A METHOD FOR TESTING ANALGESICS IN RHEUMATOID ARTHRITIS USING A SEQUENTIAL PROCEDURE

BY

J. HAJNAL, J. SHARP, AND A. J. POPERT

From the Departments of Rheumatism Research and Preventive Medicine, University of Manchester and Manchester Royal Infirmary

In the management of patients with inflammatory erosive arthritis of the rheumatoid type, general measures such as rest in bed and splintage followed by graded exercises form the basis of treatment during periods of disease activity. Medication with drugs is of value in relieving the painful symptoms of the arthritis, but there is no agreement as to which of the presently available drugs should be used and new preparations are continually being introduced. The object of this investigation was to devise a method whereby the effectiveness of drugs in controlling symptoms in patients admitted to hospital for treatment of arthritis by general measures could be rapidly assessed. It was hoped to reduce to a minimum the number of patients required for the study by the use of sequential analysis. Trials of this kind should cause the patient no hardship and should involve the minimum disturbance of ward routine and little additional labour for nursing and medical staff.

Methods

Duration.—In some preliminary studies, aspirin, a new analgesic, and placebo tablets were given for periods of 3 or 6 days with daily assessments; but it soon became clear that a longer period of therapy, fewer assessments, and the elimination of a placebo period were needed to fulfil the requirements mentioned above. The final design consisted of a 2 weeks' trial for each patient, during which each of the two drugs being compared was given for a period of one week, assessments being made on the second, fifth, and seventh day of each week of medication. In many patients such 2-week courses were repeated during the whole of their stay in hospital. The trials were all of the double-blind type; assessments being made at approximately the same time each morning by the medical resident and the technicians attached to the unit during the course of their routine duties in the ward.

Criteria of Assessment

(1) The severity of pain and stiffness, reported by the patient to the resident, each characteristic being rated by the resident on a numerical scale.

(2) The strength of grip (in mm. Hg), measured by squeezing a bag attached to a sphygmomanometer and inflated to a pressure of 20 mm. Hg, the value of three grips with each hand alternately being recorded. The figures accepted were the highest that the patient could maintain for a period of 2 to 3 seconds. The grip tests were recorded by the technical staff.

In addition, other tests (such as finger swelling measured by ring size, the time taken to tie knots, and the degree of tenderness elicited by squeezing affected joints) were used at some stage, but these proved less useful. Only the records of pain, stiffness, and strength of grip were used throughout the trials, and only these characteristics have been subjected to analysis.

Patients.—The trials were conducted in the wards of the Manchester Royal Infirmary and the Devonshire Hospital, Buxton, and all patients with active inflammatory polyarthritis of the rheumatoid type admitted to the University Centre's beds during the trial period participated unless there were special contraindications to their inclusion. There were four main reasons for such exclusion:

(a) No involvement of the hands or such deformity of the hands as to make the grip test inapplicable.

(b) Corticosteroid therapy immediately before admission.

(c) A history of previous toxic reactions to the drugs being investigated.

(d) Complicating diseases that might be aggravated by the drugs being investigated.

The patients were not told that a comparison of drugs was being made, but they were told that the object of the procedures was to determine which tablet suited them best as individuals, and this was in fact a subsidiary but important object of the trial procedure.

Drugs.—In the trials so far conducted, calcium aspirin in a daily dose of 5 g. was used as the standard medication, since this represented the routine medication used.
in our wards before the trial programme, and this has been compared with four other forms of medication:

1. A new aspirin-like agent, 4-hydroxy-isophthalic acid (HIPA), given in a dose of 6-7 g. daily.
2. Prednisone in a dose of 30 mg. daily.
3. Phenylbutazone in a dose of 300 mg. daily.
4. N-acetyl-p-aminophenol (NAPAP) marketed as "Panadol" in a dose of 7-5 g. daily.

In each trial alternate patients in order of admission were given aspirin during the first week and the other drug during the second week, the order of treatment being reversed in the other patients. When the dosage of the two drugs was similar they were made up in tablets that looked alike and they were given in five daily doses. When the dosage differed as with prednisone and phenylbutazone the patients received two lots of tablets, large and small, during each week, so that when comparing prednisone with aspirin one week’s therapy consisted of prednisone plus dummy aspirin, while during the other week it was aspirin plus dummy prednisone.

