ASSESSMENT OF DRUGS IN OUT-PATIENTS WITH RHEUMATOID ARTHRITIS

EVALUATION OF METHODS AND A COMPARISON OF MEFENAMIC AND FLUFENAMIC ACIDS WITH PHENYL BUTAZONE AND ASPIRIN

BY

R. M. MASON, D. E. BARNARDO*, W. R. FOX†, AND M. WEATHERALL‡

From the Department of Physical Medicine and Rheumatology, the London Hospital, and the Department of Pharmacology, the London Hospital Medical College

Relief of symptoms in patients with chronic rheumatoid arthritis is difficult. Conventional analgesics such as aspirin, paracetamol, and phenylbutazone have limited efficacy and are not free from toxic effects. Any alternative which was more effective, safer, or both would be welcome, but most new remedies owe such success as they achieve to therapeutic optimism which is a powerful but transient potentiator of pharmacological effects. The accurate assessment of new drugs by properly controlled trials is indispensable. Two drugs which have shown promise in laboratory studies and early clinical trials are mefenamic and flufenamic acids (Winder, Wax, Scotti, Scherrer, Jones, and Short, 1962; Goodley, 1963; Young, 1962), derivatives of xylylantranilic acid with actions broadly resembling those of aspirin. We have submitted these drugs to a controlled trial, which has already been briefly reported (Barnardo, Currey, Mason, Fox, and Weatherall, 1966). Since our trial began, several other trials have been reported. Flufenamic acid (600 mg./day) has been found to be less effective than prednisone (Fearnley and Masheter, 1966), and comparable in effect to aspirin (2,700 mg./day) over periods up to 6 months (Simpson, Simpson, and Masheter, 1966) and also comparable with phenylbutazone (300 mg./day) over shorter periods (Rajan, Hill, Barr, and Whitwell, 1967). Mefenamic acid (1,500 mg./day) has been found effective also in osteo-arthritis (Cahill, Hill, Jessop, and Hume Kendall, 1965).

The methods of clinical trials are still capable of much improvement. Particularly in trials on outpatients, the opportunities for undetected errors are very great. Problems of measurement are also considerable, at least in rheumatoid arthritis, in which all of the many measurements which can be made appear to vary somewhat independently of each other (American Rheumatism Association, 1965), so that confidence in any one is limited. One cannot decide which is the best of a number of alternative procedures without testing all of them in parallel. The present trial has been extended considerably beyond the needs of a simple comparison of drugs in order also to be informative about problems of method.

Methods

(1) Patients

All were females over 18 years of age and were selected by the six participating physicians from the attenders at the Department of Physical Medicine and Rheumatology of the London Hospital between October, 1964, and September, 1965. Each had "definite" or "classical" rheumatoid arthritis (Ropes, Bennett, Cobb, Jacox, and Jessar, 1959) for more than one year, was not pregnant nor expected to become so during the trial, and had no history of cardiac failure, hepatic disease, or proven peptic ulceration. Patients sensitive, or reacting unfavourably to, either phenylbutazone or aspirin were admitted and a provision was made as described below to avoid their receiving the drug to which they were sensitive. Patients receiving other well-stabilized treatment such as steroids or antimalarial drugs were not excluded and their treatment with these drugs was continued while they took part in the trial. The purpose of the trial was explained to each patient and anyone who did not wish to take part or who would have had practical difficulties in co-operating was excluded.
ANNALS OF THE RHEUMATIC DISEASES

(2) Design and Procedure

Patients differ considerably from one another. Variation in the severity of any given patient's arthritis is likely to be less than the variation between patients. The trial was therefore designed so that drugs were compared within patients. Each patient was to receive both new drugs and an appropriate standard treatment. The choice of a standard presented some difficulty. Many drugs have been shown to be more effective than a placebo in rheumatoid arthritis, so the use of a dummy would have raised ethical problems. Besides, one object of the trial was to compare the new remedies with standard treatments, of which aspirin and phenylbutazone seemed most appropriate. Some patients are, or believe themselves to be, sensitive to one or other of these drugs. It is for such patients that a new alternative analgesic is particularly needed. The trial therefore had to be arranged so that such patients could be included, without risk of being exposed to the drug to which they were perhaps sensitive. This objective was achieved by designing the trial so that aspirin served as the standard treatment for half the patients and phenylbutazone for the other half. Ordinarily the choice depended on the statistical design, but if a patient was believed to react unfavourably to either drug the physician requested instead a sequence omitting the drug in question. This request was made on four occasions, but the allotted sequence was, in fact, inappropriate and had to be altered only once.

During three consecutive periods, each of 4 weeks' duration, all patients were supposed to receive mefenamic acid, flufenamic acid, and either phenylbutazone or aspirin in capsules of a different colour. Three kinds of coloured capsule were used (blue, blue and white, and red and yellow) and each drug was dispensed equally often in one kind of capsule as another. It was therefore evident to patients that they were receiving different remedies in successive periods, but communication between patients or staff about the efficacy of different-looking remedies (Asher, 1948) would not lead to collective judgements related to a particular drug. Each of the twelve possible orders of treatment was allocated to three patients and the three capsule colours were arranged in a Latin square design over each such group. A complete replicate of the design thus required 36 patients. When a physician admitted a patient to the trial, the pharmacist took note of any specific request to avoid a sequence containing either aspirin or phenylbutazone and then gave the patient the next available suitable trial number. Thereafter the pharmacist dispensed the appropriate capsules, leaving both the patient and the physician unaware of the identity of the treatment.

When patients failed to complete three periods, additional patients were admitted as replacements, who followed the same drug sequence and colour code to maintain the balanced design. Finally, most patients continued in the trial for a fourth period, receiving a second treatment with whichever drug they had considered most effective in the first three periods. This replication was achieved by an appropriate instruction to the pharmacist and did not involve the physician in breaking the code.

The trial proper was preceded by a short pilot study in which the trial procedure was tested and practised and the physicians obtained experience in the completion of the forms designed for keeping records in the trial. Seven patients were admitted to the pilot study, which revealed no appreciable faults of planning. A further 36 patients were admitted; four among these failed to complete three periods of treatment and a corresponding additional number were admitted to make a complete balanced set as designed. While the results from this set were being analysed, further patients were admitted with a view to completing a second complete set of 36. The results on the first 36 indicated that no substantial difference was appearing between treatments, and the second set was therefore not completed. In all 68 patients were admitted, including the seven in the pilot study and eight who completed less than three periods. Analyses of comparison between drugs are mostly based on the balanced set of 36 patients. Other analyses were based on all the patients studied.

Many clinical trials are conducted with fixed dosage of drugs. The conclusions which can be drawn from single fixed-dose studies are very limited. If one drug produces better results than another, the difference may be due entirely to the doses chosen, and an otherwise identical trial with different doses might produce exactly the opposite result. If each drug is administered at two or more levels of dosage, and if the larger dose has larger average effects than the smaller, a reasonable basis of comparison is possible between the dose–response lines for each drug. But it is seldom possible in practice to achieve a graded response, because the range between the apparently ineffective dose and the evidently toxic dose is usually too small.

