Clusters, classification and epidemiology of interstitial lung diseases: concepts, methods and critical reflections

J.M. Antó*, #, P. Cullinan\textsuperscript{\dag}


ABSTRACT: The present article reports on two conceptual and methodological issues concerning interstitial lung disease (ILD) about which there is a lot of misunderstanding and contradiction: the investigation of epidemics and clusters and the classification and epidemiology of ILD.

In general, an epidemic is the occurrence of cases of an illness in excess of normal expectancy. The investigation of an epidemic often demands a number of sequential studies: first descriptive, then aetiological. Clusters consist of an increase in incidence of much smaller magnitude, perhaps excluding the possibility that this is merely a result of chance. In recent years, more valid statistical methods for the assessment of clusters have developed.

For interstitial lung disease in particular, different classifications exist that are sometimes inconsistent and may change with time. All diseases require a process of ascertainment, whereby they are identified, classified and perhaps registered. The two most employed epidemiological techniques for testing aetiological hypotheses are the cohort and case-referent approaches.


The aim of epidemiology is to understand the causes of disease and, by inference, develop approaches to aid prevention. Classically, its method begins with the description of the distribution of a disease within or between populations, and proceeds with analyses of the factors that determine this distribution. In this way, aetiological hypotheses are generated by observation and tested by formal study.

Such neat distinctions conceal the muddle which is more often characteristic of the process by which a disease’s aetiology is uncovered. However, the framework is a useful one, particularly in understanding where epidemiology has failed. Despite several decades of effort, the aetiologies of most interstitial lung diseases (ILDs) remain a mystery. Arguably, this is a reflection of a poor understanding of the basic distributions of these diseases and the consequent paucity of aetiological hypotheses for formal analysis.

In this context, the identification of clusters and small epidemics may eventually allow relevant risk factors to be identified. Consequently, several clusters of ILD have been reported in the literature. This article examines some of the difficulties facing epidemiologists in this area, focusing on the difficulties of establishing a clear classification of ILD, as well as the investigation of clusters.

The investigation of epidemics and clusters: conceptual and methodological issues

Definition of an epidemic

The term epidemic refers to the occurrence, in a community or region, of cases of an illness in excess of normal expectancy. Usually, such excess is so sharp and of such magnitude that chance fluctuation is a very unlikely explanation and an immediate response from the public health services is frequently necessary. However, the concept of epidemicity is relative to the usual frequency of the disease in the same area, among the specified population, and during the same season of the year [1]. From this point of view, even a small number of cases of a very rare disease may constitute an epidemic. Epidemics of respiratory disease have been reported both in the general population and in occupational groups. The former is illustrated by urban epidemic asthma [2]. Unfortunately, there are many examples of occupational outbreaks that have involved many different respiratory diseases, including silicosis, asthma, byssinosis, hypersensitivity pneumonitis and even malignant diseases, such as lung cancer.
Sequential studies in an epidemic

The investigation of an epidemic often demands the performance of a number of sequential studies [3] that may involve different study designs.

The first stage consists of descriptive studies aimed at defining the problem, describing the existing data and formulating aetiological hypotheses. This step involves a case definition, the identification of cases and the description of the distribution of cases in time and space. Some characteristic time-space models have been described in the form of a line, an area, or a point source, whose identification may help to establish aetiological hypotheses. The distribution of cases in a closed community, such as a factory or a school, may also be relevant to the generation of aetiological clues. The physiopathogenic mechanism of the disease may also help to formulate a hypothesis. However, experience has shown that epidemiological research often allows the aetiology of diseases to be identified in the absence of a well-established biological mechanism. A relevant lesson, learnt from investigation of the environmental origins of many epidemics, is that gathering clear-cut aetiological evidence may be extremely difficult and involves long-term intensive research. For example, in the case of toxic-oil syndrome, final characterization of the responsible chemical compounds and its toxicological mechanism has not yet been possible [4].

The second stage consists of an aetiological investigation, which is, at the start, retrospective by definition. In the general population, where the enumeration of persons at risk is not usually possible, cross-sectional and case-control are the appropriate study designs. By contrast, in studies on industrial outbreaks, retrospective cohort studies are sometimes possible. The aims and characteristics of the case-control study may vary in parallel with the elaboration of more refined hypotheses about the particular circumstances of the causative exposure. This is well illustrated by the investigation of the eosinophilia-myalgia syndrome [5, 6]. In the first step of this investigation, the initial cases of eosinophilia-myalgia were defined. The second step involved a case-control study, which identified an increased risk in those who reported the ingestion of tryptophan supplements. Next, two sequential case-control comparisons were conducted among subjects that reported the ingestion of tryptophan. These two analyses permitted identification of the responsible chemical compounds and its toxicological effects has not yet been possible [4].