Sequential Analysis

The length of three of the four trials reported was determined by sequential analysis. Traditional significance tests are valid only if the number of cases included in the sample is determined independently of the observations. In practice, no doubt, research workers have often gone on collecting data until a “significant” result was obtained. This procedure leads to erroneous conclusions unless special “sequential” statistical techniques are employed. Such techniques were first developed during World War II for industrial application. Their use in clinical trials has been widely discussed (for example, by Armitage (1954) and Bross (1958)) and they have been employed in a few trials listed by Bross (1958). In therapeutic trials, the idea of subjecting data to continuous analysis as they become available has special attractions, since it is often very desirable that a trial be discontinued as soon as it can clearly be established that one treatment is more effective than another.

The particular sequential technique employed in the present trials is the sequential t-test. We shall proceed to describe in non-technical language the main problems which arose in the application of this technique, and also to show the actual calculations, which are quite simple. Additional information may be found in the Appendix.

The essence of a sequential trial of two treatments is to perform a quick analysis, as each additional set of data comes in, to determine whether (i) enough data have accumulated for a decision, or (ii) more data need to be collected. Alternative (i) in the present instance could come about in two ways: we may stop because the data so far collected indicates with reasonable certainty either that there is no difference between drugs or that there is an appreciable difference.

To select a particular procedure, we must determine what we mean by "with reasonable certainty" and fix a target difference which we agree to regard as "appreciable". For the present trials "with reasonable certainty" was taken to mean that the chance of a correct decision should be nineteen out of twenty. For a sequential t-test, the target difference is fixed in terms of the variability between observations for different patients. This form of statement is not in itself very informative. That variability is measured by the standard deviation, and will depend on the index chosen to assess the effect of the drugs. In the present instance the target difference corresponded to 0-6 of the standard deviation of the observations for individual patients. The results suggest that this target is smaller than the difference between aspirin and a placebo or between prednisone and aspirin. A trial conducted in accordance with the method chosen should, therefore, detect any substance which is as much superior to aspirin as aspirin is to a placebo. It should also detect any medication which is not much better than an inert substance.

For the sequential procedure it was necessary to select a single index to measure the differences between the drugs. The main aim of statistical analysis is to disentangle, as far as possible, the differences due to the drugs from those due to other fluctuations in the observations (due to "natural" variations in the activity of the disease, changes in the patient’s mood, etc.). It was found in the preliminary experiments that differences between drugs seemed largest in relation to other fluctuations when measured by grip strength. Hence the grip strength was used as the index for the sequential test.

When a new analgesic drug is to be tested, the only information usually available relates to its toxicity and apparent analgesic properties in animals and human volunteers. Its rate of action and whether, in the case of effective drugs, its withdrawal may be followed by a “rebound” of symptoms is not known. These considerations might suggest that in the present scheme only the last assessment on each drug (i.e. when the patient has been taking the medication for 7 days) should be taken into account. However, since these and other factors may be of varying importance with different drugs, it seemed best to use the simplest method and to give equal weight to all assessments, i.e. to compare the means of the readings of the features assessed.
while the patient was receiving one drug with those
recorded while he was receiving the other drug.

Table I shows how the grip records of two patients
(the first two in the prednisone/aspirin trial) were
handled. The two patients were given the drugs
in different order. For each patient the mean grip
strength was first computed for each week. The
difference between the two means was then taken,
the mean for the prednisone week being always
subtracted from the mean of the aspirin week
(whichever week came first). The differences so
obtained, which will be called mean grip differences,*
are shown in the last line of the Table.

One further step was involved in preparing the
data for the sequential analysis. It was found that,
as a rule, the patient's condition tended to improve
considerably as the trial proceeded independently
of drug action. Consequently, if two equally
effective or ineffective drugs are compared, the one
given in the second week would usually appear to be
appreciably better. This phenomenon is further
described below. To take account of it in the
sequential analysis, the mean grip differences were
added together for pairs of successive patients;
the two members of each pair differed in the order
in which they had received the drugs.