To avoid this difficulty, doses in the present trial were selected not by an arbitrary weight of drug but by the response of the patient to treatment. Initially the doses prescribed were 720 mg aspirin, 100 mg phenylbutazone, 500 mg mefenamic acid, or 200 mg flufenamic acid, each given three times daily. In order to adjust the dose to each individual patient, all patients were seen at the end of the first week in each period; their progress was briefly reviewed and the physician had the option of decreasing or increasing the daily dose by one third (by altering the frequency of medication to twice or four times daily) if it seemed desirable on grounds of therapeutic response and toxic manifestations. Provision was also made for discontinuation of a drug if worsening of symptoms or toxic effects were severe. In such a case the patient proceeded immediately to the drug planned for the next period. Occasionally a short period without treatment was adopted before resuming the trial sequence. The rigidity of design necessary for unbiased comparison and accurate analysis was therefore combined with appreciable freedom of manoeuvre for the physician in the treatment of patients.

In addition, at the beginning of each period, each patient was given 200 tablets of paracetamol (500 mg).
with instructions to take one or two tablets as necessary every four hours if additional pain relief was required. Any established treatment with antimalarials or steroids was continued in unchanged dosage. Self medication with any other drug was discouraged, and the general practitioners attending the patients were requested not to prescribe any additional remedies.

Three methods of monitoring the consumption of drugs given in the trial were used:

(i) Containers were returned at each attendance and the number of unused capsules and tablets recorded.

(ii) Patients themselves recorded the number of capsules and tablets they had taken in the morning and afternoon each day on progress charts (which are described in section 3c below).

(iii) At each visit to hospital a specimen of urine was collected for analysis.

(3) Measurements and Records

Records were made by the physicians, the physiotherapists, and the patients themselves, and information was compiled also from records of laboratory investigations.

(a) Physician's Assessments.—Each physician completed one or other of two standard forms designed for the trial each time a patient was seen. At the first visit some relatively stable individual characteristics were recorded (age, functional capacity on a four-point scale, duration of disease also on a four-point scale, presence of positive criteria of disease according to the criteria of the American Rheumatism Association (Ropes and others, 1959)), together with a note of drugs prescribed in the previous week. These details were not repeated at subsequent visits. Otherwise the forms were largely similar. The version used at return visits is shown, with slight rearrangement for purposes of reproduction, in Fig. 1 (overleaf).

At the first visit, and at subsequent visits at which treatment was changed, the physician recorded the patient's reported duration of morning stiffness (15 minutes to greater than 2 hours on a scale 1-5), the frequency of waking at night (nil to more than four times on scale 0-5), and the presence or otherwise of spontaneous pain and pain on movement in the past week (as "absent", "intermittent", or "constant"). These are together referred to as the "rheumatoid" group of symptoms. The occurrence of certain symptoms which might be attributed to the drugs given was noted on a three-point scale. These symptoms were anorexia, nausea, vomiting, dyspepsia, diarrhoea, stomatitis, sore throat, rash, oedema, headache, vertigo, and eye symptoms. These are later referred to as "symptoms suggesting toxicity". A similarly scaled recording was made of the presence or absence of joint pain as the physician observed the patient move the following joints in turn through their full range—neck, hands, wrists, elbows, shoulders, hips, knees, ankles, and feet. Lastly, at the end of the second and third periods, the physician recorded the period containing the drug which the patient felt had benefited her most. The drug given in period \(d\) was according to the overall preference of the patient after periods \(a\), \(b\), and \(c\). Sometimes the drug remembered as having given most pain relief was not preferred overall because of the presence of unwanted symptoms during its administration. No detailed records were made at visits when dosage was reviewed unless the drug was discarded altogether and a new period of treatment was started.

(b) Physiotherapist's Assessments.—These were made initially and at the end of each period of treatment with a given drug. They consisted of:

(i) recording the patient's weight,

(ii) measuring the proximal interphalangeal joint sizes by means of jewellers' rings (Hart and Clark 1951),

(iii) measuring the maximum grip generated by each hand (using a sphygmomanometer cuff inflated to 30 mm.Hg and recording the average of three readings),

(iv) measuring the time taken to perform a simple dexterity test with each hand in turn (in which marbles were taken from a tray alongside and placed in the holes of a solitaire board situated in front of the patient),

(v) measuring the angle of maximum active shoulder abduction,

(vi) measuring the time taken to walk a standard distance (47 feet) using whatever aid was necessary (and the same aid on each subsequent occasion).

(c) Patient's Assessments of their Day-to-Day Well-being.—These were recorded on a weekly progress chart. Four charts were supplied at the beginning of a period together with reply-paid envelopes so that each could be returned by post as soon as it was completed. Entries were made by marking appropriate squares in answer to the questions "Do you feel better?", "... much the same?", or "... worse?" Patients were asked to make their entries in the morning and at the end of the day, and to state in the appropriate squares the number of capsules and tablets they had taken during the morning and afternoon of the day.

(d) Laboratory Investigations.—On entry to the trial and at the end of each period, the erythrocyte sedimentation rate, haemoglobin, and white cell count with differential were performed. In addition, the latex and sheep-cell agglutination titres were estimated on the first attendance if these had not been done during the preceding 3 months. On entry to the trial and at both the one-week and four-week (end of period) visits, a specimen of urine was analysed "blind" for glucose and protein and for the presence of paracetamol, aspirin, and the metabolic products of phenylbutazone and both mefenamic and flufenamic acids. Paracetamol and aspirin could be recognized by the addition of ferric chloride to the urine. Metabolites of the feniamic acids fluoresced in ultraviolet light but could not accurately be distinguished from each other.
THE LONDON HOSPITAL
DEPARTMENT OF PHYSICAL MEDICINE & RHEUMATOLOGY
Clinical Evaluation (on admission to trial use form L.55.2)

1. Name ..........................................................  
   Trial Number ................................................  
   Date ..................................................................  
   Current Trial Period a b c d  

2. No. of weeks completed in this period

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>&gt;4</th>
<th>pills</th>
<th>capsules</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>no</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

3. Complete this section one week after starting new treatment

<table>
<thead>
<tr>
<th>Drug satisfactory</th>
<th>Lower dose needed</th>
<th>Higher dose needed</th>
<th>Drug must be changed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

4. Comment:

5. Duration of morning stiffness (Minutes)

<table>
<thead>
<tr>
<th>0</th>
<th>15</th>
<th>30</th>
<th>60</th>
<th>120+</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

6. Number of times waking at night

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7. Symptoms suggesting toxicity

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>?</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td>anorexia</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>nausea</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>vomiting</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>dyspepsia</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>diarrhoea</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>stomatitis</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>sore throat</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>rash</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>oedema</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>headache</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>vertigo</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>eye symptoms*</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>other*</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

* specify

8. Spontaneous pain

<table>
<thead>
<tr>
<th></th>
<th>absent</th>
<th>intermittent</th>
<th>constant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

9. Pain on movement

<table>
<thead>
<tr>
<th></th>
<th>absent</th>
<th>intermittent</th>
<th>constant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

10. Joint pain now on full range of active movement

<table>
<thead>
<tr>
<th></th>
<th>right</th>
<th>left</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>neck</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>hand</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>wrist</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>elbow</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>shoulder</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>hip</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>knee</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ankle</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>foot</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

11. Physician

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr. Tegner</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Dr. Mason</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Dr. Currey</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Dr. Duggal</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Other (specify)</td>
<td></td>
<td>8</td>
</tr>
</tbody>
</table>

12. At end of periods b and c, encircle patient's choice of treatment which has relieved symptoms most effectively.

<table>
<thead>
<tr>
<th></th>
<th>after period b</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>after period c</td>
<td>a</td>
<td>b</td>
</tr>
</tbody>
</table>

13. Treatment prescribed at this visit:

Further Comment:

Fig. 1.—Form used by physicians to record progress of each patient.
(4) Analysis of Findings

(a) Rating System

The various items reported by the physicians (see Fig. 1) were later summarized as three ratings or scores computed as follows:

(i) Rheumatoid symptom rating = \[
\frac{100}{21}(2a + b + 2c + 2d - 6)
\]

where \(a\) = duration of morning stiffness (1-5)
\(b\) = times waking each night (0-5)
\(c\) = spontaneous pain (1-3)
\(d\) = pain on movement (1-3)

(The constants in the equation ensure that the rating takes values from 0 to 100 only and determine the relative weights attributed to different symptoms. The weighting was arbitrary, based on clinical judgement: moderate changes in the weighting factors have no appreciable influence on the results of the trial).

