Statistics in Question

SHEILA M GORE

ASSESSING CLINICAL TRIALS— BETWEEN-OBSERVER VARIATION

“Poor repeatability implies poor validity”¹—as in epidemiology, so too in clinical trials. Unless a satisfactory answer can be given to the question: How reproducible are results by the same observer?—for example, on a second occasion or on repeat specimens—the assessment should be rejected in favour of another that measures the same thing but which is of proved repeatability. This rule applies in pathology (selecting assays or staining techniques), radiography, ultrasonography, and other subjects, and even to recording clinical signs. A sign carries no information if doctors assess it differently when re-examining the patient.

Even when the field has been narrowed to repeatable measurements that are also valid—that is, they measure what they purport to measure—the problem of between-observer variation remains. A doctor eliciting signs in respiratory disease,² a neurologist making a tentative diagnosis of multiple sclerosis,³ a geriatrician assessing stroke rehabilitation,⁴ or an anaesthetist determining fitness for operation⁵ is each making a judgment that might be made differently by another doctor. Variation between observers in a clinical trial can seriously compromise the research findings. The worst example is when all patients who are given drug A are assessed by one doctor, while a different doctor evaluates patients having drug B. The outcome of such a trial is completely uninformative because any result can be accounted for by the difference between observers rather than between drugs. To avoid the problem one can arrange for the same observer to assess all the patients, but this is not always practicable because of considerations such as the efficient running of wards and clinics and the doctors’ duty rota.

At least ensure that the same observer makes serial assessments on a given patient, and when there is more than one observer identify which observer assesses which patient (observer code).

(25) *How can variation between observers be reduced?*

- use an expert panel or reference laboratory
- compare rival assessment schemes in a pilot study
- find out in what way the observer may be biased
- train observers
- standardise the techniques and criteria for making judgments
- consider: how serious is the disagreement?
- randomise patients to observer
- defer the trial
- appoint an assessor to the research team

COMMENT

Observer variation may be tackled by appointing a specialist panel to review all trial diagnoses—pathological, radiological, scintigraphic, etc. Ideally this evaluation should be done before randomising patients to treatment to avoid retrospective exclusion. Pathological, haematological, or biochemical trial results may often be submitted to a reference laboratory. Such results should of course be reviewed blind, with no knowledge of the treatment assignment.

Another approach is to compare rival assessment schemes in a pilot study to find which scheme leads to the greatest consensus among observers. Thus Oldham⁶ showed that classifying chest radiographs by the presence or absence of tuberculosis or by the clinical significance of the shadows could be repeated reliably more often than classifying them by disease activity or inactivity. When designing this type of pilot study remember that patients who have shown dramatic improvement or appreciable deterioration will be correctly identified by most observers on almost any criteria. It is the patients who have shown a partial response who are most difficult to classify. A good assessment scheme leads to consensus on these patients also.

Be suspicious of any rating—such as behaviour rating in schizophrenia⁷ by nurses on a linear scale (scored 0 to 10)—which has not been studied for observer variation. Remember also that consensus alone is not enough: there must be evidence that the assessment is a reliable diagnostic test or measure of recovery. For example, speech therapists independently assessed school entrants⁸ to whom doctors had given a simple speech screening test. In the field trial an accuracy rate of 92% for the 438 children was reported, and the referral rate by the doctors was 14%. Speech therapists subsequently observed or treated 44 children: nine of them had passed the simple screening test.

Understanding in what way the observer may be biased may reduce observer variation. It has been suggested that an unconscious bias towards lowering or raising a patient's blood pressure is created by making an arbitrary division between normal and high blood pressure⁹—likewise the awareness of risk factors, such as obesity, or a threshold for clinical trial entry. Wilson *et al*⁵ estimated the degree of optimism and pessimism shown by 10 anaesthetists in assessing whether patients were fit for elective surgery. Another aspect of measurement error is that sometimes an observer may prefer even values when instruments such as manometers and applanation tonometers are marked off at even pressures; and a preference for the terminal digits 0 and 5 has also been noted. The tendency to avoid the extremes on any scale is illustrated by this example (for which I thank Dr S J Daldry): a class of 105 statistics undergraduates were asked to write down an integer in the range 0 to 9. The histogram below is certainly not consistent with uniform preference. In particular, the numbers 0 and 9 were avoided.

