intra-articular steroid injections (usually hydrocortisone acetate) may work wonders, especially when combined with muscle exercises. Intra-articular steroids work particularly well in the knee, shoulder, and in tenosynovitis of the flexor tendons of the hand, and my use of them increases year by year, but never at more than three-month intervals. If repeated injection is needed I tend to use intra-articular radioisotopes to ablate the synovium in older patients and surgical synovectomy in younger ones.

Misuse or severe damage of one joint often leads to problems in nearby joints. Subluxed and painful metatarsal heads often predispose to knee or ankle problems, or a flexion deformity of the knee leads to hip or ankle problems in the opposite side. Identification of the problem and correct treatment will reduce the need for unnecessary tablets.

Systemic corticosteroids

The vogue of systemic steroids in rheumatoid arthritis is long over. No one will deny the temporary benefit given to patients, but their long-term side effects outweigh their advantage, even in low dosage. Long-term trials show this effect to be lost after the first year, after which steroid-treated patients do worse. Most rheumatologists feel they see less "vasculitis" and less amyloidosis now that systemic steroids are used less, and the higher rate of wound infections, vertebral collapse, and skin fragility in steroid-treated patients deters me from using them. Besides, many patients spontaneously request me not to start steroids. I do use them for two groups: firstly in wage-earners about to lose their job and housewives unable to cope while I wait for the gold or penicillamine to take effect; and secondly in patients in whom all else has failed and whose quality of life is poor. Ideally I try to use them on an alternate-day basis or as a morning daily dose of not more than 5 mg, as both of these minimise the steroid side effects more commonly seen with thrice daily or night-time regimens.

Remissions

Remission of inflammation may occur spontaneously or with disease-modifying drugs. At this stage NSAID may be reduced or stopped, or conversion to pure analgesic drugs such as Distalgesic may relieve the pain of damaged joints.

I acknowledge much valuable discussion with my colleagues, and the assistance of the Pharmacy at Northwick Park Hospital in the preparation of table I.

References

[1] Huskisson, E C, *Pharmatherapeutica*, 1976, **1**, 30.
[2] Kay, A G L, *British Medical Journal*, 1976, **1**, 1266.
[3] Gumpel, J M, *Rheumatology and Rehabilitation*, 1976, **15**, 217.

Epidemiology for the Uninitiated

Repeatability and validity

GEOFFREY ROSE, D J P BARKER

British Medical Journal, 1978, **2**, 1070-1071

An ideal survey technique is repeatable (it gives the same answer when the subject is re-examined) and also valid (it measures what it purports to measure). Poor repeatability implies poor validity, because only one answer can be the right one. But a consistent answer may also be wrong: a laboratory test might yield persistently false-positive results, or a highly repeatable psychiatric questionnaire might be an insensitive measure of, for example, "stress." To assess how much weight to attach to epidemiological results calls for numerical estimates of both their repeatability and their validity.

Department of Medical Statistics and Epidemiology, London School of Hygiene and Tropical Medicine, London WC1E 7HT
GEOFFREY ROSE, DM, FRCP, professor of epidemiology

Faculty of Medicine, University of Southampton, Southampton General Hospital, Southampton S09 4XY
D J P BARKER, PHD, MRCP, reader in clinical epidemiology

Repeatability

The principles of repeatability testing were described in last week's article. Results for numerical variables such as blood pressure may be expressed as the *standard deviation* of replicate measurements or as the *coefficient of variation* (standard deviation÷mean). Separate estimates may be given for within and between observers, or for variability between measurements made consecutively and those made on separate occasions.

For qualitative attributes, such as clinical symptoms and signs, the results are first set out as a contingency table:

		Observer 1	
		Positive	Negative
Observer 2	Positive	a	b
	Negative	c	d

The overall level of agreement could be represented by the proportion of the total falling in cells a and d. This measure unfortunately turns out to depend more on the prevalence of the condition than on the repeatability of the method. This is

because in practice it is easy to agree on a straightforward negative; disagreements depend on the prevalence of the difficult borderline cases. *Repeatability (for the individual subject)* is usually therefore defined as:

$$\frac{\text{Number agreed positive}}{\text{Number positive to either observer}} = \frac{a}{a+b+c}$$

This measure is largely independent of prevalence. It states the probability, given one positive test, of the second also being positive.

Epidemiological conclusions are more concerned with groups than individuals, and the above measure is less important than an estimate of *observer or test bias*:

$$\frac{\text{Number positive to observer 1}}{\text{Number positive to observer 2}} = \frac{a+c}{a+b}$$

Note that agreed positives are necessarily fewer than the positives for a single observer. Higgins's law states that "*the frequency of any condition is inversely proportional to the number of investigators (or investigations) required to establish its presence.*"

Measuring validity

A sphygmomanometer's validity can be measured by comparing its readings with intra-arterial pressures, and the validity of a thermographic diagnosis of breast cancer can be tested (if the woman agrees) by biopsy. More often, however, there is no sure reference standard. The validity of a questionnaire for diagnosing angina cannot be fully known: the best clinical opinion is subject to observer variation, and even coronary arteriograms may be normal in true cases or abnormal in symptomless people. The pathologist can describe postmortem structural changes, but these may say little of the patient's symptoms or functional state. Measurements of disease in life, whether clinical or epidemiological, are often incapable of full validation.