(ii) Joint pain on movement rating = 100 \(f\)

(iii) Symptoms suggesting toxicity rating = 100 \(e\)

\(a, b, c, d, e, \) and \(f\) are the actual ratings for the items so lettered, and \(e\) and \(f\) are the means of the individual ratings respectively for painful joints and symptoms suggesting toxicity (rating present as 1, doubtful as 0-5, and absent as zero). Each rating lies between zero and 100 per cent., representing the range between complete absence and maximal intensity respectively of all items.

(b) Use of Computer

Data from every form completed by a physician, physiotherapist, or patient (2,312 forms comprising in all about 10,000 items) were punched on paper tape suitable for use with an Elliott 803 computing system. The tape was then read into the computer, and the data checked for errors and inconsistencies before being stored in a more permanent form on magnetic tape.

Calculation of the scores (Section 4a), listing of specific items of information, and much of the analysis were then obtained from this store by means of computer programmes either written by two of the authors (M.W. and W.F.), or taken from the Elliott programme library. A further account of the principles followed in analysis by computer has been published elsewhere (Weatherall, 1967).

Results

(1) Patients admitted to the Trial

Details are shown in Table I of the principal measurements made on all patients at the beginning of the trial. These observations represent the state at one moment in the course of a chronic disease in each of 68 patients: moreover, nearly all the patients were receiving analgesic drugs and sometimes steroids or antimalarial drugs. Consequently, the range of scores for rheumatoid symptoms and for joints painful on movement extended to zero, although other criteria observed in the patients with zero scores left no doubt about the diagnosis and gave opportunity for therapeutic improvement to appear. As the ranges show, most of the measurements follow a positively skewed distribution, and appropriate caution has been adopted in later statistical analysis.

(2) Variation between Assessors in the Trial

Part of the apparent variation between patients is probably due to the assessors rather than the patients themselves. In Table II (overleaf) some of the observations shown in Table I are analysed according to the physician who admitted the patient to the trial. Two physicians each saw very few patients and their results are combined as one doctor, "E". Three of the rows of the Table refer to assessments derived from observations of the physicians themselves; that is, the rheumatoid symptom score, the

### Table I

**Patients on Admission to Trial**

<table>
<thead>
<tr>
<th>Clinical Particulars</th>
<th>Mean of 68 Patients Admitted</th>
<th>Mean of 8 Patients who Withdraw</th>
<th>Range in 68 Patients</th>
<th>Standard deviation of single observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (yrs)</td>
<td>54</td>
<td>60</td>
<td>19-76</td>
<td>11-2</td>
</tr>
<tr>
<td>Rheumatoid Symptoms Score (per cent.)</td>
<td>57</td>
<td>49</td>
<td>0-100</td>
<td>15-9</td>
</tr>
<tr>
<td>Pain on Joint Movement Score (per cent.)</td>
<td>41</td>
<td>31</td>
<td>0-100</td>
<td>22-9</td>
</tr>
<tr>
<td>Symptoms Suggesting Toxicity Score (per cent.)</td>
<td>11</td>
<td>20</td>
<td>0-53</td>
<td>13-6</td>
</tr>
<tr>
<td>Weight (lb.)</td>
<td>135</td>
<td>135</td>
<td>85-221</td>
<td>26-2</td>
</tr>
<tr>
<td>Mean Ring Size (mm. diameter)</td>
<td>19</td>
<td>20</td>
<td>11-29</td>
<td>4-1</td>
</tr>
<tr>
<td>Grip Strength (mm. Hg)</td>
<td>146</td>
<td>145</td>
<td>44-300</td>
<td>67-0</td>
</tr>
<tr>
<td>Dexterity test (sec.)</td>
<td>27</td>
<td>29</td>
<td>16-55</td>
<td>9-0</td>
</tr>
<tr>
<td>Abduction Angle</td>
<td>281</td>
<td>258</td>
<td>79-360</td>
<td>64-8</td>
</tr>
<tr>
<td>Walking Time (47ft) (sec.)*</td>
<td>21</td>
<td>38</td>
<td>9-55</td>
<td>14-3</td>
</tr>
<tr>
<td>ESR (mm./hr)</td>
<td>40</td>
<td>45</td>
<td>6-90</td>
<td>21</td>
</tr>
<tr>
<td>Haemoglobin (g./100 ml.)</td>
<td>12</td>
<td>12</td>
<td>9-16</td>
<td>1-7</td>
</tr>
<tr>
<td>White Blood Cells (per c. mm.)</td>
<td>7,600</td>
<td>8,600</td>
<td>4,000-20,500</td>
<td>2,600</td>
</tr>
</tbody>
</table>

*Excluding one patient with a walking time of 172 seconds.
The illness of the physician can be included in the experimental design are very limited, but this isolated observation suggests that neither variable can safely be ignored in any investigation of responses to treatment. Other differences appeared between physicians. Physician A, and to some extent B and E, reported a much higher mean number of certain symptoms (such as nausea, anorexia, or sore throat) in the patients they admitted than did C and D. Indeed, symptoms not obviously related to the patient’s disease were recorded exceptionally rarely by C, who found no such symptoms in 65 per cent of patients. The corresponding figure for all other doctors is 18 per cent (χ² = 11·02; P < 0·001). Differences in assessment between different physicians are well known (Cochrane, Chapman, and Oldham, 1951; Joyce, 1962b), but they are not always explicitly considered in reports of clinical trials. Any systematic transfer of patients from one physician to another could produce significant changes in scores which might wrongly be attributed to treatment given; for as instance, if the patients were all admitted by a consultant and subsequently reviewed by a registrar. In this trial patients were seen by the same physician throughout except when absence from the clinic made this impossible. In all, seventeen patients were seen at least once by a physician other than the one who had admitted them to the trial, but twelve of these changes occurred at the end of the fourth period and have little influence on the main assessment of the drugs. Most of the remaining five changes involved physicians whose ratings did not show evident differences, and in any case the number is such a small fraction of the total that any influence of the change of rates can safely be ignored in the subsequent analysis.
It is also evident that some differences in rating occurred between physiotherapists, but these were not of sufficient magnitude nor so systematic as to cause serious errors of interpretation.

