A recent review of diagnostic criteria for upper limb musculoskeletal disorders found 27 published classification systems, no two of which were the same. The differences related not only to the criteria by which individual disorders were specified and the names by which they were identified, but also to the range of diagnoses distinguished. Such diversity of classification, which is by no means confined to rheumatology, presents a major challenge to the epidemiologist. Optimal case definition is important in the design of studies, but can be problematic when there is no satisfactory gold standard against which to assess potential diagnostic criteria. In this paper, we examine the basis for diagnostic classifications in medicine and suggest a way of addressing the difficulties that arise in the absence of agreed gold standards.

Diagnosis as a descriptor of disease

When medical students embark on their clinical training, they are taught first how to elicit a history from a patient and how to conduct a relevant physical examination. The aim is to establish a ‘correct’ diagnosis so that prognosis can be predicted and appropriate treatment can be given. Often, a firm diagnosis cannot be made on the basis of only symptoms and signs, and further investigation is required to distinguish between several possible differential diagnoses, for example by radiology or blood tests. Sometimes, an initial diagnosis is revised in the light of new findings at operation, at autopsy, or as the patient’s illness evolves over time, because it no longer appears ‘correct’. Implicit throughout is the notion that one or more specific disease processes are responsible for the patient’s presenting complaints and are waiting to be discovered, or alternatively that in reality there is no underlying pathology. Diseases are conceived as having an independent existence, and clinicians use clinical data to identify the diseases that are most likely to be present.

This model has obvious limitations. Illness is not simply a manifestation of disease. (Different authors have variously defined terms such as ‘illness’ and ‘disease’. For definitions of these and other key terms as used in this paper, see Table 1.) It...
arises from a complex interplay of pathological, physiological, psychosocial, and cultural influences. Nevertheless, epidemiologists have widely embraced the concept of diseases as objective natural phenomena that can be observed, classified, and investigated, and most epidemiological textbooks include sections on assessment of the ‘accuracy’ of diagnostic tests. We recognize that many diseases occur in a continuous spectrum of severity such that the borderline of normality is ill defined and somewhat arbitrary (e.g. osteoporosis, sensory-neural deafness), and that the means by which we classify a person as having a disease are generally imperfect and sometimes subjective. However, these problems do not detract from our belief in the objective nature of the diseases that we study. Thus, in formulating a case definition, our aim is to distinguish individuals in whom a specific disease is present or has occurred at a specified level of severity.

Thinking of case definition in this way works well when disorders are associated with clearly defined pathology that can be established as present or absent with reasonable confidence in at least a representative sample of the study population. For example a fracture of the femoral neck will almost always be demonstrable radiologically, and the presence of e-antigen on serological testing normally provides a reliable index of active infection by the hepatitis B virus. Even if it is not practical or ethical to apply the relevant diagnostic criteria in everyone who is studied, they provide a gold standard against which other case definitions can be tested. Any lack of sensitivity or specificity can then be taken into account when interpreting the results of investigations that use them. For example, thorough histological examination of tissues is generally regarded as providing reliable evidence for or against a diagnosis of cancer (i.e. a gold standard). However, many valuable epidemiological discoveries have been made using less accurate clinical diagnoses of cancer from death certificates (most notably in cohort studies using cancer mortality as an outcome), the effect of any diagnostic misclassification generally being to bias risk estimates towards the null.

Problems arise when there is no satisfactory diagnostic gold standard for the pathology that is presumed to underlie a disorder, and still more when the underlying pathology is unknown.

Pathology cannot always be diagnosed unequivocally

Raynaud’s disease is characterized by episodes in which the distal parts of one or more fingers become pale, cold and numb, and then red and painful. It is believed to result from intermittent spasm of blood vessels in the digits, and can be identified objectively if a patient is observed during an attack. However, currently there is no sure method of excluding the pathology when episodes are not witnessed. How in situations such as this, can we compare potential case definitions and optimize the choice?

One criterion might be the repeatability (consistency) of diagnosis between observers. If there is substantial disagreement between adequately trained independent observers in the classification of cases, then a method of diagnosis cannot be considered entirely satisfactory. However, the fact that a diagnostic test is objective and repeatable does not necessarily imply that it is meaningful, and recognition of this limitation has in some cases led to a paradox. Thus, in exploring the use of cold challenge for the diagnosis of Raynaud’s phenomenon, researchers have assessed its sensitivity and specificity against reported symptoms as a standard. However, if symptom history provides a reliable gold standard then there should be no need for more elaborate investigations such as cold challenge.