The procedure is illustrated in Table II, which
relates to the prednisone/aspirin trial. For succes-
sive patients who received prednisone in the first
week and completed the trial, we denote the mean
grip differences by \( x_1, x_2, x_3 \ldots \); for patients who
started on aspirin and completed the trial we denote
the mean grip differences by \( y_1, y_2, y_3 \ldots \). Thus,
as the observations become available, we compute

\[ z_i = \frac{(\sum y_i)^2}{\sum d_i^2} \]


* As patients entered the trial they were alternately assigned
to start on prednisone or aspirin. If they had all completed the trial,
the observations would have come in in the order \( x_1, y_1, x_2, y_2, x_3, y_3 \ldots \). However, some patients were withdrawn because of
suspected toxicity, faulty diagnosis, etc., and it happened, therefore,
that two complete records suitable for sequential analysis became
available for prednisone-starters, with no intermediate aspirin
starter.

### Table I

**SPECIMEN ANALYSIS OF GRIP RECORDS**

<table>
<thead>
<tr>
<th>Serial No. of Assessment</th>
<th>Patient A</th>
<th>Patient B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Medication</td>
<td>Grip Strength (mm. Hg)</td>
</tr>
<tr>
<td></td>
<td>Prednisone</td>
<td>190</td>
</tr>
<tr>
<td></td>
<td>Aspirin</td>
<td>183</td>
</tr>
<tr>
<td>Mean of First Week Mean of Second Week</td>
<td>Prednisone</td>
<td>212</td>
</tr>
<tr>
<td>Mean Difference between Weeks</td>
<td>Aspirin minus Prednisone</td>
<td>-10:4</td>
</tr>
</tbody>
</table>

*Note: The grip strength at each assessment is itself the mean of six readings, three being taken on each hand.*

### Table II

**SEQUENTIAL ANALYSIS OF PREDNISONE TRIAL**

<table>
<thead>
<tr>
<th>Serial No. of Pair of Patients</th>
<th>Mean Grip Differences*</th>
<th>Sum of Differences for Successful Pairs</th>
<th>Cumulative Total of Previous Column</th>
<th>Squared Sums</th>
<th>Cumulative Total of Previous Column</th>
<th>Sequential Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>( x_i )</td>
<td>( y_i )</td>
<td>( d_i = x_i + y_i )</td>
<td>( \Sigma d_i )</td>
<td>( d_i^2 )</td>
<td>( \Sigma d_i^2 )</td>
</tr>
<tr>
<td>1</td>
<td>-10</td>
<td>-30</td>
<td>-40</td>
<td>-200</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>3</td>
<td>-2</td>
<td>-50</td>
<td>-30</td>
<td>-300</td>
<td>900</td>
<td>2,116</td>
</tr>
<tr>
<td>4</td>
<td>-7</td>
<td>-39</td>
<td>-46</td>
<td>-150</td>
<td>2,166</td>
<td>6,332</td>
</tr>
<tr>
<td>5</td>
<td>-16</td>
<td>-40</td>
<td>-56</td>
<td>-336</td>
<td>1,098</td>
<td>8,268</td>
</tr>
<tr>
<td>6</td>
<td>-18</td>
<td>-51</td>
<td>-33</td>
<td>-227</td>
<td>1,089</td>
<td>9,357</td>
</tr>
<tr>
<td>7</td>
<td>-46</td>
<td>-78</td>
<td>-32</td>
<td>-259</td>
<td>1,024</td>
<td>10,381</td>
</tr>
</tbody>
</table>

* The mean grip strength for the prednisone week has been subtracted from the mean for the aspirin week.*
successively $x_1 + y_1$, $x_2 + y_2$, $x_3 + y_3$, \ldots The sums so obtained will be denoted by $d_1$, $d_2$, $d_3$, \ldots In particular, from Table I, we have for the first pair

$$10 \cdot 4 + (-30 \cdot 4) = -200$$

(The figures in Table II have been rounded for convenience.)