(3) Previous and Concurrent Treatments

All but three of the 68 patients had been taking analgesics regularly, mostly acetylsalicylic acid or phenylbutazone alone or in compound tablets (Table III). Seven patients had already received mefenamic acid from their hospital doctor or general practitioner. In 21 patients another drug or physiotherapy was continued unchanged for the duration of the trial (Table IV). Prednisone was used in eleven subjects in a daily dose between 2.5 and 7.5 mg. each. Chloroquine (200 mg. daily) was used in three subjects.

<table>
<thead>
<tr>
<th>Analgesic</th>
<th>No. of Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetylsalicylic acid alone or in combination</td>
<td>45</td>
</tr>
<tr>
<td>Phenylbutazone alone or in combination</td>
<td>27</td>
</tr>
<tr>
<td>Paracetamol alone or in combination</td>
<td>8</td>
</tr>
<tr>
<td>Mefenamic Acid alone or in combination</td>
<td>7</td>
</tr>
<tr>
<td>Indomethacin</td>
<td>5</td>
</tr>
<tr>
<td>None reported</td>
<td>3</td>
</tr>
</tbody>
</table>

(4) Attendance during the Trial and Withdrawals

Table V sets out the actual duration of the 244 treatment periods. Of these 189 (77 per cent.) were within 3 days of the intended 4-week duration.

Nearly always these periods were consecutive for a given patient but on five occasions a period (of average duration 19 days) intervened between the use of a drug which had been prematurely discontinued and the start of the next treatment.

Eight patients did not complete three periods and were considered as "withdrawals" (Table VI). Four withdrew for ostensibly social or personal reasons; two of these were receiving aspirin and one each phenylbutazone and flufenamic acid at the time of withdrawal from the trial. Four withdrew for apparent medical reasons: one who was receiving phenylbutazone suffered an exacerbation of her arthritis; one who was receiving aspirin developed symptoms related to hypertension; one who was receiving steroids as well as mefenamic acid developed symptoms of an acute gastric ulcer which was radiologically confirmed; and the fourth, also receiving mefenamic acid, was taken to another hospital than our own after gastro-intestinal bleeding* at home. These two patients with gastric ulcers recovered with conservative treatment and withdrawal of mefenamic acid. The eight patients who withdrew included a higher proportion over the age of 60 than the rest of the patients, and a higher proportion (6/8) graded in functional capacity 3 than the patients who remained in the trial (9/60): the latter difference, which is highly significant ($\chi^2 = 11.5$; d.f. = 1; $P<0.001$), was supported also by the several independent measures of functional capacity.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>No. of Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prednisone (2.5 to 7.5 mg. daily)</td>
<td>11</td>
</tr>
<tr>
<td>Physiotherapy</td>
<td>7</td>
</tr>
<tr>
<td>Chloroquine (200 mg. daily)</td>
<td>3</td>
</tr>
<tr>
<td>Gold</td>
<td>0</td>
</tr>
<tr>
<td>No concurrent therapy</td>
<td>47</td>
</tr>
<tr>
<td>Total</td>
<td>68</td>
</tr>
</tbody>
</table>

(5) Physicians' Anticipation of Progress

When admitting a patient to the trial, the physician indicated in completing his form whether he thought the patient would improve, remain unchanged, or deteriorate during the trial. No qualification of this statement was asked for. Only two patients fell into the last category; 33 were expected to improve and 33 to remain unchanged.

* It was initially reported that this patient was admitted to hospital after an haematemesis (Barnardo and others, 1966). Further enquiries did not confirm this report and it now appears that the patient was admitted with melaena and a haemoglobin of 7.4 g/100 ml. No ulcer was reported at a barium meal about 2 weeks later.
No consistent pattern distinguished the properties of patients expected to improve from those not expected to change. No mean values of any variable differed significantly between these two groups. The greatest contrast was shown in the reports of symptoms suggesting toxicity (lower in those expected to improve) and pain on movement (lower among those not expected to change). One physician (A) anticipated improvement in all but one of his eight patients, while two physicians (listed jointly as E) put none of their six patients in this category.

Fig. 2 illustrates the behaviour during the trial of four measurements according to the two main predictive groups. After one 4-week period in the trial, the painful joint score and the grip strength improved slightly more among patients expected to improve than among those not expected to change. After three trial periods, however, this anticipation was no longer borne out.

(6) Progress in Trial and Response to Treatment

When the results of all patients are combined, the means of most measures improved, at least over the first period and often also over the remaining periods of the trial (Fig. 3). The gain was not large compared with the random variation, but the consistent occurrence of a trend in the same direction in most measures suggests that the effect was real. In some measures, for instance of dexterity, the improvement is probably due to practice in performing the test. In others, such as the rheumatoid symptom score, therapeutic optimism engendered by the use of new remedies may have contributed. It will be noted that the improvement was not shared by the ESR, which increased (insignificantly) throughout the first two periods. The means of measurements after the third and fourth period are not precisely comparable with earlier means, because some patients had withdrawn. These patients usually had unfavourable scores and their absence resulted in apparent improvement of the means, including that of the ESR. Even if the eight patients who withdrew before completing three periods are excluded, the remaining sixty do not constitute a balanced set in that some treatments occurred more commonly in one period than another. In view of the improvement in successive periods, comparisons on the sixty introduce some undesirable bias.

The most satisfactory assessment of the effect of drugs is therefore based on the results of the 36 patients who received three periods of treatment according to the original trial design (Table VII, opposite).

The initial improvement generally observed in the first or first two periods of the trial is distributed over all the treatments, and no one stands out as clearly better than any or all of the others. Analyses of variance shows no significant differences in the
METHODS OF DRUG ASSESSMENT

381

TABLE VII
MEAN VALUES AT ENDS OF PERIODS OF TREATMENT
PATIENTS 1–36, PERIODS a, b, and c

<table>
<thead>
<tr>
<th>Values Tested</th>
<th>On Admission (mean ± S.E.)</th>
<th>Periods of Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MEAN</td>
<td>Aspirin</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phenylbutazone</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Flufenamic Acid</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mefenamic Acid</td>
</tr>
<tr>
<td>No. of Patients</td>
<td>36</td>
<td>18</td>
</tr>
<tr>
<td>Rheumatoid Symptoms Score (per cent.)</td>
<td>55.0 ± 2.7</td>
<td>49.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>49.5</td>
</tr>
<tr>
<td>Pain on Joint Movement Score (per cent.)</td>
<td>44.9 ± 4.2</td>
<td>38.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>33.9</td>
</tr>
<tr>
<td>Symptoms Suggesting Toxicity Score (per cent.)</td>
<td>4.5 ± 1.1</td>
<td>8.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.5</td>
</tr>
<tr>
<td>Mean Ring Size (mm.)</td>
<td>19.9 ± 0.7</td>
<td>18.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>18.6</td>
</tr>
<tr>
<td>Grip Strength (mm. Hg)</td>
<td>143 ± 11</td>
<td>138</td>
</tr>
<tr>
<td></td>
<td></td>
<td>154</td>
</tr>
<tr>
<td></td>
<td></td>
<td>157</td>
</tr>
<tr>
<td>Dexterity Test Time (sec.)</td>
<td>26.2 ± 1.2</td>
<td>24.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25.2</td>
</tr>
<tr>
<td>Abduction Angle</td>
<td>278 ± 12</td>
<td>289</td>
</tr>
<tr>
<td></td>
<td></td>
<td>285</td>
</tr>
<tr>
<td></td>
<td></td>
<td>293</td>
</tr>
<tr>
<td>Walking Time (sec.)*</td>
<td>18.6 ± 1.8</td>
<td>15.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>17.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>18.7</td>
</tr>
<tr>
<td>ESR (mm./hr.)</td>
<td>44.7 ± 3.3</td>
<td>35.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>49.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>43.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>47.4</td>
</tr>
<tr>
<td>Haemoglobin (g./100 ml.)</td>
<td>12.1 ± 0.2</td>
<td>12.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11.7</td>
</tr>
<tr>
<td>White Blood Cells (per cu. mm.)</td>
<td>7,800 ± 500</td>
<td>8,400</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7,300</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7,500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7,700</td>
</tr>
</tbody>
</table>

*Excluding one patient with walking time on admission 172 seconds.

