PAPERS AND SHORT REPORTS

Normality and reliability in the clinical assessment of backache

GORDON WADDELL, CHRIS J MAIN, EMYR W MORRIS, ROBERT M VENNER, PETER S RAE, SAMIR H SHARMY, HELEN GALLOWAY

Abstract

Backache is common yet its routine medical assessment is imprecise, unreliable, and poorly interpreted. Reproducibility studies on 475 patients improved the reliability of clinical interview and examination in backache, while studies of 335 normal subjects defined the limits of normality. Assessment of nerve function was found to be reliable but assessment of the back itself had to be considerably modified and examination improved by incorporating actual measurements. The validity and clinical utility of the information were analysed to determine the minimum amount of information which should be collected to permit clear diagnosis and management.

Introduction

Medical management of the individual patient always has been, always should be, and indeed can only safely be based on the clinical history and physical findings. This assumes that the information obtained from interview and examination provides a reliable measure of abnormality, distinguishes normal from abnormal, and permits valid interpretation. In routine clinical assessment of backache all of these assumptions may be questioned. The low reproducibility of clinical findings among even experienced doctors has been documented in several clinical conditions,¹⁻⁵ including backache,⁶ though it has not received the wide attention it deserves. The few attempts to improve data collection techniques in backache⁶⁻⁸ have generally

been of limited scope or concentrated on specific aspects of the assessment.

In this report a large number of sequentially modified studies (see table I) progressively and systematically improved the reliability of all aspects of both physical and psychological assessment in backache while parallel studies of normal subjects defined the limits of normality. The aim was to improve the quality of clinical information in backache to give a sounder basis for diagnosis and management.

TABLE I—Groups of patients studied

Data collected	Groups and numbers of subjects		Type of reliability study		
	Normal subjects	Reproducibility in patients with backache	Patient/patient	Patient/doctor	Doctor/doctor
Clinical history	—	{ 25 25 52 25	+	+	
Pain scale	—	25	+	+	
Pain drawing	—	{ 25 25	+		
Disability	{ 40* 60† 40*	30		+	
Inappropriate symptoms	{ 40* 60† 60†	20			+
Physical impairment	{ 30 55 70	{ 25 25			+
Inappropriate signs	{ 50 30	50‡§			+
Full clinical history, examination, diagnosis	—	30			+
Psychological questionnaires	—	40	+		
Psychological ratings: Doctor/doctor	—	{ 50‡§ 30			+
Doctor/psychologist	—	50			+
Total	335	475	190	155	230

*Same groups.

†Same groups.

‡Same groups.

§Canadian Workmen's Compensation Board.*

||Separate report by G Waddell *et al* ("Psychological assessment in backache"; in preparation).

Orthopaedic Department, Western Infirmary, Glasgow G11 6NT

GORDON WADDELL, MD, FRCS, consultant orthopaedic surgeon

CHRIS J MAIN, MA, MPHIL, senior clinical psychologist

EMYR W MORRIS, MB, FRCS, spinal research fellow

ROBERT M VENNER, MB, FRCS, consultant in accident and orthopaedic surgery

PETER S RAE, MB, FRCS, senior orthopaedic registrar

SAMIR H SHARMY, DS(ORTH), MD(ORTH), consultant orthopaedic surgeon

HELEN GALLOWAY, medical student

Subjects and methods

A total of 475 patients with backache and 335 normal subjects were studied (table I). The results in each group were analysed and the techniques modified and improved before proceeding to the next group. The techniques of history taking and examination presented here are mainly the final versions developed by the end of these studies. For the clinical history the patient completed a questionnaire while awaiting consultation. The doctor used the same questionnaire independently as a basis for the clinical interview and the patient repeated the questionnaire 24 hours later (return rate over 90%). Physical examination and diagnosis were recorded independently by two doctors. Psychological questionnaires were completed by the patient and repeated 24 hours later. Psychological data were recorded by two doctors independently and also by a doctor and a psychologist in separate interviews. In all studies the order of examiners was randomised. Most patients were reinterviewed or re-examined immediately except for a small subgroup, who were re-examined after 24 hours. There was no detectable difference caused by the order of examiners or by this time interval between examinations.