The remaining computations for the sequential analysis are also given in Table II. The sequential criterion computed at each stage (denoted by $z_i$) is given by the formula

$$z_i = \frac{\left(\Sigma d_i\right)^2}{\Sigma d_i^2}$$

Here $\Sigma d_i$ denotes the total of all the $d$s so far available, and $\Sigma d_i^2$ denotes the total of the squares of all the $d$s so far available.

The sequential criterion $z$ thus computed is compared with two predetermined numbers, an upper and a lower stopping limit. If the value of $z_i$ exceeds the upper limit, the trial ends because the data are sufficient to indicate that there is an appreciable difference between the drugs. If $z_i$ falls short of the lower limit, the trial is ended because the data are sufficient to show that there is no appreciable difference between the drugs. So long as $z$ remains between the upper and the lower limits, additional observations are taken.

Stopping limits for the present trials are shown in Table III up to the 30th pair of patients. The limits change as the number of observations increases. By referring back to Table II, it may be seen that in the prednisone/aspirin trial the sequential criterion stood at 5.51 after the sixth pair, which is below the upper stopping limit at that point (5.58). After the sixth pair, however, the sequential criterion jumped to 6.46 and thus exceeded the upper stopping limit for the seventh pair (5.50). In addition to the stopping limits, Table III shows the course of the sequential analysis of the phenylbutazone/aspirin and NAPAP/aspirin trials.

The position of the stopping limits is such that the trials have the characteristics specified above, i.e. essentially a 1 in 20 chance of a wrong decision. As is explained in the Appendix, lists of stopping limits are readily available for trials for which it is desired to specify other characteristics.

<table>
<thead>
<tr>
<th>Serial No. of Pair of Patients</th>
<th>Phenylbutazone Trial</th>
<th>NAPAP Trial</th>
<th>Stopping Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sums of Mean Grip Differences $d_i$</td>
<td>Sequential Criterion $z$</td>
<td>Sums of Mean Grip Differences $d_i$</td>
</tr>
<tr>
<td>1</td>
<td>-3</td>
<td>1.00</td>
<td>-27</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1.00</td>
<td>+43</td>
</tr>
<tr>
<td>3</td>
<td>+21</td>
<td>0.75</td>
<td>-8</td>
</tr>
<tr>
<td>4</td>
<td>-10</td>
<td>0.13</td>
<td>-9</td>
</tr>
<tr>
<td>5</td>
<td>-23</td>
<td>0.18</td>
<td>+17</td>
</tr>
<tr>
<td>6</td>
<td>-14</td>
<td>0.48</td>
<td>+4</td>
</tr>
<tr>
<td>7</td>
<td>+11</td>
<td>0.21</td>
<td>+2</td>
</tr>
<tr>
<td>8</td>
<td>-18</td>
<td>0.74</td>
<td>+2</td>
</tr>
<tr>
<td>9</td>
<td>+4</td>
<td>0.58</td>
<td>+23</td>
</tr>
<tr>
<td>10</td>
<td>-31</td>
<td>1.46</td>
<td>+11</td>
</tr>
<tr>
<td>11</td>
<td>-17</td>
<td>2.12</td>
<td>+32</td>
</tr>
<tr>
<td>12</td>
<td>+29</td>
<td>0.69</td>
<td>+23</td>
</tr>
<tr>
<td>13</td>
<td>+3</td>
<td>0.60</td>
<td>+6</td>
</tr>
<tr>
<td>14</td>
<td>+5</td>
<td>-17</td>
<td>+6</td>
</tr>
<tr>
<td>15</td>
<td>-17</td>
<td>+6</td>
<td>2.29</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>2.29</td>
<td>+4</td>
</tr>
<tr>
<td>17</td>
<td>+8</td>
<td>3.52</td>
<td>+4</td>
</tr>
<tr>
<td>18</td>
<td>+23</td>
<td>3.52</td>
<td>+4</td>
</tr>
<tr>
<td>19</td>
<td>+10</td>
<td>4.49</td>
<td>+3</td>
</tr>
<tr>
<td>20</td>
<td>+33</td>
<td>5.48</td>
<td>+2</td>
</tr>
<tr>
<td>21</td>
<td>+36</td>
<td>5.48</td>
<td>-3</td>
</tr>
<tr>
<td>22</td>
<td>-2</td>
<td>5.36</td>
<td>+1</td>
</tr>
<tr>
<td>23</td>
<td>+10</td>
<td>6.76</td>
<td>+2</td>
</tr>
<tr>
<td>24</td>
<td>-33</td>
<td>7.05</td>
<td>+6</td>
</tr>
<tr>
<td>25</td>
<td>+6</td>
<td>6.59</td>
<td>+2</td>
</tr>
<tr>
<td>26</td>
<td>+3</td>
<td>7.39</td>
<td>-2</td>
</tr>
<tr>
<td>27</td>
<td>+21</td>
<td>8.17</td>
<td>+20</td>
</tr>
<tr>
<td>28</td>
<td>+20</td>
<td>3.03</td>
<td></td>
</tr>
</tbody>
</table>