105 undergraduates select a number in the range 0–9

Training and practice reduce inter-observer variation. In the clinical assessment of stroke⁴ observer agreement was increased from 55% at the first attempt after receiving instruction and definition to 68% at the third test (estimated from 84 assessments). Observers who differ in experience are likely to differ systematically in their assessments on that account. Reports from painstaking assessors will be subject to less random error.

Inter-observer variation can be limited by standardisation—that is, by adhering to particular routines, such as agreed guidelines for physical examinations and selection of diagnostic

tests, specified reagents for use in laboratory assessments, and carefully worded questions asked in a given order. In ophthalmology ocular tension is measured systematically first in the right eye and then in the left; visual acuity is measured in standard lighting conditions and at a prescribed distance from the Snellen chart; and patients are encouraged to read as far down the chart as they can. The standardisation of blood pressure measurement has been well described⁹ but has not been adequately reported in medical journals. Two-thirds of 96 papers relating to blood pressure that had been published in four prestigious general medical journals failed to report on one or more of the following¹⁰: the type of equipment used; the diastolic end-point; the number of readings per measure-

The case for standardisation has not been made strongly enough by statisticians or editors

ment. Standardisation was in doubt in 15 out of 38 clinical trials that I reviewed from the *Lancet* (July to December 1977).

In *Measurement in Medicine*¹¹ Oldham distinguished between qualitative and quantitative disagreement. The latter is disagreement on the intensity of the evidence—for example, in support of a diagnosis of multiple sclerosis—and not on the interpretation—for example, the presence or absence of a clinical sign. The figure shows that certain disagreements (shaded) have less serious implications than others in the diagnosis of multiple sclerosis by two neurologists,³ and these instances of observer variation may be considered more leniently.

Diagnostic classification for multiple sclerosis for 149 patients by two neurologists (observers)

		Neurologist B				Total	Proportion
		Certain	Probable	Possible	Doubtful		
Neurologist A	Certain	38	5	0	1	44	0.30
	Probable	33	11	3	0	47	0.32
	Possible	10	14	5	6	35	0.23
	Doubtful	3	7	3	10	23	0.15
Total		84	37	11	17	149	
Proportion		0.56	0.25	0.07	0.11		

There are other solutions to between-observer variation. By randomising patients to the observers all of the different types of patients should be proportionately represented in the sample that is assessed by each observer. Each report should have an observer code or signature on it to identify which observer has assessed which patient. Subsequent observations on a particular patient must be made by the observer who recorded the patient's first result to avoid confounding of observer and treatment effects. I was told the story of the senior house officer who after having examined a patient compared his finding with the consultant's notes and was alarmed by the apparent deterioration. The consultant reviewed the case and pronounced an improvement. Beware between-observer variation. It may be necessary to defer the clinical trial until a better measurement scheme has been worked out or to appoint someone to do all the trial assessments.

(26) In what way do the observations by clinician C differ from the others?

—systematic error by observer C

Circum-corneal hyperaemia (scored 0,1,2,3,4) by four ophthalmologists												
Ophthalmologist	Patient											
	1	2	3	4	5	6	7	8	9	10	11	12
A	3	2	2	2	2	2	1	2	2	1	1	2
B	3	3	2	2	3	2	2	2	2	2	2	2
C	4	4	3	3	4	4	2	3	3	3	3	4
D	3	3	1	2	2	2	1	3	3	1	2	2

COMMENT

Circum-corneal hyperaemia (only for the lower half, avoiding the area of the cataract incision) was defined as: 0, normal; 1, conjunctival oedema; 2, conjunctival oedema + small venous engorgement; 3, conjunctival oedema + large venous engorgement; 4, unable to define underlying sclera.