Table, and it is clear that any superiority which one drug may have over another is too slight to be apparent against the residual variation which has not been excluded by the trial design.

As the trial was conducted so that the dose of each drug was adjusted according to the response of the patient, it is not surprising that little difference should emerge. The fact that all measurements follow a similar trend is important: it shows that the various drugs are probably acting by a common mechanism and gives no ground for distinguishing, for instance, an "anti-inflammatory" action of phenylbutazone or flufenamic acid from an "analgesic" action of aspirin or mefenamic acid.

There is, however, one disconcerting aspect of this evaluation. When records of drug consumption are taken into account, it appears that only eighteen of the 36 patients considered here were on all counts evidently taking their intended medication throughout the three periods, and in at least seven of the 108 patient-periods the intended medication was undoubtedly not being consumed, because excessive numbers of unused capsules of the drug issued were returned to the pharmacy. When these patients are considered separately, no obvious difference in their progress is apparent. This problem will be discussed more fully elsewhere (Barnardo and Greenwood, in preparation), but it appears that at least three explanations are possible.

The simplest is that the drugs are entirely ineffective: this is most unlikely in view of the evidence already cited that all these drugs are more effective than dummy medication. The second is that, whatever the physician may direct, the patient adjusts her dosage according to the state of the disease, and consequently reduces dosage as improvement occurs. The third, which is supported by some evidence from analyses of urine, is that some patients obtain or continue to consume already available drugs from other sources, and so diminish their requirement for the medication supplied in the trial. Whatever the explanation, much more research into the behaviour of patients is clearly necessary before any accurate appraisal of drugs is possible in outpatient trials of this sort.

(7) Adjustment of Dosage during Treatment

Whatever patients may have done about dosage without reference to their physician, the quantity of drug prescribed to be taken was quite commonly altered after the first week of treatment. When a change was made, it was usually to a higher dosage; changes were not appreciably commoner with one drug than another (Table VIII, overleaf). On the supposition that the revised doses were all close to the optimal, the mean doses so obtained for the four drugs can be taken as equipotent (Table VIII, last line). On this basis, the potencies of the new drugs relative to phenylbutazone are 0.5 (flufenamic acid) and 0.2 (mefenamic acid): on the same basis, aspirin is about one-seventh as potent as phenylbutazone.
Treatment was abandoned with particular drugs because of unwanted effects on thirteen occasions. Most of these occurred with the newer drugs. The difference might be fortuitous (P = 0·24), but the numbers involved are small and it would be unjustified to conclude that there is no difference in toxicity between the various drugs in the doses used. The incidence of individual symptoms is shown in Table IX. Any apparent differences between drugs in the incidence of a symptom are insignificant, except that seven of the nine reported rashes occurred while the patients were receiving mefenamic acid. This greater incidence is significant (by Fisher’s exact test; P = 0·02), but among so many comparisons a single significant finding might well appear at this level by chance. However, mefenamic acid also appears in Table IX as a cause of vomiting, stomatitis, and diarrhoea, and it must be recalled that two subjects who withdrew, one with proven and one with presumed peptic ulceration, both did so after receiving mefenamic acid. It seems probable that this drug is least satisfactory in its gastro-intestinal effects. It is important also not to dismiss differences as non-existent because they do not reach conventional levels of significance with the relatively small numbers observed here.

Table IX

UNWANTED EFFECTS IN 66 PATIENTS IN 192 PERIODS

<table>
<thead>
<tr>
<th>Side-Effects</th>
<th>per cent. Incidence</th>
<th>Drug with Highest Incidence* (per cent.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>On Entry to Trial</td>
<td>During Trial</td>
</tr>
<tr>
<td>Anorexia</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>Nausea</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Vomiting</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Dyspepsia</td>
<td>12</td>
<td>29</td>
</tr>
<tr>
<td>Diarrhoea</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Stomatitis</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Sore throat</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Rash</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Oedema</td>
<td>22</td>
<td>26</td>
</tr>
<tr>
<td>Headache</td>
<td>10</td>
<td>17</td>
</tr>
<tr>
<td>Vertigo</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Eye symptoms</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>Other</td>
<td>9</td>
<td>16</td>
</tr>
</tbody>
</table>

*The differences between drugs do not reach conventional levels of statistical significance. See text for discussion.

(8) Paracetamol Consumption

In addition to a prescribed quantity of trial drug, sometimes consumed faithfully and sometimes not, patients were provided also with tablets containing 500 mg. paracetamol as a supplementary or reserve medication. Consumption of these tablets, as judged by the numbers not returned, varied greatly from patient to patient and usually varied less between periods for a given patient than it varied between patients. Interpretation was complicated, as expected, by some failures to return boxes of tablets at the end of periods. Such failures were not associated with any particular trial drug. The mean quantity of paracetamol apparently consumed did not differ significantly in relation to the trial drug taken at the same time.

(9) The Fourth Period of Treatment

56 patients remained in the trial for a fourth period. During this period, the treatment prescribed was a repetition of whichever drug had appeared most satisfactory in the first three periods. The repetition served two purposes. It allowed an objective display of the patients’ preferences, and if any one drug had been particularly popular, it would appear as the most frequent choice in this period. Also it provided duplication of the responses of each patient to one treatment, and so had special significance in evaluation. If that treatment was consistently more (or less) effective than the other treatments, the fourth period results would agree more closely with the previous results on the same treatment than with the results on other treatments.

Neither of these criteria indicated any difference between the drugs. The frequency of preference is shown in the penultimate line of Table VIII. The preferred treatment was not necessarily the same as the one in which objective measures showed most improvement. Phenybutazone was preferred on the highest proportion of possible occasions and

TABLE VIII

PREFERENCES AND DOSE ADJUSTMENTS

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Aspirin</th>
<th>Phenylbutazone</th>
<th>Flufenamic Acid</th>
<th>Mefenamic Acid</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Patient Periods</td>
<td>35</td>
<td>30</td>
<td>62</td>
<td>65</td>
<td>192</td>
</tr>
<tr>
<td>No. in which Dose was</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unaltered</td>
<td>17</td>
<td>19</td>
<td>35</td>
<td>31</td>
<td>102</td>
</tr>
<tr>
<td>Increased</td>
<td>13</td>
<td>10</td>
<td>21</td>
<td>27</td>
<td>71</td>
</tr>
<tr>
<td>Reduced</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Abandoned</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>Preferences in Fourth Period</td>
<td>10</td>
<td>11</td>
<td>15</td>
<td>20</td>
<td>56</td>
</tr>
<tr>
<td>Mean Dose finally Adopted (g./day)</td>
<td>2.4</td>
<td>0.33</td>
<td>0.67</td>
<td>1.7</td>
<td></td>
</tr>
</tbody>
</table>

(8) Paracetamol Consumption

In addition to a prescribed quantity of trial drug, sometimes consumed faithfully and sometimes not, patients were provided also with tablets containing 500 mg. paracetamol as a supplementary or reserve medication. Consumption of these tablets, as judged by the numbers not returned, varied greatly from patient to patient and usually varied less between periods for a given patient than it varied between patients. Interpretation was complicated, as expected, by some failures to return boxes of tablets at the end of periods. Such failures were not associated with any particular trial drug. The mean quantity of paracetamol apparently consumed did not differ significantly in relation to the trial drug taken at the same time.

