of Health, wrote:

> *"I do not wish to give the impression that I reject retrospective studies as a method of investigating the etiology of chronic diseases. The retrospective method, in theory, can provide data of reliability and validity comparable to that obtained from prospective studies. But, as usually applied, it does not do this. The fundamental defect of many retrospective studies is that they are based on an unspecified sample of persons chosen by an unknown method of sampling from an unidentified population"* (Dorn, 1959, p. 577).

The paper was rather favorable to case-control designs but it set very high standards for a valid design. Modern epidemiology has clarified which were these standards.

It became clear that the key criterion for the validity of the case-control study was that the cases and controls originate from the same source population. All case-control studies became viewed as being nested within cohorts, whether the cohort has been actually enumerated and characterized as a cohort study (i.e., nested case-control studies) (Doll, Part II; Doll, 1998), or whether the cohorts are hypothetical. This led to a new conceptualization of the different ways of sampling controls and to the

decline of the "rare disease assumption" formulated by Cornfield to equate the case-control study odds ratio with the cohort study relative risk.

5.4.4. Concepts

Modern epidemiology is the first phase comprising a coherent set of methods and concepts spanning the different circumstances that the researcher faces when trying to establish the causal nature of an association: methods for comparing groups, sources of biases, presence of multiple independent causes (i.e., confounding), presence of multiple interrelated causes (i.e., interaction), and causal inference.

Without going into details, the theory of confounding became enriched by: a) a more rigorous formulation of the conditions under which confounding occurs; b) a theory of matching in cohort and case-control studies; c) the simultaneous treatment of multiple confounders.

The theory of biases was further developed with a better understanding of the effects of losses to follow-up in cohort studies and selection in case-control studies, the implications of misclassification of exposure, disease and confounders, whether differential or not, across the compared groups.

The concept of interaction became, as we also saw, much more refined, with a distinction between interaction of attributable risks (i.e., additive interaction or public health interaction), and interaction of relative risks (i.e., multiplicative interaction).

Causal inference did not evolve much in this new phase even though it was intensively debated. The debate turned in particular around the question of the relevance of the "falsification of hypotheses" approach proposed by the epistemologist Karl Popper (1902–1994) for the design and interpretation of epidemiologic studies (Greenland, 1987a; Rothman (ed), 1988). The Popperian approach did not succeed in replacing the traditional Humean approach formalized by classic epidemiologists.

5.4.5. Definition of epidemiology

What is the definition of modern epidemiology? The second edition of "Modern Epidemiology" states that:

> *"the ultimate goal of most epidemiologic research is the elaboration of causes that can explain patterns of disease occurrence"* (Rothman and Greenland, 1998, p. 29).

This definition is a good reflection of the state of the discipline, because it relates its methods (elaboration of causes or "etiology") to its subject matter (disease occurrence). The evolution of epidemiologic methods and concepts has been driven by the search for causes of human diseases. It is likely that this will remain the driving force of epidemiologists and of epidemiology. Nevertheless, it is important to note that at

that stage of abstraction, concepts and methods become independent of specific issues, such as public health or simply health-related problems. They can be applied in any field in which combining group comparisons and population thinking can be an appropriate mode of knowledge acquisition.

5.5. What will come next?

If the scheme of analysis used to describe the genesis of epidemiology is correct, modern epidemiology is a transitory phase, just like the early and classic epidemiology phases were. The new theoretical tools allow us to address problems of increasing complexity both in the biological and the social dimensions. These will require further theoretical developments and lead to a new phase, the nature of which can probably be perceived in the latest theoretical work of epidemiologists.

6. Conclusion

In this essay, I have attempted to show that:

1) epidemiology is characterized by the combination of population thinking and group comparisons aiming to discover the determinants of human health;
2) the set of methods (study design) and concepts (measures of disease occurrence, confounding, interaction, bias) have evolved since the 17th century. This evolution is consistent with Piaget's theory of genetic epistemology.
3) In this evolution, we can identify four phases characterized by qualitative leaps in formalization and abstraction of the methods and concepts. After a preformal phase, in which epidemiology was discovered intuitively by scientists, most of all physicians, epidemiology has gone through an early, a classic and a modern period.

History and epistemiology are not a type of a general culture, a knowledge that it is nice to have but that is not essential for the active life of the epidemiologists. On the contrary, I argue that they are an integral part of the background of epidemiologists. Understanding the origin and evolution of epidemiologic methods and concepts can stimulate Scientific creativity. Methods and concepts are tools. These tools improve with time. Each epidemiologist should be ready to contribute to this improvement by adapting the methods and concepts to the solution of new or more complex problems than those which the available methods have contributed to solving in the past.

What will be the next phase in the evolution of epidemiologic methods and concepts? I am not sure we can tell yet. But the discipline is certainly undergoing a period of uneasiness. Our current methods have been successful for the discovery of

major, relatively independent determinants of health in the environment (e.g., germs, tobacco smoke, radiation, asbestos, or social inequalities), in our behaviors (e.g., physical activity, vitamins, alcohol, drugs) and in our biology (e.g., genes, lipids, blood pressure, obesity). They are not that well adapted to assess the more complex manifestations of many (e.g., genetic, social, infectious) health determinants. These contradictions between old methods and new problems should induce theoretical developments, and make epidemiology enter into a new qualitative phase in order to continue to improve human health.