In practice, therefore, validity may have to be assessed indirectly. In epidemiology two approaches are available. A test which has been simplified and standardised to make it suitable for use in surveys may then be compared with the best conventional clinical assessment. A self-administered psychiatric questionnaire, for instance, may be compared with the majority opinion of a psychiatric panel. Alternatively, a test may be validated by its ability to predict some other finding or event, such as the ability of glycosuria to predict an abnormal glucose tolerance test, or of a questionnaire to predict future illness. Validation, especially by predictive ability, tends to be more difficult and to require much larger numbers than the testing of repeatability.

Analysing validity

The same subjects are classified as positive or negative, first by the survey and then by the reference test, and the findings can then be expressed in a contingency table:

Survey test	Reference test: Positive	Reference test: Negative	Totals
Positive	True-positives, correctly identified = (a)	False-positives = (b)	Total test positives = (a + b)
Negative	False-negatives = (c)	True-negatives, correctly identified = (d)	Total test negatives = (c + d)
Totals	Total true-positives = (a + c)	Total true-negatives = (b + d)	Grand total = (a + b + c + d)

From this table several important statistics can be derived.

Sensitivity—A sensitive test detects a high proportion of the true cases, and this quality is measured here by a/a + c.

Specificity—A specific test has few false-positives, and this quality is measured by d/b + d.

Systematic error—For epidemiological rates it is particularly important for the test to give the right total count of cases. This is measured by the ratio of the total numbers positive to the survey and the reference tests, or (a + b)/(a + c).

Predictive value—This is the proportion of test positives that are truly positive. It is important in screening, and will be discussed further in the article on that subject.

Sensitive or specific? A matter of choice

If diagnostic criteria are stringent there will be few false-positives but the test will be insensitive. Conversely, if criteria are relaxed there will be fewer false-negatives but the test will be less specific. In a recent survey of breast cancer alternative diagnostic criteria were compared in relation to a reference test (positive biopsy). Clinical palpation by a doctor yielded fewest false-positives (93% specificity), but missed half the cases (50% sensitivity). Criteria for "a case" were then relaxed to include all the positives identified by doctor's palpation, nurse's palpation, or *x*-ray mammography: few cases were now missed (94% sensitivity), but specificity fell to 86%.

By choosing the right test and cut-off points it may be possible to get the balance of sensitivity and specificity that is best for the particular study. In a survey to establish prevalence this might be when false-positives just balance false-negatives. In a study to compare rates in different populations the absolute rates are less important, the primary concern being to avoid systematic bias: a specific test is likely to be preferred, even at the price of some loss of sensitivity.

Eventually this series will be collected into a book and hence no reprints will be available from the authors.

WORDS We are lucky that English is becoming an international language in technical and scientific affairs. It lessens the need to learn foreign languages while imposing a corresponding burden on those whose mother-tongue is not English. The established practice of using classical Greek and Latin roots in forming new medical terms gives them widespread intelligibility, as speakers of Romance languages will understand all the Latin-based and some of the Greek-based words, while many of the university-educated speakers of Teutonic languages will have some knowledge of these ancient languages. In coining new medical terms we should, within reason, stick to this custom.

When I see words such as breakdown for analysis, see-through for transparent, set-up for organisation, blow-up for (photographic) enlargement, and hold-up for obstruction—all acceptable in brisk, graphic everyday speech—I begin to wonder whether this trend could be the thin end of the wedge to penetrate medical terminology. But does it matter? See what has happened to the German language and let that be a warning. New words formed from Latin or Greek have often been translated root for root into Teutonic equivalents, thus erecting an unnecessary barrier to international comprehension. A few examples follow, with the centre column showing a literal translation.

hydrogen	water-stuff	*Wasserstoff*
carbon	coal-stuff	*Kohlenstoff*
television	far-seer	*Fernsehen*
technology	industry-science	*Gewerbekunde*
aluminium acetate	vinegar-acid clay-earth	*Essigsäure Tonerde*
anaemic	blood-poor	*blutarm*
duodenum	twelve-finger-bowel	*Zwölffingerdarm*

I have heard that mitral stenosis became *Mitralstenose* only by the skin of a cusp, as it were, having escaped from a recommendation that it be named *Bischofsmützeklappenunzugänglichkeit*, bishop's-cap-(mitre)-valve-inaccessibility.