(9) The Fourth Period of Treatment

56 patients remained in the trial for a fourth period. During this period, the treatment prescribed was a repetition of whichever drug had appeared most satisfactory in the first three periods. The repetition served two purposes. It allowed an objective display of the patients’ preferences, and if any one drug had been particularly popular, it would appear as the most frequent choice in this period. Also it provided duplication of the responses of each patient to one treatment, and so had special significance in evaluation. If that treatment was consistently more (or less) effective than the other treatments, the fourth period results would agree more closely with the previous results on the same treatment than with the results on other treatments.

Neither of these criteria indicated any difference between the drugs. The frequency of preference is shown in the penultimate line of Table VIII. The preferred treatment was not necessarily the same as the one in which objective measures showed most improvement. Phenybutazone was preferred on the highest proportion of possible occasions and
(10) One Year Later

13 months after the first patient had completed the trial, a survey was begun of the progress of all patients who had left the trial. No exact timing of this survey was arranged, but all patients were reported on at some point between 1 and 2 years after they had completed or withdrawn from the trial. Re-examination by physician and physiotherapist was intended, but changes of staff in the intervening year resulted in a large proportion of assessments not being made by the original observers. Several patients were not specifically re-assessed because they were seen by new staff who were not familiar with the requirements of the trial. Eight patients were not seen at the clinic, but reports from doctors elsewhere were obtained about seven of these. One, aged 68, had died of congestive heart failure after a myocardial infarct 23 weeks after leaving the trial. The remainder were in varying health. None had more disability than expected from their state during the trial, and some had improved modestly. None had recognizable persisting ill-effects which might be attributed to the drugs they had received. The final missing patient had not seen her own doctor and could not be found at home, but was reported by neighbours to be very active and at work most of the day and evening. The recorded changes in the patients seen at our clinic are not presented in detail because they are influenced by uncontrolled variation between observers and contain only a potentially biased selection of all the patients (Table X).

(11) Use of Multiple Measurements

Treatment in rheumatoid arthritis can be assessed in many ways. In this trial at least fourteen criteria have been used without exhausting the possibilities. Different measurements have different significance. For instance, the number of joints painful on movement, the duration of morning stiffness, or the erythrocyte sedimentation rate might each be taken alone as measures of the patient's illness. If so, it would be desirable to show that changes in any of the measurements were significantly correlated with changes in the rest. Otherwise the concept of any single measurable process implicit in the words "the patient's illness" would become nebulous and it would be desirable to think again about what was being achieved by treatment.

Apart from the measurements which are closely related to arthritis, others (such as the haemoglobin concentration or the white cell count) have more diffuse significance, and some (such as the special collection of symptoms suggesting toxicity) are mainly evidence of undesirable effects of treatment. If a drug consistently produced unwanted effects at the same time as it reduced rheumatoid symptoms, the "toxic" score would be correlated with the direct criteria of the rheumatoid disease. Consequently the absence of such correlation would be more encouraging than its presence.

It can also be argued that so many sources of inaccuracy have been disclosed that the data collected were not worth formal analysis. However, different kinds of data are vulnerable in different ways. If the errors were of such magnitude as to make a particular kind of measurement useless, correlation

### Table X

PROGRESS OF NINETEEN PATIENTS SEEN AFTER ONE YEAR*

<table>
<thead>
<tr>
<th>Measurements</th>
<th>Mean Values of Eight Measurements at Each Stage</th>
<th>S.E. of Each Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>At Entry</td>
<td>After Period</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4</td>
<td></td>
</tr>
<tr>
<td>1. Rheumatoid Symptoms Score (per cent.)</td>
<td>53</td>
<td>49 51 51 45</td>
</tr>
<tr>
<td>2. Pain on Joint Movement Score (per cent.)</td>
<td>42</td>
<td>37 34 40 31</td>
</tr>
<tr>
<td>3. Mean Ring Size (mm.)</td>
<td>20</td>
<td>19 18 18 17</td>
</tr>
<tr>
<td>4. Grip Strength (mm. Hg)</td>
<td>144</td>
<td>143 151 147 144</td>
</tr>
<tr>
<td>5. Dexterity Test Time (sec.)</td>
<td>27</td>
<td>26 25 25 26</td>
</tr>
<tr>
<td>6. Abduction Angle</td>
<td>286</td>
<td>294 280 293 292</td>
</tr>
<tr>
<td>7. Weight (lb.)</td>
<td>128</td>
<td>129 129 129 130</td>
</tr>
<tr>
<td>8. Walking Time (sec.)</td>
<td>19</td>
<td>20 18 18 18</td>
</tr>
</tbody>
</table>

*These patients were all from the complete replicate 1–36

(1) excluding one patient value 61 sec.
(2) excluding one patient value 20°
(3) excluding one patient value 100 sec.
would disappear between that and other measures. Some sort of overall analysis was therefore justifiable and was in fact reassuring that the trial measurements had not been overwhelmed by random events. Initially it seemed desirable to see how all the measurements behaved collectively. Correlation coefficients were therefore calculated between all possible pairs of fourteen measurements made at the beginning of the trial, both on the 36 patients who completed the formal design and on the 32 others. Each analysis yielded 91 correlation coefficients. With so large a number of coefficients, a few would be expected by chance to exceed conventional levels of significance. To minimize the influence of such random associations, only those correlations which were significant (at the 10 per cent. level) in both sets of patients were taken as important. The choice of 10 per cent. rather than the conventional 5 or 1 per cent. level followed from the double-testing process, and gives a final probability of a fortuitous correlation of at most 1 per cent. These correlations are shown in Fig. 4. They involved particularly the walking time, the abduction angle, the dexterity test time, the grip strength, and the ESR. The signs of the correlation coefficients depend on whether a low (e.g. grip) or high value (e.g. ESR) meant greater severity, and they were consistent.

![Diagram](image)

**Fig. 4.**—Significant correlations between principal measurements. The figure shows pairs of variables between which the coefficients of correlation were significantly different from zero at beyond the 10 per cent. level in two independent sets of data (viz. patients 1-36 and 37-72). The sign of the coefficient is also shown. Note that a negative sign indicates numerical decrease in one measurement as the other increases, and is as indicative of correlation as a positive sign.

It therefore appeared likely that all these correlated measurements were contributing information about some common process, and a series of fuller analyses was performed to express the measurements in terms of as few components as possible. The method adopted (Harmer, 1961; Seal, 1964) extracted principal components from a correlation matrix based on most of the variables of Fig. 4, but excluding age, functional capacity, and duration of disease, and including the score for symptoms suggesting toxicity. The twelve variables used were all measured at five stages in the trial (initially, and after each 4-weekly period). All except the three laboratory tests were also analysed in measurements from 44 patients at the follow-up visit from 12 to 18 months after entering the trial.