All studies of both normal subjects and patients with backache were carried out on equal numbers of white men and women. All patients were aged 20-55 years—the common age of presentation of backache—and normal subjects were stratified over the same age range. Except for one group of 50 patients studied in a Canadian Workmen's Compensation Board⁹ (table I), all studies were carried out in a largely industrial area of western Scotland. All patients were referred to an orthopaedic outpatient clinic with backache. The clinical history was studied in newly referred patients who had never been admitted for backache and had no spinal abnormality or psychiatric history. Examination techniques were studied in selected patients with some detectable lumbar or neurological abnormality. Psychological features were studied both in patients newly referred from family doctors and in patients with long, complex histories of failed treatment referred to a regional problem back clinic. Normal subjects were patients with minor hand injuries, hospital visitors, and hospital staff. No normal subject had a history of time off work or consultation with a doctor about backache, a psychiatric history, or a current illness. Relatives of patients with backache were excluded. Subjects with a particular physical fitness background such as physiotherapists were excluded from studies of physical impairment. Hospital staff and subjects with medical knowledge or who had close contact with backache or psychiatry were excluded from psychological studies.

All examiners were orthopaedic registrars or consultants working together in the back clinic. Before each study there was detailed discussion and agreement on the exact information to be gathered and its format, qualifications, and exclusions. Four examiners participated in the reproducibility studies, though for most study groups only two examiners were compared. Five examiners participated in the normal studies, with only minor variations in their findings.

Finally, the complete assessment system including all aspects of history, full physical examination, and clinical and psychological diagnosis was carried out independently by two observers (GW and RMV) in 30 selected patients referred to the problem back clinic.

CLINICAL HISTORY

Analysis was possible only after specifying the exact information to be gathered and the form in which it was to be recorded. A questionnaire originally developed in the West of Scotland Neurosurgical Unit (W B Jennett, personal communication) was simplified so that it could be filled in by patients without help. The same questionnaire was used as a guide to the information to be gathered in the medical interviews and to record that information. Clinicopathological distinctions between acute and chronic, referred and root leg pain, and disc prolapse and degeneration had to be clarified in the design of certain questions.¹⁰ Information was gathered about the onset of symptoms, the time course and pattern of pain, anatomical localisation of pain, severity and effects of pain, mechanical characteristics of pain, and effect of treatment (table II, first column). The format of the answers was presented either as a dichotomous Yes/No, as a forced-choice classification, or as an open-ended question or scale (table III). Four successive versions of the questionnaire were used. Table IV gives an example of successive formats tried for one question. The first and second columns of table II summarise the final questionnaire.

Several versions of a pain scale¹¹ were incorporated. A pain drawing¹² was used to analyse anatomical patterns of pain. Separate questions were asked about functional disability in activities of daily living. A simple doctor-administered scale of nine dichotomous items (G

Waddell, C J Main, report in preparation) was tested and also found to correlate closely with a previously described disability questionnaire of 10 forced-choice items.¹³

Information on behavioural and psychological features was obtained by questionnaires measuring personality traits, depression, somatic

TABLE II—Final format and results of clinical history questionnaire

Question	Final format*	% Agreement patient/patient
Onset:		
Sudden/gradual	C	79
Accident/pregnancy/neither	C	80
Bend-lift/blow/neither	C	79
Time:		
Duration of back trouble	O	96
Duration of present attack	O	74
No of separate attacks in past year	O	80
Duration of attacks, days/weeks	C	96
Longest spell between attacks	C	75
In morning back better/worse	C	72
By evening back better/worse	C	92
Time off work now	O	83
Time off work in past year	O	82
Anatomical patterns:		
Pain ever below knee	D	100
Pain ever into foot	D	92
Numbness/pins and needles below knee	D	95
Back pain/leg pain worse	C	88
Pain drawing (anatomical pattern)	O	90
Hip ever stick out to one side	D	84
Severity:		
Pain scale (within 5 mm)	O	77
Time off work now	O	83
Time off work in past year	O	82
Ever confined to bed	D	73
Between attacks forget about/careful of back	C	92
Disability—nine activities of daily living	D	73-90
Physical actions affecting pain:		
Pain worse bending forward/arching back	C	100
Pain better/no different/worse:		
Twisting	C	92
Lying flat	C	88
Lying curled on side	C	84
Walking affects back pain/leg pain	C	84
Does your back ever "lock"?	D	88
Relief from conservative treatments:		
Considerable/slight/no change/worse	C	84
Mean		85.5 ± 7.55