Observer C systematically scored patients higher for circum-corneal hyperaemia than the other ophthalmologists did. Observer B scores in the middle of the range.

(27) Comment on the assessment of iris hyperaemia by the four ophthalmologists.

—poor agreement

—carelessness (code 9)

Iris hyperaemia (scored 0,1,2,3,4) by four ophthalmologists												
Ophthalmologist	Patient											
	1	2	3	4	5	6	7	8	9	10	11	12
A	1	0	0	0	0	3	1	2	0	0	0	1
B	1	0	1	4	1	1	1	1	4	1	0	1
C	4	0	0	0	0	4	0	4	9	0	0	0
D	3	3	1	1	2	2	1	2	2	1	2	2

COMMENT

Iris hyperaemia was coded as follows: 0, nil; 1, proximity of iridectomies only—that is, less than 90°; 2, at least 90° and less than 180°; 3, at least 180° and less than 270°; 4, all quadrants.

The ophthalmologists vary unacceptably in their scoring of iris hyperaemia. Observer C uses only the extremes of the scale and introduces a spurious code—9—which has not been defined; observer D avoids the extreme codes. There is no

patient on whom there is total agreement, and only three for whom the difference between the highest and lowest score is one. Observer variability is too large for this assessment to be of value in a clinical trial.

(28) Observer variation in the clinical assessment of stroke was studied by Garraway *et al.*⁴ Twelve patients were examined by consultants A, B, C, and D. The order of examination (figure) was balanced with respect to carry-over. Explain what a carry-over effect due to consultant A in the assessment of mental function means.

—consultant A leaves patients more (or less) able to answer the mental function questions than he found them

Examination	Patient											
	1	2	3	4	5	6	7	8	9	10	11	12
1	A	B	C	D	A	B	C	D	A	B	C	D
2	B	A	D	C	C	D	A	B	D	C	B	A
3	C	D	A	B	D	C	B	A	B	A	D	C
4	D	C	B	A	B	A	D	C	C	D	A	B

COMMENT

The design selected by Garraway *et al.*⁴ was balanced for order of examination because patients might tire in an afternoon during which they had been examined four times. It was also balanced for carry-over from one examiner to the next, so that consultant A was followed equally often by consultants B, C, and D. Garraway *et al.* found that whoever examined after consultant A scored mental function highly. An explanation was sought. It transpired that consultant A, a kindly man, put patients at their ease by telling them the answers to the questions on which they had failed. The information was retained over the short interval to the next examination but did not affect later assessments.

A carry-over from consultant A occurs if consultant A leaves patients more receptive to examination—or less receptive—than he found them.

The figure under question 25 showing observer variation between two neurologists has been reproduced from J R Landis and G G Koch. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159-74, with permission of the Biometric Society.

References

- Rose G, Barker DJP. Repeatability and validity. *Epidemiology for the uninitiated*. London: British Medical Association, 1979.
- Smyllie HC, Blendis LM, Armitage P. Observer disagreement in physical signs of the respiratory system. *Lancet* 1965;ii:412-3.
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159-74.
- Garraway WM, Akhtar AJ, Gore SM, Prescott RJ, Smith RG. Observer variation in the clinical assessment of stroke. *Age and Ageing* 1976;**5**:233-40.
- Wilson ME, Williams NB, Baskett PJF, Bennett JA, Skene AM. Assessment of fitness for surgical procedures and the variability of anaesthetists' judgments. *Br Med J* 1980;**280**:509-12.
- Cochrane AL, Garland LH. Observer error in the interpretation of chest films—appendix C (PD Oldham). *Lancet* 1952;iii:505-9.
- Yorkston JJ, Gruzeliier JH, Zaki SA, Hollander D, Pitcher DR, Sergeant

- HGS. Propranolol as an adjunct to the treatment of schizophrenia. *Lancet* 1977;ii:575-8.
- ⁸ Rigby MJ, Chesham I. A trial speech screening test for school entrants. *Br Med J* 1981;282:449-51.
- ⁹ O'Brien ET, O'Malley K. ABC of blood pressure measurement—the observer. *Br Med J* 1979;iii:775-6.
- ¹⁰ Lehane A, O'Brien ET, O'Malley K. Reporting of blood pressure data in medical journals. *Br Med J* 1980;281:1603-4.