The purpose of such an analysis was to find a number of weighted combinations, called components, of all twelve variables. The weights in the first component are calculated so that the component accounts for as much as possible of all the variability in the data; the weights in the second so that it accounts for as much as possible of the remaining variation after the first component has been allowed for, and so that it is not correlated with the first component, and so on. If the number of components which account for a significant amount of the variation is small, say one or two, then the problem is reduced from one of twelve correlated variables to a single component or to two uncorrelated components.

In these analyses, only one component was consistently statistically significant, though the second and third components had sufficiently clear origins to be given some tentative meaning. The variables most highly weighted in the first component include the grip strength, walking time, rheumatoid symptom score, abduction angle, and painful joint score. All these variables are clearly associated with the patient's disease, and so the rating for a particular patient on this component gives a single numerical measurement of the severity of that patient's illness relative to the other patients included in the same analysis. This component accounted in several analyses for about 29 per cent. of the total variation in all the measurements, so its isolation, though valuable, still leaves much to be accounted for. (If all the twelve variables had been entirely uncorrelated, and twelve factors had been extracted, each factor would have accounted for one variable, i.e. about 8 per cent, of the total). In fact, the second component accounted for about 16 per cent., and was associated particularly with the patient's weight and also with mean ring size. These two variables are significantly (and not surprisingly) correlated with each other and not much with the remaining measurements. The third component was related
particular to the white cell count, and negligibly to any other measurement.

This method of analysis was applied once again, on this occasion to the changes in each kind of measurement during a period of treatment, instead of to the measurements themselves. Two informative findings might be obtained in this way. First, if the variation in scores from period to period represented a real change in severity of the rheumatoid process, the changes should be correlated in much the same way as the measurements themselves. If, on the other hand, the variation was due to an assortment of unrelated influences, no such correlation would be expected. Secondly, given real changes (in the sense just outlined), the components obtained by this analysis would give the best single measure of the magnitude of the change and so provide a basis for determining the effects of different drugs.

In fact, the changes in different measurements were not appreciably correlated. Correlation coefficients between pairs of differences exceeded + 0·4 or were less than − 0·4 on only eight of 220 possible occasions, and did not recur consistently in consecutive periods. Evidently, therefore, the variation in particular measurements within the framework of this trial had little or no general significance about the progress of the disease, and no additional information about the efficacy of drugs was obtained by this unifying analytical procedure.

(12) Serological Tests and Severity of Disease

If the component isolated by the foregoing analysis is meaningful, it would be expected to relate to other measures of disease which had not been included in its calculation. Two such measures available were the titres obtained in the latex test and the sheep cell agglutination test (Table XI). These titres are related to each other, in that sera which agglutinated sheep cells at a high dilution always also gave a positive latex test at high dilution; but the reverse was not true. The association is significant at about the 5 per cent. level of probability. When the titres were compared with individual weightings on the principal component analysis using the Wilcoxon 2-sample rank test, the association with the latex titre was significant (P < 0·05) but not that with the SCAT. The association is not strong but in the expected direction.

Discussion

From the point of view of therapeutic benefit, the results of this trial are straightforward. No appreciable difference has been found between aspirin, phenylbutazone, mefenamic acid, and flufenamic acid used for 4 weeks at a time, in doses adjusted to suit the individual patient. Indeed, the results with each drug are so similar that approximately equipotent doses can be specified: these are 2,400 mg. aspirin, 330 mg. phenylbutazone, 1,700 mg. mefenamic acid, and 670 mg. flufenamic acid per day. The ratio of doses for aspirin and flufenamic acid (3·6 : 1) agrees reasonably well with the 4·5 : 1 estimated by Simpson and others (1966), and the ratio for mefenamic acid and phenylbutazone (5·1 : 1) is practically identical with the estimate (5·1 : 1) given by Cahill and others (1965) in osteo-arthritis. Other trials, reported after this work was completed, have reached similar conclusions (Rajan and others, 1967; Symposium, 1967). To these findings may be added two impressions, neither carrying the weight of statistical significance but both open to confirmation or the reverse as more evidence accumulates. One is that phenylbutazone seemed marginally more popular with patients and seemed repeatedly to appear in various analyses as perhaps a little more often effective than any of the other three drugs. The same modest preference over aspirin and over flufenamic acid is clearly shown in the sequential diagrams of Rajan and others (1967). The other finding is that mefenamic acid more often caused gastro-intestinal disturbance than the remainder.

The lack of discrimination between the drugs may mean that there is really no important difference between them, apart from dosage. It may also mean that all our methods of measurement are so insensitive that they fail to detect real differences. From this point of view, the absence of a dummy or placebo treatment from the trial is regrettable. But differences between phenylbutazone and dummy have been demonstrated clearly with less detailed and no more sensitive measures than those used.
here (Joyce and Mason, 1961; Mason, 1962) and repetition of such a test of validity would not have been in the patients' interests. The use of a single adjustable dose of each drug also prevents any estimation of the slope of the dose-response curve. But it is unlikely that the very poorly controlled conditions of dosage in a trial of this kind would allow any such curve to be constructed.

The uniformity in the patients' responses regardless of treatment ought to have been foreseen as a likely outcome which would interfere with the comparative assessment of different methods of measurement. The correlation and component analysis is useful in showing how much different measurements contribute to the overall assessment of patients at the particular stage here considered, but a similar analysis applied, say, to comparable patients treated with corticosteroids, or to the present patients when their disease had progressed for a further 5 years, might give quite different weights to different measurements. For this reason, no attempt has been made to develop a "rheumatoid index" based on appropriately-weighted combinations of scores. More ground work is necessary before such a development would be generally useful. But the methods of analysis used here are very simple to apply, given adequate computing facilities, and provided that basic work of obtaining clinical data has been done with rigorous thoroughness.

Most clinical trials depend to some extent on the assessment of patients by physicians. The observations reported above (p. 378) are a reminder that variation between assessors may be one of the largest sources of error in a trial. Without objective measurements to put alongside physicians' assessments, the magnitude of such error is easily overlooked. Also, if patients are allocated to physicians by the ordinary processes of some clinics, the variation due to differences between physicians can become confused with differences between patients selected for the physicians. Such selection does not appear to have had demonstrable effects in the present trial, but in other circumstances has turned out to be the largest single source of variation in an entire trial (Reynolds, Joyce, Swift, Tooley, and Weatherall, 1965).

Anyone who may still wish to find evidence of the worthlessness of ill-designed, uncontrolled, or inadequately controlled trials will discover plenty in the preceding pages. Particularly, one may note the subjective improvement observed regardless of treatment in most of the patients, the entirely spurious appearance of improvement created by the falling out of some patients, who happened to be among the more severely-ill participants in the trial, and the contribution of the physician to assessments. All these sources of bias are well known. In at least two recent trials of treatments in rheumatoid arthritis (Currey, 1965; Donnelly, Lloyd, and Campbell, 1967), substantial improvement was achieved both by the active treatment and by the dummy medication. Failure of patients to take drugs as directed is another major source of error, particularly in trials with out-patients (Dixon, Stradling, and Wootton, 1957; Joyce, 1962a; Willcox, Gillan, and Hare, 1965). The error appears more likely to affect interpretations of the trial than health of the patients, as the present results suggest that non-taking of drugs may be a sign of spontaneous improvement rather than of delinquency. Evidently there is a large field for study of the reasons why patients reject prescribed treatment. Many influences can be suggested, and until they are examined and understood the value of this kind of trial as evidence about the therapeutic efficacy of drugs is limited.