Note: Values of $d_i$ have been rounded, so that the $z_i$ calculated from them would not necessarily be in precise agreement with those given.
METHOD FOR TESTING ANALGESICS IN RHEUMATOID ARTHRITIS

Figure.—Boundaries for "Difference" and "No Difference" between drugs for the sequential test employed, and sample paths of the prednisone phenylbutazone, and NAPAP trials.

(A) Drug better or worse than Aspirin
(B) No difference between the Drug and Aspirin

It is informative to follow the progress of a sequential trial on a chart such as that shown in the Figure. The stopping limits are represented by two heavy black lines. Successive values of the sequential criterion $z_i$ can be entered as the trial proceeds. The trial continues so long as $z_i$ remains between the heavy black lines. The crossing of one of the lines is the signal for ending the trial. In the Figure, the course of all the three sequential trials reported in the present paper is shown in this way.

Table IV summarizes the sequential trials in regard to duration and outcome. The number of patients required till a decision was reached varied widely. This is in part a matter of chance. However, the difference in effectiveness between the medications being tested is an important determinant. If the difference between them is as great as the "target" difference which the procedure selected is designed to detect, the average number of pairs required is likely to be about fifteen. If the difference exceeds the "target" then the trial usually ends earlier, as was the case in the prednisone trial, because such a large difference can be quickly detected. (No doubt the prednisone trial happened to be a particularly short one even for a comparison

<table>
<thead>
<tr>
<th>Table IV</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DURATION AND RESULTS OF SEQUENTIAL TRIALS</strong></td>
</tr>
<tr>
<td><strong>Drug</strong></td>
</tr>
<tr>
<td>Prednisone</td>
</tr>
<tr>
<td>Phenylbutazone</td>
</tr>
<tr>
<td>NAPAP</td>
</tr>
</tbody>
</table>

* Each drug was compared with 5 g. calcium aspirin per day.
† Only patients withdrawn because of possible toxic effects of the medication are counted here. Withdrawals for other reasons (e.g. faulty initial diagnosis) are not included (see text).
where the true difference is considerable.) In the case of phenylbutazone there was almost no difference between the drugs. Such a trial usually takes somewhat longer than one where the difference is at the "target" level. The longest runs, however, will usually occur when the difference between the drugs is about half as big as the "target" difference. The NAPAP trial was probably long partly because the difference between NAPAP and aspirin as measured by grip was small, but not negligible.

Analysis by Traditional Procedures

Numbers of Patients.—The sequential analysis was used only to determine when sufficient patients had been included in the trials. After each trial had been terminated, a full analysis of the results was made by traditional statistical methods. The results of this analysis are summarized in Tables V to X (below).

These Tables include a considerably larger number of patients than that used for the sequential procedure. There are two main reasons for this. First, observations were recorded for all patients, but initially only those with an erythrocyte sedimentation rate of over 30 mm./hr (Westergren) and without radiological changes in the hands were used in the sequential analysis. (The first restriction was later abandoned, and the second considerably relaxed.) Secondly, some patients were included intentionally in all trials in excess of those required for decision by the sequential procedure, because patients who had already started the trial when the sequential analysis indicated that an answer had been obtained completed their courses. Moreover, since it was not always possible to switch immediately to the testing of a new drug, the old trial was in two instances continued for some additional weeks.