*D = Dichotomous. C = Forced choice. O = Open-ended question or scale.

NB: In all instances $p < 0.01$; and, in almost all, reliability coefficient $K > 0.60$.

TABLE III—Comparison of different formats for clinical history questions

Type of information	% Reproducibility (patient/patient)		
	Dichotomous Yes/No	Classify forced choice	Open-ended scale
Onset	96	80	—
Time patterns	—	76	96
Anatomical patterns	96	60	90
Severity	92	57	77
Physical activities affecting pain	92	88	—
Treatment	83	84	—

TABLE IV—Importance of design in forced-choice questions

	% Reproducibility	
	Patient/patient	Patient/doctor
(a) In the past two years how many attacks have you had of:		
Back pain alone (none, 1-3, 4-6, >6)?	60	44
Leg pain alone (none, 1-3, 4-6, >6)?		
Back and leg pain together (none, 1-3, 4-6, >6)?		
(b) How many attacks have you had in the past year?		
No previous attacks	86	58
1-3		
4-6		
More than 6		
Continuous		
(c) How many separate attacks of back trouble have you had in the past year?	88	80
Where is your pain worse?		
In your back?	96	88
In your leg?		
Back pain and leg pain about the same?		

awareness, hypochondriasis, invalidism, and pain behaviour (G Waddell *et al*, report in preparation) supplemented by psychological and clinical interviews. The clinical interviews included a group of inappropriate or magnified descriptions of symptoms of backache which were statistically separable from and independent of the physical findings and which appeared to have psychological rather than physical significance (G Waddell *et al*, report in preparation).

PHYSICAL EXAMINATION

Examination provides information about the lumbosacral spine, about lumbar and sacral nerve function, and also about the patient's illness behaviour. Information available from examination of the back itself is limited. Some techniques had to be considerably modified and incorporate actual measurement to give reliable findings. The only instrument used was an ordinary centimetre tape-measure in a spring-loaded case.

Spinal deformity and lumbar movements were measured with the patient standing erect but relaxed with the feet together and with the back to the examiner. Loss of lordosis was measured by tensing a tape-measure between the thoracic and sacral prominences in the midline. If the maximum distance between the concavity of the lumbar lordosis and the tape-measure was less than 2 cm estimated by eye against the centimetre markings of the tape, then lumbar lordosis was considered to be reduced. A sciatic list where the shoulders were offset from the pelvis was distinguished from a true structural scoliosis with an element of rotation and compensatory curves above and below. A list was measured by dropping the tape as a plumbline from the lower thoracic convexity and estimating in centimetres how far the tape hung to the side of the gluteal cleft.

Lumbar flexion was measured using the skin-marking technique of Moll and Wright.⁷ A mark was made on the skin in the midline at the level of the dimples of Venus, which approximates to L5. A second mark was made in the midline 10 cm above and a third mark made 5 cm below this initial mark. The 15 cm between upper and lower marks was taken as the resting measure. The patient was then asked to bend straight forwards and reach down with both hands as far as possible towards the toes, and the distance between the upper and lower marks was remeasured. The difference between the two measurements was taken as the measure of lumbar flexion and was considered to be restricted if less than 5 cm.

A "catch" was recognised as a break in the normal lumbar-pelvic rhythm on returning from the flexed to the erect position.

Lateral flexion was measured in the midaxillary line between two skin marks 10 cm apart at the levels of the two upper marks used for measuring lumbar flexion. The patient was then asked to lean straight across to the opposite side with the fingers reaching down the side of the thigh, and the increase in distance between the two marks was measured. The normal increase was at least 3 cm. This technique was not valid in obese patients with thick rolls of subcutaneous fat.