- ¹¹ Oldham PD. Variation of observers' judgements. *Measurement in medicine*. London: English Universities Press, 1968.

Sheila M Gore, MA, is a statistician in the MRC Biostatistics Unit, Medical Research Council Centre, Hills Road, Cambridge CB2 2QH.

No reprints will be available from the author.

Dealing with the Disadvantaged

Problems of short stature

I—Adults

MARY LINDLEY

Adults of restricted growth are as varied in their personalities, occupations, and medical concerns as any other adults. They expect to be treated as such, and not to be mistaken for children, clowns, or creatures from another planet. Whether patients are 2.38 m (4 feet 10 inches) or 1.34 m (3 feet 5 inches), they naturally merit the same courteous treatment afforded to other patients, being recognised as individuals rather than interesting cases of rare syndromes. But their rarity may make special demands on a doctor's skill. Even if achondroplasia is a memory from student days, it is likely that many doctors have never come across any of the other forms of restricted growth. So when a very small person appears for a consultation, a doctor may have much to learn.

Understanding the problems

A general practitioner may find that a patient has already been diagnosed accurately, knows a great deal about the condition causing his restricted growth, and is in the regular care of a consultant. In such a case the patient will be grateful to feel that the general practitioner is interested in his progress and understands when special care may be needed. While patients of restricted growth mostly require only routine care, there is always the possibility that an apparently straightforward problem, such as chest infection, obesity, or a broken bone, is associated with or could be aggravated by the cause of the restricted growth. The patient's own knowledge may be valuable in such circumstances, and doctors should be prepared to listen. When referral is necessary, it is helpful to make sure that the consultant knows the full details of the restricted growth. When the underlying cause of the restricted growth is unknown and the patient understands little about his condition, it is worth suggesting that he should seek an accurate diagnosis so as to understand any special problems when they might arise. If needed, a specialist can be recommended by the medical committee of the Association for Research into Restricted

Growth (ARRG), a British self-help group that has been active for the past 10 years.

Family planning may entail more discussion than usual. Before deciding whether to have children, people of restricted growth should have the opportunity to seek the advice of a genetic counsellor so as to know what to expect. Much confusion and distress may be prevented by a clear explanation at this stage. Patients may worry unnecessarily if told that they "have bad genes," with no attempt to explain the simple link between a defective gene, a defective enzyme, and the reality of an inherited disease, whether the precise defect is known or not. A woman of restricted growth who becomes pregnant may require special antenatal care. On the other hand, adoption has been shown to be perfectly possible and may be worth pursuing.

Apart from medical care, people of restricted growth may need a doctor's support in seeking help with daily problems. They may, for example, wish to apply for a parking permit from the local authority or a mobility allowance from the Government, or they may wish to have financial help in modifying their homes. Such help may be the key to an independent life, enabling a very small person to play a full role in society. Those who have never thought of making use of some of the opportunities available to them may appreciate a doctor's encouragement and advice.

In the consulting room

When talking to patients of restricted growth, a doctor should remember above all that they are ordinary people who happen to be very small. The following simple recommendations may help to smooth the encounters.

- (1) If you have several chairs in the room, invite the patient to choose the most comfortable.
- (2) Talk to the patient as to any other adult, judging intelligence according to conversation, knowledge, questions asked, and so on and not according to appearance.
- (3) Do not ask, "Do you work?" ask instead, "What is your job?" Suppress your surprise when told that your four-foot-tall patient is a television producer or a managing director.
- (4) Do not refer to dwarfs and midgets in conversation but to people of restricted growth or short stature.
- (5) If practical difficulties arise and, for example, the patient

Mary Lindley died on 7 May 1981.

London SE18 1BS

MARY LINDLEY, MA, chairman, Association for Research into Restricted Growth