Katz et al. have proposed that where there is no satisfactory diagnostic gold standard, the best proxy against which epidemiological case definitions can be assessed may be the opinion of an expert clinician. Classification would then be determined by an empirical approach that identified the elements of history, physical examination, and laboratory tests that best assigned subjects to diagnostic categories specified by the clinician. However, expert clinicians do not always agree, and even when there is a consensus of experts, we cannot be certain that they are right, or that their opinion on what constitutes a case will not change over time. The fourth edition of Brain’s textbook of neurology, published in 1951, described Alzheimer’s disease as a pre-senile dementia, affecting people between the ages of 40 and 60 years. Now, of course, it is thought to be the most common cause of cognitive decline in the elderly, and older age would not be a reason for excluding the diagnosis.

Disorders for which the underlying pathology is unknown

The challenges that confront the epidemiologist are even greater when the pathogenesis of disease is uncertain. Disorders such as schizophrenia are distinguished on the basis of clinical findings, in the belief that the illnesses of those diagnosed arise from a similar, although as yet undefined, pathological mechanism. Historically, such hunches have sometimes turned out to be correct. For example, scurvy was recognized as a distinct clinical entity long before James Lind showed (in 1747) that it could be successfully treated with citrus fruit, and even longer before vitamin C was isolated and synthesized (1932). However, in the absence of a postulated disease mechanism, how can one set of diagnostic criteria be evaluated in comparison with another?

Statistical techniques

One approach that has been advocated when there is no satisfactory diagnostic gold standard is the use of latent class analysis. In this technique a mathematical model is constructed in which a ‘latent’ variable is assumed to represent an individual’s ‘true’ disease status, and the model is used to estimate errors for different diagnostic tests that might be applied. However, the findings depend on assumptions about the inter-dependence of the errors from each test, and this cannot be established empirically.

A utilitarian view of diagnosis

An alternative way out of the difficulty is to think of diagnoses not necessarily as labels for disease processes, but more generally as a useful method of classifying people for the ultimate purpose of preventing or managing illness. This approach has been advocated previously in psychiatry.
could usefully be applied more widely. The value of a case definition is then determined by its practical utility in distinguishing groups of people whose illnesses share the same causes or determinants of outcome (including response to treatment), and competing case definitions can be compared according to their performance against these criteria.

There is evidence, for example, that numbness and tingling in the hand which is confined to the sensory distribution of the median nerve and affects most of that area, differs in its association with risk factors from the same symptoms occurring in other anatomical patterns (being more strongly associated with activities that physically stress the wrist, and showing no association with psychological risk factors). If correct, this may form the basis for a useful diagnostic distinction. Another example is the classification of non-Hodgkin’s lymphoma, in which methods based solely on histological appearances have been superseded by schemes that incorporate data on immunological cell surface markers, with improved prognostic accuracy.

In psychiatry, diagnostic classifications based entirely on history and clinical examination have been developed through successive iterations following a process described by Robins and Guze as ‘continuing self-rectification and increasing refinement leading to more homogeneous diagnostic groupings’. These have been shown to distinguish successfully between patients with different prognosis and response to treatment. Similarly, in the classification of musculoskeletal disorders such as low back pain, useful schemes have not required the specification of pathology, but have described the pattern of symptoms and patient experience.

If diagnostic criteria are viewed in this way, then optimizing their utility will involve a trade-off between ‘lumping’ and ‘splitting’. When diagnostic categories are too inclusive, the effects of risk factors and determinants of outcome may be diluted to the extent that they are missed or ignored. But if case definition is too discriminatory, the statistical power of investigations may be compromised, as well as the scope for exploiting results in all of the circumstances to which they are relevant. It would be unfortunate, for example, if the benefits of a treatment for non-Hodgkin’s lymphoma in general were missed because it had been investigated in only one subtype of the tumour.