Tenderness was examined in the standard manner with the patient lying prone and relaxed. Lumbar tenderness was taken as localised tenderness to firm, deep palpation over L1-S1 or over the para-vertebral muscles within 5 cm of the midline. Gluteal tenderness was observed similarly if localised within the gluteal muscle mass. Widespread and non-localised tenderness was regarded as inappropriate (see below).

Straight-leg raising was estimated by passive raising of the leg with the knee held straight by the examiner and the patient lying supine and relaxed with the head on a single pillow. The onset of pain was found to be a less reliable measure, and the maximum tolerated straight-leg raising was used. Patients could distinguish between straight-leg raising limited by hamstring tightness, back pain, and radiating leg pain.

Neurological deficit was examined in the standard manner. Diminished sensation was tested with light touch and pinprick, emphasising minor differences between the two legs. Localised dermatomal changes were distinguished from inappropriate regional changes (see below). Motor weakness was estimated by comparison with the normal leg and graded MRC 0-5. Localised myotomal weakness was again distinguished from inappropriate regional giving way.

Inappropriate or magnified responses to physical examination consisted of signs elicited on physical examination which did not fit accepted anatomical and neurological patterns but appeared to have a predominantly non-physical basis. These have been described⁹ as superficial and widespread non-anatomical tenderness, pain on simulated rotation or axial loading, straight-leg raising improving

with distraction, regional sensory changes, regional giving way, and overreaction to examination.

DIAGNOSIS

The main clinical diagnosis was made with use of a flow chart¹⁰ (fig). The diagnosis of a prolapsed disc was based on satisfying three of the following criteria: (a) predominant complaint root pain, (b) root irritation signs, (c) root-compression signs, (d) matching metrizamide myelography.

For direct comparison doctors and psychologist used their separately gathered information to judge non-physical and psychological features on the same scales. The importance of psychosocial features in the clinical presentation and disability was rated as minor, moderate, or major. Psychological suitability for surgery (independent or whether or not surgery was indicated physically) was rated as suitable, doubtful, or unsuitable.

DATA ANALYSIS

The most widely used method of measuring and describing reproducibility is the percentage agreement between two independent observations. This figure includes the cases in which both observations are positive and also those in which both are negative. The main limitation of such a percentage agreement is when the incidence of positive observations is low. If, for example, only 10% are positive then a percentage agreement of over 80% may be achieved by chance even when there is almost complete disagreement about the positive observations. Hence even a high percentage agreement requires checking, and a reliability coefficient (K) was calculated based on an unweighted kappa coefficient which emphasised the positive findings. Random answering will give 50% agreement on a dichotomous question, 25% agreement on a four-alternative forced-choice, and 10% agreement within five points on a 100-point scale. A probability level was therefore calculated for the likelihood of obtaining the observed level of agreement by chance. To confirm a 70% agreement K was required to be more than 0.30 and p less than 0.05.^{14 15} In clinical practice reproducibility is frequently less than 70%,⁶ though for scientific analysis the aim should be over 90%.

By the end of these studies most of the information described could be improved to a reproducibility of 80-90% ($K > 0.05-0.7$; $p < 0.01$). Due to statistical limitations a Pearson correlation coefficient was substituted for K in open-ended questions.

Results

Table II summarises the best reproducibility achieved for patients repeating the final version of the history questionnaire. Patient/doctor agreement on the same questions was closely comparable, with an identical average agreement of 86%. Table III compares the different formats tested for these questions, and table IV illustrates the importance of design for forced-choice questions. (The final version of the history questionnaire is available on request.)