The additional cases are especially helpful, because it is not strictly correct to analyse data collected in a sequential trial by traditional statistical procedures. The risk of error on the ground is, however, greatly diminished by the additional cases; standard errors may, therefore, be interpreted fairly confidently as though the length of the trials had been fixed without looking at the data as they accumulated.

Although the patients for whom complete records were obtained exceed those used in the sequential analysis, the total numbers admitted to the trials were greater still. Apart from those who had to be withdrawn from the trials because of the toxic effects of the drugs, there were patients whose treatment was altered or who were transferred to another hospital during the trial period for reasons unconnected with the trials or in whom the original diagnosis was shown to be incorrect after they had started on the trial. There were also a few in whom the assessments were not all made, or were improperly entered, etc. In all, 35 patients were admitted to the prednisone trial, 48 to the phenylbutazone trial, and 100 to the NAPAP trial.

Tables VI, IX, and X (as also the Appendix Tables) present, in addition to the three trials which were analysed sequentially, data from the HIPA/aspirin trial. HIPA was used in preliminary trials in which various schemes (differing in length of administration of drugs, frequency of assessments, etc.) were tried. In all, 27 patients were involved and thirteen patients participated in the final series of these preliminary trials in which the procedure was identical with the subsequent trials (i.e. one week on each of two drugs, three assessments per week), but sequential analysis was not yet employed. The results for these thirteen patients are reported below. As explained in the Appendix, this experience provided material helpful in the choice of appropriate sequential methods.

Comparison of Medications.—The results of the NAPAP trial are summarized in Table V. In reading this Table it should be remembered that an improvement in the patient’s condition is reflected in an increase of the score for grip, but in a decrease of the scores for pain and stiffness. Hence the superiority of aspirin is indicated by positive figures for grip, and negative figures for pain and stiffness.

| Table V |
| DIFFERENCE IN MEAN DAILY ASSESSMENTS BETWEEN NAPAP AND ASPIRIN (ASPIRIN RATINGS MINUS NAPAP RATINGS) |
|----------------|----------------|----------------|
| Assessment      | (a) Patients who Started on NAPAP | (b) Patients who Started on Aspirin | Mean of (a) and (b) |
| Grip (mm. Hg)   | +11             | 0              | ± 5.5 |
| Pain*           | -0.29           | -0.06          | -0.175 |
| Stiffness*      | -0.25           | -0.08          | -0.085 |
| No. of Patients | 38              | 37             |       |

* Pain and stiffness were assessed in a rating of patients' statements on an arbitrary scale running from 0 (no pain or stiffness) to 3 (very severe pain or stiffness).

† Numbers of patients are those for pain and stiffness. Grip records were somewhat less complete and the figures were based on 34 and 35 patients respectively (for the two order groups).

The order in which the drugs are given makes an appreciable difference, as explained earlier, since the assessments tend to improve between the first and second weeks independently of medication. If these results for the patients who first received aspirin and those who first received NAPAP are averaged, the
METHOD FOR TESTING ANALGESICS IN RHEUMATOID ARTHRITIS

195

effect of the "spontaneous improvement" may be assumed to cancel out. The figures given in the last column of Table V are therefore estimates of the difference in effect between aspirin and NAPAP.

Figures derived in the same way as the last column of Table V (correct to two places of decimals) are shown for all four trials in Table VI (overleaf). The method of calculation of the standard errors is detailed in Section 2 of the Appendix. The directions of the differences correspond, of course, to the conclusions listed in Table IV.

The observations on grip, pain, and stiffness were consistent; i.e.
...

Drug Toxicity.—The numbers of patients developing manifestations which could be attributed to toxic effects of the drugs are indicated in Table VII (overleaf), and further details are given in Table VIII (overleaf).

In these trials any symptoms or signs which might represent side-effects of the treatment were recorded as toxic effects of the medication being given at the time. The recorded incidence of toxic effects is probably therefore unduly high, since some of these manifestations may have been unrelated to the medication. Such errors would, however, be likely to occur with about equal frequency with aspirin and with the drug being compared.