Assessment of clusters

In recent years, more valid statistical methods for the assessment of clusters have developed. This development has been of such magnitude that the epidemiological and statistical analysis of clusters has become a specialized field that has been covered by several journals [11–13]. Recently, a cooperative effort has permitted a first comparative validation experiment on the methods for cluster analysis, and has shown that one test cannot be relied on and that false-positive results cannot be totally excluded [14]. The latter is consistent with experience accumulated over many years, that many of the clusters that have been studied are attributable to chance, and that an association between cases and well-defined exposures can rarely be established. In a review of almost 400
clusters, conducted in the state of Minnesota, only five were subjected to a complete epidemiological survey and only one deserved to be published in a specialized journal [15]. It should be emphasized that the study of such clusters represents a considerable human and financial effort, and that making appropriate decisions about when further investigation is warranted is of utmost importance. Use of the guidelines developed by the Centres for Disease Control (CDC) for the investigation of clusters is highly recommended for the latter [16].

The investigation of clusters in respiratory diseases may be difficult. One illustration of this is seen in the study of sarcoidosis in fire-fighters. Kern et al. [17] reported a unique time-space cluster of three cases of sarcoidosis among 10 fire-fighters who had trained together. The investigators put forward the hypothesis that fire-fighters may have an increased risk of sarcoidosis, and undertook a study of 1,282 active and retired fire-fighters and police officers in search of common exposures to smoke and infectious agents. The study did not show any increased risk relating to the suspected exposure. Obviously, three cases of sarcoidosis in this large population represented such a small number that is was difficult to exclude clustering or chance as alternative explanations. Another type of clustering is seasonality, which has recently been reported in sarcoidosis presenting with erythema nodosum [18].

The possible difficulties in the investigation of clusters are further illustrated by two outbreaks of ILD that have been reported in textile workers. Lougheed et al. [19] described the occurrence of five cases of ILD among 88 workers at a Canadian textile factory and attributed the outbreak to the inhalation of aflatoxins. Kern et al. [20] reported eight cases of chronic ILD among 165 workers in a US plant of the same Canadian Company. In the latter study, it was possible to fully enumerate the cohort of workers, which included 162 males and three females, who accounted for 660 and nine person-yrs at risk, respectively. Through retrospective cohort analysis, the authors estimated an incidence of 1,495 cases per 100,000 person-yrs calculated from a population-based register in New Mexico [21]. These two studies also illustrate the difficulties of performing aetiological case-control or cohort analyses in order to elucidate the potentially causative exposures in a small number of cases. Ad hoc environmental measures did not find elevated levels of aflatoxins in the US plant and suggested that the disease could have been caused by the inhalation of respirable-size nylon fragments.

Conclusions

Organizing the research of an epidemic can be a rather complex task that demands cooperation between different professionals and institutions. The results of the research can have relevant social, economic and legal implications, and it is therefore no wonder that these sometimes lead to diverse points of view amongst researchers. When this happens, it is of utmost importance to have an appropriate comprehensive study protocol in place. Epidemics are likely to keep occurring due to the continuous introduction of new hazards into the environment, and the insufficient development of public health systems in most countries. In addition to facilitating the environmental control of the causative factors, the investigation of clusters and outbreaks is likely to produce new aetiological knowledge. The identification of a new cause as a result of the investigation of an epidemic, may prevent its further occurrence in other places. Environmental epidemiology also has a relevant role in this preventive process.

Critical reflections on epidemiology and classification

Classifications

ILDs are classified by their aetiology, pathology, radiology, associations with other diseases, responses to therapy, or even by the absence of an alternative diagnosis (fig. 1). In many cases, combinations of these features are used. Systems of classification change with bewildering frequency and are often inconsistent between different countries or between physicians in the same country. In some cases, for example the histological evaluation of biopsy specimens, their reproducibility is open to doubt. As a result, unambiguous case definitions are often difficult to establish.

In broad, aetiological terms, ILDs may be classified by three categories of understanding. 1) There are several instances where close associations between ILDs, particularly fibrotic ILDs, and exposures to a variety of specific environmental stimuli are seen. The clearest relationships are seen after occupational exposures, presumably through inhalation, to both inorganic and organic agents. Several inorganic fibrogens (particularly asbestos and beryllium) have clear exposure-response relationships, at least under conditions of high exposure. As with virtually all diseases, these relationships are modified by individual susceptibility, an area exemplified by berylliosis [22].