Thinking of diagnoses as a useful method of classifying people does not preclude their being related to underlying pathology. Diagnoses that group people whose illnesses arise through the same identified pathological process are likely to be useful for preventive or therapeutic purposes. However, knowledge of underlying pathology is not an essential requirement. Another starting point might be an observed syndrome—i.e. an unusual clustering of symptoms, signs or other clinical features in certain individuals. The fact that clinical features cluster indicates that they may arise as part of the same disease process. And even if there is no shared pathology, there may be shared psychosocial or cultural causes that are amenable to manipulation.

Clinical consensus can also provide a useful initial basis for classification in so far as it represents what clinicians perceive to be a useful way of categorizing patients. However, the value of such classifications cannot be assumed without empirical evidence of utility. The Quebec task force on Spinal Disorders used a consensus process and systematic literature review to arrive at a classification of low back disorders based primarily on symptoms and signs. This has helped our understanding of the differences in prognosis of different presentations of low back pain, and has advanced efforts to manage the disorder better.

**Optimal case definition can vary**

An important implication of adopting a utilitarian approach to diagnosis is that the optimal case definition for a disorder may vary according to the circumstances in which it is applied. Thus, from analysis of clinical findings in patients who had consulted a physician with hip pain, and using clinical diagnosis as the gold standard, the American College of Rheumatologists concluded that osteophytosis was the best radiographic discriminant of hip osteoarthritis. This suggests that the presence or absence of osteophytosis would be a useful component of case definition for osteoarthritis in studies of hospital patients with hip problems, among whom the prevalence of inflammatory arthritis may be substantial. However, in community-based studies, where the prevalence of inflammatory hip disease is much lower than that of osteoarthritis, the most important diagnostic requirement is to differentiate cases not from people with other arthritides, but from people who have no hip disease at all. For this purpose, the presence of joint space narrowing may be a more useful criterion. Another example is the classification of chronic renal disease. In aetiological studies there is probably value in distinguishing patients with glomerulonephritis from those with other underlying pathologies, whereas this distinction may be less important when investigating the clinical management of end-stage renal failure.

**How can case definitions be optimized in practice?**

When faced with a need to classify illness for which there is no satisfactory diagnostic gold standard, researchers can explore the merits of potential case definitions empirically. For example, they may investigate the discriminatory potential of different diagnostic groupings in relation to patterns of association with known or suspected risk factors (as in the example of sensory disturbance in the hand described earlier), or in predicting prognosis and clinical outcome when a sample of subjects is followed longitudinally. In this exploratory phase it will usually be sensible to start with the finest classification that can meaningfully be analysed statistically, and then aggregate categories for which no clear distinction is apparent. When using this approach, however, it is important to be aware that some apparent similarities or differences between potential diagnostic groups may be a product of chance, and that when applied to other samples of subjects, the aggregated categories may therefore not perform as well. This problem can be addressed by developing the diagnostic classification in a random subset of subjects, and then testing its performance in the remainder.

Another aspect of methodology that may require care is the choice of measures of association for comparison of relationships to risk factors. In cross-sectional studies of common disorders such as shoulder pain and low back pain, associations with risk factors are often summarized by prevalence ratios. However, the maximum value of a prevalence ratio is constrained by the...
prevalence of the disorder in unexposed subjects, since by definition prevalence cannot be greater than 100%. This makes prevalence ratios less satisfactory when comparing the associations of a given risk factor with disorders that differ markedly in their prevalence, and in these circumstances it may be preferable to use odds ratios.

Conclusion
We recommend that epidemiologists should view diagnoses not necessarily as labels for diseases but more generally as a useful method of classifying people for the purpose of preventing and managing illness. Where there is good reason to believe that a category of illness results from a defined pathological process and the presence or absence of this pathology can be established accurately, then this provides a gold standard against which other case definitions can be assessed. However, where there is no established underlying pathology or no credible gold standard for the presumed underlying pathology, case definitions should be evaluated according to their practical utility in the elucidation of preventable causes and the optimization of clinical care.

KEY MESSAGES
• Optimal case definition is important in epidemiology, but can be difficult when there is no satisfactory diagnostic gold standard.
• One approach to this problem is to think of diagnoses not necessarily as labels for disease processes, but more generally as a useful method of classifying people for the ultimate purpose of preventing or managing illness.
• The value of a case definition is then assessed by its ability to distinguish, usefully, groups of people whose illnesses share the same causes, or prognosis and response to treatment.

References