There was no significant difference (p > 0.5) between the proportions of patients in whom toxic effects were ascribed to aspirin in the three trials, namely 23 per cent. (7/31) in the prednisone trial, 22 per cent. (10/45) in the phenylbutazone trial, and 15 per cent. (14/94) in the NAPAP trial, and this suggests that such effects were recorded consistently in each trial. A satisfactory comparison of the proportions of those who continued to take the different medications in alternate weeks who developed toxicity after the first fortnight is not possible, since the factors determining whether patients continued with this programme varied from trial to trial and sometimes during the course of a single trial.

Aspirin.—In the 170 patients who received aspirin, toxic effects were ascribed to the drug in 48 (28 per cent.). Five were withdrawn from a trial during the first course and five during later courses, and three others completed one or more 2-week courses, but were not given further courses on account of these effects, a total of thirteen (8 per cent.).

28 (17 per cent.) suffered from either deafness, tinnitus, or both; four were withdrawn during a trial course, and another was not given a third course on account of these symptoms. A further patient complained of dizziness.

Gastro-intestinal disturbances, usually anorexia and nausea, but occasionally epigastric pain, vomiting, or diarrhoea, were noted in twenty patients (12 per cent.) necessitating withdrawal during a trial course in four and at a later stage in two others who continued on the trial regime after the first fortnight.

Mild skin rashes occurred in three patients receiving aspirin and all were withdrawn from the trials. One occurred at the end of the first week of the trial; it cleared rapidly when the aspirin was discontinued but recurred some weeks later when aspirin was resumed in lower dosage. In the others the rashes occurred during the 4th and 8th weeks respectively. In each case the rash cleared up rapidly and its relationship to the medication appeared doubtful; in one of them patch tests with aspirin and NAPAP were negative and the other received aspirin subsequently without adverse effects.

One patient complained of headache and nausea and another of headache and "fullness of the face" while taking aspirin during the second fortnight; one was withdrawn during and the other after the completion of the second fortnight.

One patient complained of depression on aspirin, but this may have represented a withdrawal effect of prednisone, which she had received in the previous week.

One patient had spontaneous epistaxis while on aspirin which had also occurred during her previous week on NAPAP; the plasma prothrombin concentration was found to be 50 per cent. and there was other evidence of liver damage. The trial regime was not continued, but subsequently she tolerated aspirin in lower dosage without ill-effects.

Prednisone.—Only one of the thirty patients receiving prednisone developed toxic effects necessitating withdrawal from the trial; she suffered from essential hypertension and when on prednisone in
### TABLE VI

**DIFFERENCES IN MEAN DAILY ASSESSMENTS BETWEEN EFFECT OF ASPIRIN AND OTHER DRUGS TESTED**

(RATINGS FOR DRUGS SPECIFIED SUBTRACTED FROM THOSE FOR ASPIRIN)

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Drug Compared</th>
<th>HIPA</th>
<th>Prednisone</th>
<th>Phenylbutazone</th>
<th>NAPAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grip (mm. Hg)</td>
<td>Total</td>
<td>-25±5</td>
<td>-15±4</td>
<td>+1±2</td>
<td>-6±1</td>
</tr>
<tr>
<td>Pain (Arbitrary Rating)</td>
<td>Starting First Course</td>
<td>+0±21</td>
<td>+0±3</td>
<td>+0±3</td>
<td>+0±3</td>
</tr>
<tr>
<td>Stiffness (Arbitrary Rating)</td>
<td>Starting Further Course(s)</td>
<td>+0±12</td>
<td>+0±3</td>
<td>+0±3</td>
<td>+0±3</td>
</tr>
<tr>
<td></td>
<td>On Drug Specified</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>On Aspirin</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of Patients</td>
<td>Total</td>
<td>13</td>
<td>29</td>
<td>37</td>
<td>75</td>
</tr>
</tbody>
</table>

**Notes:** Numbers of patients are those for pain and stiffness. The numbers included who were first given one or the other drug differed at most by 1. Grip records were slightly less complete. The numbers of patients tested for grip may be seen in Table X.

For Grip Strength, a positive sign indicates that aspirin gave a superior result, and a negative sign indicates that the drug tested gave a superior result.