Table V gives the incidence of physical findings in normal subjects, which was used to decide the limits of normality which would best discriminate between patients with backache and subjects without symptoms. Table V also gives full reproducibility statistics for the final examination techniques described above. Several commonly mentioned signs were so unreliable that they were completely discarded—for example, muscle spasm and guarded movements. Earlier trials showed

TABLE V—Normality and reproducibility of physical examination

	Normal subjects (%)	Agreement	Reliability coefficient K	Significance (p)
Operative scars	0	100	1.0	<0.001
Loss of lordosis (<2 cm)	5	88	0.71	<0.001
List (>1 cm)	3	78	0.53	<0.05
Limited flexion (<5 cm)	3	91	0.81	<0.001
Catch	2	92	0.82	<0.001
Limited lateral flexion (<3 cm)	2	70	0.41	<0.05
Tenderness { Lumbar	8	100	1.0	<0.001
Buttock	4	81	0.63	<0.001
Straight-leg raising <75°	2	77	0.56	<0.001
Root irritation signs	0	86	0.45	<0.01
Root compression signs	0	93	0.62	<0.001

that simple clinical "eyeballing" detected only gross examples of loss of lordosis, list, limited flexion, and limited lateral flexion and therefore the measurements described above were substituted. Although there was good agreement on the presence or absence of localised tenderness (79%), anatomical localisation of tenderness was so unreliable as to be clinically valueless (30%). Neurological examination was much more reliable than examination of the spine itself.

Diagnostic classification of the main clinical presentation as back pain/referred leg pain/root pain/?pathology was 93% reproducible between two observers ($K=0.77$; $p<0.001$). Diagnostic criteria for a prolapsed disc were also 93% reproducible ($K=0.68$; $p<0.001$).

Table VI gives full reproducibility statistics for the psychological and behavioural assessment. In all aspects good discrimination was achieved from normal subjects (G Waddell *et al*, report in preparation).

For simplicity tables II-IV give only percentage agreements, though all figures quoted had corresponding K and p values.

TABLE VI—Reproducibility of psychological and behavioural assessment

	Agreement	K	p
Pain drawing: "magnification"	92	0.84	<0.001
Inappropriate symptoms	80-90	0.58-0.67	<0.01
Inappropriate signs	78-86	0.55-0.71	<0.001
Psychological questionnaires:			
Depression	—	0.94*	<0.001
Somatic awareness	—	0.89*	<0.001
Ratings:			
Psychosocial component to presentation:			
Doctor/doctor	83	0.71	<0.001
Doctor/psychologist	68	0.44	<0.001
Psychological suitability for surgery:			
Doctor/doctor	72	0.51	<0.001
Doctor/psychologist	64	0.27	<0.05

*Pearson correlation coefficients.

Discussion

Clinical information should meet the criteria of normality, compliance, reliability, validity, utility, and cost effectiveness. The limits chosen for normality should give a low incidence of positive findings in normal subjects and clearly discriminate between patients with symptoms and normal subjects. They should give acceptable reproducibility and clinically meaningful information. The limits of normality for physical examination suggested here meet those criteria (table V).

Questions must be simple, clear, unambiguous, and whenever possible phrased in ordinary everyday English, avoiding medical jargon and inappropriate cultural assumptions. The information sought must be within the patient's knowledge and the patient must be able and willing to answer the question. Some North American psychological questionnaires are so unacceptable that only 70% of British patients can or will complete them adequately (G Waddell *et al*, report in preparation). If such "compliance" is less than 90% then the information obtained may be biased. If positive findings are less than 5% then the effort of collection is high and the information must be clinically very useful to be cost effective.

Information should be reliable—the patient should answer consistently on different occasions, and different observers should independently obtain the same answer. Patient/doctor reproducibility comparing a self-report with a structured interview allows the doctor to check and amplify how the patient has understood and answered the question and may show that what doctors mean by a particular term or question can be different from what patients understand from it.¹⁶ Thus reproducibility may depend on the reproducibility of the patient's recall and on the doctors' administration of the question, the patient's understanding of the question, the way in which the patient or doctor decides to classify the answer, and observer bias.