Fig. 1. – Possible classifications of interstitial lung diseases by aetiological certainty and by clinical, radiological or histological features. RASS: rheumatoid arthritis-systemic sclerosis; MCTD: mixed connective tissue disease; CFA: cryptogenic fibrosing alveolitis.
Environmental factors may also dominate the aetiology of extrinsic allergic alveolitis, although evidence of this is less clear. Therapeutic drugs may also cause ILDs, presumably reaching the lung through the circulation. In some cases, especially antinecancer and anti-arrhythmic therapies, where doses are often high and treatments prolonged, dose-response relationships may be discernible [23]. Often, however, pulmonary disease only develops in a small proportion of treated patients, in which case, the abnormal response appears to be idiosyncratic and determined by host susceptibility factors. In many cases the pathology resolves once treatment is stopped. 2) Beyond these disease states lie a wide range of clinical and/or pathological states, the aetiologies of which remain either largely or wholly unknown. Most are uncommon, some excessively rare, but two (cryptogenic fibrosing alveolitis (CFA), synonymous with idiopathic pulmonary fibrosis (IPF), and sarcoidosis) are less so. 3) Occupying an uneasy (to the epidemiologist) position somewhere between these two classes of understanding are the collagen-vascular associated ILDs, the causes of which are presumed to be the same as those underlying the related connective-tissue condition.

Alternative, pathological classifications of these diseases do not neatly coincide with this aetiological grouping (fig. 1). For example, fibrosis resulting from heavy exposure to asbestos may be indistinguishable, at a microscopic level, from that labelled as cryptogenic. Similarly, the histological features of the collagen-vascular pulmonary fibroses may be very like those of some idiopathic conditions, although they appear to follow a different and often less aggressive clinical course [24]. Sarcoïdosis and berylliosis are similar, granulomatous conditions, which are distinguished chiefly by a demonstrable immunological response to beryllium and an appropriate occupational history. There may also be similarities with pulmonary diseases not generally classified as interstitial; for example, sarcoidosis shares many of the features of tuberculosis.

Such overlaps presumably reflect the limited and stereotyped responses of the lung to toxic stimuli. If so, there are at least three important implications for the epidemiologist. First, observations of disease distribution that depend entirely on pathological patterns may be misleading; grouping conditions by their homogenous, microscopic pathology alone may obscure important differences in distribution. Thus, Loeffler’s syndrome may be aetiologically quite distinct from other forms of pulmonary sarcoidosis. Second, the extent to which the aetiological understanding of diseases in category 1 can be translated to those of similar pathology in categories 2 and 3 is debatable. Third, the current pathological classifications can lead to logical absurdities, such as the difficulty in making a diagnosis of CFA in a male with a history of occupational asbestos exposure.

**Ascertainment**

In order to be recognized, all diseases require a process of ascertainment whereby they are identified, classified and perhaps registered. This process may be more or less straightforward, depending on the clarity with which a disease is clinically manifest and distinguishable from other diseases. In all cases, however, the apparent distribution of a disease within a population is, to a greater or lesser extent, a reflection of its ascertainment. In general, diagnoses of ILDs can only be made using sophisticated technology. Thus, their recognition and consequent ascertainment are closely linked with access to such services. This has important implications, not only for descriptive epidemiology, but also for the familiar difficulty surrounding the choice of referent populations in analytical, case-control studies. A similar probability of ascertainment in each group seems essential for such studies; where cases are derived from hospital populations, it is arguably more appropriate to select referents from the same study base, in which the processes of recognition are broadly similar.

Sarcoïdosis provides a reasonable example of some of these difficulties. The disease is commonly without clinical manifestations [25] and thus, may only be detectable by chest radiography. Outside campaigns of mass radiography (which are themselves selective), recognition of the disease is partly a reflection of the probability of having a chest radiograph. For example, in Rochester, MN, USA, sarcoïdosis was more commonly identified as "asymptomatic" among recent immigrants and healthcare professionals, presumably because these groups undergo chest radiography more frequently [26]. Similarly, part of the apparent clustering of the disease on the Isle of Man [27] may be a reflection of the high number of cases among nurses working at the island’s main hospital.

For the same reasons, temporal patterns apparent in the frequencies of these diseases should be viewed with caution. The increase in prevalence of sarcoïdosis on the Isle of Man between 1977–1983 [28] almost certainly reflects better case ascertainment, and similarly, the reported increases in male mortality from CFA in England and Wales [29] are within the bounds of improved diagnostic methods.