On the other hand, it is ultimately the only relevant way of establishing whether drugs in fact are effective in out-patient conditions, and further improvements in methodology are urgently needed.

Summary

Flufenamic acid and mefenamic acid have been compared with aspirin and phenylbutazone over periods of 4 weeks in women with rheumatoid arthritis.

Evaluation was based on records made by physicians, measurements made by physiotherapists, laboratory estimations, and records kept by patients themselves.

The trial was conducted so that patients and physicians were not aware which treatment was being prescribed at any time. Consumption of prescribed treatments was verified by counts of returned drugs, tests of urine, and patient's own records.

The doses used were adjusted for each drug in each patient to find a suitable daily amount.

The mean daily doses used after adjustment were aspirin 2·4 g., phenylbutazone 0·33 g., mefenamic acid 1·7 g., and flufenamic acid 0·67 g.

No differences were detected in the effectiveness of the four drugs, so that these doses can be regarded as equipotent.

Differences in toxicity were not statistically convincing, but suggested that mefenamic acid was particularly liable to cause gastro-intestinal disturbances.

The relative value of different measurements was
assessed and methods of combining them into a single criterion of progress were examined. These statistical analyses indicated that much of the variation in the measurements could be accounted for by a single criterion, but that trials at other stages of rheumatoid arthritis were necessary to develop a generally applicable system of measurement.

Major sources of variation include differences in the assessment of patients by different physicians, subjective improvement of patients during the early part of a trial regardless of treatment, and failure of patients to consume drugs as prescribed. Failure to control these variables does not make a trial worthless, but reduces its sensitivity and is likely to obscure the modest but possibly useful therapeutic benefits of particular drugs.

This trial would not have been possible without help from many people. We should like to thank particularly Dr. P. V. Piggott, of Parke, Davis and Company, and Dr. P. Fowler, of Geigy (U.K.) Limited, for arranging for supplies of drugs to our exacting requirements, and for much other help and continued interest; to Mr. W. Abbot and the staff of the London Hospital Computer Unit; to Dr. R. M. Greenwood for supervising the testing of urines for drugs, and for help in analysis of data; to the physicians and physiotherapists in the Departments of Physical Medicine and Rheumatology; to Mrs. I. Howard for extensive clerical assistance; and to the patients themselves who undertook numerous additional tests and visits to hospital for the purposes of the trial. The work was supported by a grant (to M.W.) from the Medical Research Council.

REFERENCES


L'évaluation des médicaments chez des malades externes atteints d’arthrite rhumatismale

RéSUMÉ

On compara l’acide flufenamique et l’acide méfénamique à l’aspirine et à la phenylbutazone pendant des périodes de quatre semaines chez des femmes atteintes d’arthrite rhumatismale.

On basa l'évaluation sur des notes prises par les médecins, des mesures prises par les physiothérapeutes, des tests de laboratoire et des notes des malades.

L’essai fut arrêté en sorte que ni les malades ni les médecins puissent identifier le médicament ordonné. Pour vérifier la consommation médicamenteuse on compa les médicaments retournés, analysa l’urine et examina les notes des malades.

Pour arriver à la dose quotidienne appropriée on ajusta individuellement la dose de chaque médicament.

La dose quotidienne moyenne après l’ajustement fut de 2,4 g. d’aspirine, 0,33 g. de phenylbutazone, 1,7 g. d’acide méfénamique et 0,67 g. d’acide flufenamique.

On ne trouva pas de différence en ce qui concerne l’efficacité des quatre médicaments, de manière qu’aux doses indiquées leur puissance peut être considérée égale. En ce qui concerne la toxicité, du point de vue statistique les différences furent peu convaincantes, mais on pense que l’acide méfénamique aurait particulièrement une tendance à produire des désordres gastro-intestinaux.

On évalua la valeur relative de différentes mensurations et on examina les méthodes pour les combiner en un seul critère du progrès. Ces analyses statistiques indiquèrent que beaucoup de variations des mensurations pourraient s’expliquer par une seule cause, mais qu’on aurait besoin d’autres essais à d’autres stades de l’arthrite rhumatismale pour développer un système de mensuration d’application générale.

Parmi les sources majeures des variations on trouve des différences en l’évaluation du malade selon le médecin, l’amélioration subjective des malades au début de l’essai indépendamment de la médication et le fait que les malades ne prennent pas leurs médicaments de la manière prescrite. Le défaut de contrôler ces variables n’annule pas la valeur de l’essai, mais réduit sa précision et peut masquer des avantages thérapeutiques modestes et possiblement utiles d’un médicament particulier.

La valoración de medicamentos en pacientes ambulantes con artritis reumatoide

SUMARIO

Se compararon el ácido flufenámico y el ácido mefenámico con la aspirina y la fenilbutazona durante periodos de cuatro semanas en enfermos con artritis reumatoide.

La valoración fue basada sobre notas de los médicos, medidas de los fisioterapeutas, investigaciones de laboratorio y apuntes de los enfermos.

La investigación fue conduida de manera que ni los enfermos ni los médicos conocieron el medicamento recetado. Para averiguar la toma de los productos recetados se contaron los medicamentos devueltos, la orina fue analizada y se examinaron los apuntes de los enfermos.

Para llegar a la dosis diaria apropiada se ajustó individualmente la dosis de cada medicamento.

La dosis diaria media después de ajuste fue de 2,4 g. de aspirina, 0,33 g. de fenilbutazona, 1,7 g. de ácido mefenámico y 0,67 g. de ácido flufenámico.

No se encontraron diferencias respecto a la eficacidad de los cuatro medicamentos, de modo que, a dosis indicadas, su poder puede considerarse igual.

Respecto a la toxicidad, las diferencias fueron estadísticamente poco convincentes, se sospecha sin embargo que el ácido mefenámico en particular puede ocasionar disturbios gastrointestinales.

Se consideró el valor relativo de diferentes modos de valoración y se examinaron métodos para incorporar estos modos en un solo criterio. Análisis estadísticos indican que muchas variaciones de las mediciones pueden posiblemente explicarse por una sola causa, pero se necesitan otras investigaciones de otras etapas de la artritis reumatoide para elaborar un sistema de valoración generalmente aplicable.

Otras fuentes de variaciones incluyen diferencias en la valoración de los enfermos por diferentes médicos, mejoría subjetiva de los enfermos al principio de la investigación independiente de la medicación y el hecho de que los enfermos dejan de tomar los medicamentos de la manera ordenada. El hecho de que estas variables no se pueden controlar no anula el valor de la investigación pero reduce su precisión y puede obscurecer los beneficios terapéuticos modestos pero posiblemente utiles de un medicamento particular.
Assessment of drugs in out-patients with rheumatoid arthritis. Evaluation of methods and a comparison of mefenamic and flufenamic acids with phenylbutazone and aspirin.

R M Mason, D E Barnardo, W R Fox and M Weatherall

Ann Rheum Dis 1967 26: 373-388
doi: 10.1136/ard.26.5.373