Clinical information may be obtained from the patient, from relatives, by direct observation of the patient, or by laboratory investigations, and each of these has different error and bias. The history of the presenting illness, descriptions of present symptoms, physical signs, psychological assessment of illness behaviour, and laboratory investigations are again subject to differing errors. Memory is unreliable, and current events will be described more reliably than information from more than one or two years before. Clear-cut factual information is more likely to be reliable and valid than asking the patient to interpret events or ascribe meaning or causation. Answers may be obtained as a simple dichotomous Yes/No, as a forced-choice classification, or as an open-ended question or scale (table III). Yes/No questions are often answered most reliably but the information obtained is severely limited. When appropriate, open-ended questions give good reliability, particularly for duration or anatomical patterns. Scales are less reliable for questions of judgment such as severity. Where dichotomous questions are inadequate and open-ended questions or scales inappropriate, forced-choice questions give more useful information. This particularly applies to ratings or judgments. Forced-choice questions require very careful design (tables III, IV). The options should be clear, logical, and mutually exclusive and designed to give a fairly even distribution of answers. Qualifications and exclusions should be avoided. For many types of clinical rating or judgment none/minor/moderate/major will be the most useful classification.

Data must be valid—it must actually report or measure the event it purports to assess. If a question is misunderstood the answer may obviously be invalid. Pathological interpretations of clinical data may require independent confirmation. X-ray studies of lumbar flexion show that how far the fingers reach to the ground is a compound measure of lumbar flexion, hip flexion, hamstring tightness, and psychological influences. Some patients with ankylosing spondylitis with a solid lumbar fusion can touch the ground. Fingers to the ground is therefore an invalid measure, and the skin-marking technique is much more valid and specific and correlates well with radiological measurements of lumbar flexion.⁷

The usefulness or utility of clinical data lies in diagnosis and management. Many items of clinical information simply duplicate other data and confuse analysis. Moreover, the greater the amount of information gathered the lower is its average reproducibility, while careful concentration on a limited quantity of information improves reproducibility.⁶ Redundant information which adds nothing to diagnosis and decision making should be omitted. The selection of which information to retain will depend on its relative usefulness balanced against the effort and risk of collecting it—its cost effectiveness.

The implication of reliability studies is considerable both for medical practice and for research. There is obvious scope for improvement in the standard of clinical information in backache. Though a great deal of work is required in both the choice and format of the information to be gathered, the final form can be simple, quick, and applicable to routine clinical practice. The above examination of the lumbar spine is just as quick as traditional examination techniques and entirely practicable in a busy clinic. The implications for medical research are even more imperative. Much clinical information is so unreliable as to be

scientifically unusable. Any medical information which is to be the basis of scientific analysis must therefore be checked—and improved—by reproducibility studies on a representative group of patients *before* starting the main study.⁵ The standard of clinical data collection must be improved if it is to match increasingly sophisticated radiological and laboratory investigations and if clinical history and examination are to remain the foundation of medical practice.

We thank the Medical Research Council and the Back Pain Association for Mr E W Morris's fellowship and for their generous support.

The final version of the history questionnaire may be obtained from Mr Gordon Waddell.

References

- ¹ Cochrane AL, Chapman PJ, Oldham PD. *Lancet* 1951;ii:1007.
- ² Kilpatrick GS. *J Med Educ* 1963;38:38.
- ³ Gill PW, Leaper DJ, Guillou PJ, Staniland JR, Horrocks JC, de Dombal FT. *Methods Inf Med* 1973;12:108.
- ⁴ Hall R, Horrocks JC, Clamp SE, De Dombal FT. Observer variation in assessment of results of surgery for peptic ulceration. *Br Med J* 1976;ii:814-6.
- ⁵ Yorkshire Breast Cancer Group. Observer variation in recording clinical data from women presenting with breast lesions. *Br Med J* 1977;iii:1196-9.
- ⁶ Nelson MA, Allen P, Clamp SE, de Dombal FT. *Spine* 1979;4:97.
- ⁷ Moll J, Wright V. In: Jayson M, ed. *The lumbar spine and back pain*. London, 1976:ch 6.
- ⁸ Million R, Wall W, Nilsen K, Baker R, Jayson M. *Spine* (in press).
- ⁹ Waddell G, McCulloch JA, Kummel E, Venner RM. *Spine* 1980;5:117.
- ¹⁰ Waddell G. *Hospital Medicine* (in press).
- ¹¹ Sternbach RA. *Pain: a psychophysiological analysis*. New York: Academic Press, 1968.
- ¹² Ransford AO, Cairns D, Mooney V. *Spine* 1976;1:127.
- ¹³ Fairbank JCT, Davies JB, Couper J, O'Brien JP. *Physiotherapy* 1980;66:271.
- ¹⁴ Cohen J. *Educational and Psychological Measurement* 1960;20:37.
- ¹⁵ Fleiss JL, Cohen J, Everitt BS. *Psychol Bull* 1969;72:323.
- ¹⁶ Boyle CM. Difference between patients' and doctors' interpretation of some common medical terms. *Br Med J* 1970;ii:286-9.