**Distributions: age and sex**

At its most basic, descriptive epidemiology is concerned with observations of age and sex distributions among systematically collected patient populations. When these are available they can provide useful, if limited, aetiological information.

Sarcoïdosis has a peak incidence in young adulthood. In Scandinavian and British populations, a second peak is apparent among 50–60-yr-old females. The disease is rare in childhood and among the elderly. Its sex distribution is almost equal, although certain manifestations may be more common among females. These features would be unusual in a disease that is the result of dust inhalation, particularly occupational dust inhalation. They are more typical of an infective aetiology, albeit one with an unusual distribution and perhaps a strong immunological component. Epidemiological and microbiological attempts to support this have not yet been successful. The
search for epidemiological markers of infection, essentially the demonstration of disease clustering in time and space (including within families), has been unsuccessful and, as argued above, is open to the vagaries of differential ascertainment. The pathological similarities between sarcoidosis and tuberculosis (and the similarities in their age and sex distributions, where the latter is endemic) have led to many attempts to identify mycobacterial antigens in sarcoid tissues, a quest fuelled by the availability of polymerase chain reaction and similar techniques. Overall, the results have been inconsistent and ultimately disappointing, which may be due to the extreme sensitivity of the detection techniques and the near-ubiquity of nontuberculous mycobacteria. Recent reports of propionibacterial deoxyribonucleic acid (DNA) in sarcoid tissue may, however, provide an avenue for further, more successful investigation [30].

Conversely, CFA is largely a disease of later life, and is probably more common among males than females. A number of occupational aetiologies for the disease have been proposed, and some formally tested. For example, associations with wood- and metal-work have been reported [31], although only within a relatively confined geographical area. The size of the resultant odds ratios suggests that only a small proportion of the disease can be attributed to these occupations. Disease misclassification must also be considered, particularly since these occupations are also strongly linked with mesothelioma. More broadly, however, if occupational exposures were responsible for all or a substantial proportion of cases of the disease, then an increasing male/female incidence ratio with age would be anticipated. Available evidence from death certificates in England and Wales suggests that this pattern is only weak. Conversely, infective aetiologies for the disease are required to invoke some form of delayed response to explain the disease’s age distribution [32].

Other distributions

Remarkably, little is known about other distributions. It is currently impossible to make detailed, geographical comparisons of CFA, as international recognition of the disease, at least in terms of death certification, varies [33]. Those observed in sarcoidosis, generally from comparison of mass radiographic surveys between different countries, but occasionally from occupational screening programmes [34], are not easy to interpret. Overall, these studies suggest that northern latitudes increase the prevalence of the disease [25], but again, it is not clear to what extent this (and similar geographical patterns) reflects biases in ascertainment. Similar difficulties are encountered in the assessment of temporal trends. Racial variations in the incidence of sarcoidosis have been reported with some consistency and probably cannot be wholly explained by geographic or socioeconomic confounding. Smokers are under-represented among patients with sarcoidosis [25], and over-represented among those with CFA [31]. In the former disease, there is evidence that this may be explained, in part, by greater access of young, nonsmokers to chest radiography [35]; a similar, but opposite bias may be acting among older populations where CFA is prevalent.

Epidemiological techniques

Formal, epidemiological techniques for testing aetiological hypotheses are generally nonexperimental; the two most frequently employed are the cohort and case-referent (-control) approaches. Cohort studies are inefficient where the outcome is rare, and have been used infrequently in the study of ILDs. Case-referent studies, in which the point of entry is defined by disease state, have been used much more widely. Overall, the results have been unconvincing; BRESNITZ and STROM [34], for example, list over 20 such studies in sarcoidosis, none of which has led to consistent or sustainable aetiological concepts. This may, in part, reflect the previously discussed difficulties in referent selection. However, it is more probable that this results from the use of this methodology to test inadequately grounded hypotheses. Case-referent studies are seldom, if ever, successful in the generation of hypotheses; those that are based on vague questions concerning gene-environment interactions usually produce equally vague answers.

Conclusions

Faced with the above, it is tempting to adopt an attitude of despair toward the epidemiology of interstitial lung diseases. These issues, however, are not unique to interstitial lung disease and indeed, to a greater or lesser extent, are relevant in the study of all disease. It is important that these issues are recognized and used to inform all observations of the distribution of interstitial lung disease. From such observations, either from a formal epidemiological exercise or careful clinical practice, testable aetiological hypotheses should emerge.