(Accepted 25 March 1982)

Hypertension: comparison of drug and non-drug treatments

GAVIN ANDREWS, S W MACMAHON, ANNE AUSTIN, D G BYRNE

Abstract

Thirty-seven reports of the treatment of hypertension by non-pharmacological means were compared with the results of treatment by standard drug regimens. Treatment by drugs produced the greatest lowering of blood pressure. Treatment by weight reduction, yoga, and muscle relaxation each produced smaller, but appreciable, changes in blood pressure. The effects of meditation, exercise training, blood pressure biofeedback, and salt restriction were inferior to those of the other regimens and were not significantly different to the effects of placebo treatment.

Large comparative trials of pharmacological and non-pharmacological treatments are needed before definite conclusions can be made.

Introduction

Recent trials have shown that using drugs to lower blood pressure in mild hypertension reduces the death rate.^{1,2} While this is heartening, the prospect of asking some 17% of the adult population to take antihypertensive drugs for years is somewhat daunting, particularly with regard to possible unforeseen side effects and the considerable personal and public cost of such treatment. More attention needs to be focused on non-pharmacological interventions^{3,4} but it is not clear from the published

reports just which approach the doctor should take in treating the patient with mild hypertension. A randomised, controlled trial of non-drug and drug regimens would help to clarify this issue, on which only slight evidence exists,⁵ though doctors are already advising patients to use non-drug treatments.⁶

Rational decisions need to be made on the basis of existing reports but it is difficult to review the results of a collection of studies in which research design, subject characteristics, treatment modality, and measures of outcome all differ. Glass⁷ introduced an approach called meta-analysis in which the properties of several studies could be recorded in quantitative terms and descriptive statistics applied to derive an overall conclusion. Thus, reviewing the published works ceases to require the judgment of Solomon and becomes a quasiempirical procedure. We used the meta-analytic technique to review non-pharmacological treatments for hypertension. Three questions were explored: (a) how effective are non-drug treatments compared with standard drug regimens? (b) which of the non-pharmacological treatments produce the greatest reductions in blood pressure? and (c) are the reductions in blood pressure maintained?

Method

We studied the published reports on outcome of non-drug treatment of mild hypertension—that is, diastolic blood pressure >90 mm Hg, systolic blood pressure >140 mm Hg. Studies of drugs and placebos were included for comparison. Treatment studies from which means and standard deviations could not be calculated were excluded. In the meta-analytic technique the effects of treatment are expressed as an effect size⁸ calculated from the difference between mean blood pressures (mm Hg) in treated and control groups, standardised by the variability of blood pressures in the control group. The effect size, a measure of the extent of improvement in standard deviation units, is treated as the dependent variable, the other features of the studies being coded numerically and considered as independent variables likely to contribute to the effect size.

By assuming that the frequency distribution of the measures of outcome—usually blood pressure levels—were normally distributed,

University of New South Wales, Prince Henry Hospital, Little Bay 2036, Australia

GAVIN ANDREWS, MD, FRANZCP, associate professor of psychiatry
S W MACMAHON, BSC, MA, research scientist, school of psychiatry
ANNE AUSTIN, BSC, research psychologist

Department of Psychology, Australian National University, Canberra, Australia

D G BYRNE, PHD, reader