For Pain and Stiffness the opposite is the case.

### TABLE VII

**INCIDENCE OF DRUG TOXICITY**

<table>
<thead>
<tr>
<th>Daily Dosage of Drug compared with Calcium Aspirin 5 g.</th>
<th>Number of Patients</th>
<th>Average Number of Days</th>
<th>Toxicity (Number of Patients)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Starting First Course</td>
<td>Satisfactory Datare Toxicity*</td>
<td>Starting Further Course(s)</td>
</tr>
<tr>
<td>Prednisone 30 mg.</td>
<td>35</td>
<td>31</td>
<td>7</td>
</tr>
<tr>
<td>Phenylbutazone 300 mg.</td>
<td>48</td>
<td>47</td>
<td>21</td>
</tr>
<tr>
<td>NAPAP 7.5 g.</td>
<td>100</td>
<td>98</td>
<td>68</td>
</tr>
</tbody>
</table>

* One patient was excluded because the diagnosis of rheumatoid arthritis was found to be incorrect shortly after her entry into the trial, and the others because the forms on which toxicity was recorded were lost.

† In two further patients in the phenylbutazone trial and one in the NAPAP trial the trial courses were completed, but further courses were not given owing to toxic effects of calcium aspirin.

### TABLE VIII

**TOXIC EFFECTS**

<table>
<thead>
<tr>
<th>Trial</th>
<th>No. of Patients Experiencing Toxicity</th>
<th>Symptoms and Signs of Toxicity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Causing Withdrawal from Trial</td>
</tr>
<tr>
<td>Prednisone</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Aspirin</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Phenylbutazone</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>Aspirin</td>
<td>21</td>
<td>20</td>
</tr>
</tbody>
</table>

Note: Figures in brackets indicate the numbers of patients withdrawn from the trial or in whom the trials were discontinued because of the toxic manifestations indicated.
METHOD FOR TESTING ANALGESICS IN RHEUMATOID ARTHRITIS

the second week of the trial her blood pressure rose abruptly and the prednisone was discontinued. The blood pressure then fell rapidly to its previous level.

One patient complained of slight tinnitus and another of slight deafness in one ear, in each case at the end of the first week of the trial when they had received prednisone. Symptoms continued in the latter throughout the week on aspirin but rapidly subsided after she changed back to prednisone for a further week. In the other the tinnitus subsided after 2 days on aspirin. In view of the precautions that were taken to ensure that the patients received the correct drug (each dose of tablets for a patient being indicated on a chart which was initialled by the nurse giving it), it is unlikely that these two patients had received aspirin in error, but also unlikely that these symptoms were attributable to the prednisone. Wolf and Pinsky (1954) have called attention to the unexpected symptoms and occasionally signs which may follow administration of even an inert placebo.

Phenylbutazone.—Toxic effects were recorded in only three (7 per cent.) of the 44 patients who received it. In this trial, in view of the potentially serious side-effects of phenylbutazone, seven patients who gave a history of adverse reactions when phenylbutazone had been given previously and ten others who had conditions predisposing them to serious side-effects were excluded from the trial. The latter included five with known or suspected peptic ulcers, two with severe hypertension, two with auricular fibrillation, and one with a refractory anaemia. But for this selection the incidence of toxic effects observed with phenylbutazone would almost certainly have been higher. Thus one patient, who was included in the trial despite a history of having developed soreness of the mouth when previously taking phenylbutazone, again developed ulcerative stomatitis after 5 days on phenylbutazone in his first course and had to be withdrawn from the trial.

One patient who suffered from chronic eczema of the hands complained of increased itching without objective change in the rash after 3 days on phenylbutazone and was also withdrawn.

The third patient who complained of transient nausea while receiving phenylbutazone was not withdrawn from the trial.

NAPAP.—Toxic effects were encountered in 21 (22 per cent.) of the 97 patients receiving NAPAP, seventeen of them in the first course; six (6 per cent.) were withdrawn from the trial on this account, four of them during the first course. Seven of these 21 patients also experienced toxic effects from aspirin.

Gastro-intestinal disturbances, similar to those encountered with aspirin occurred in eighteen patients, in sixteen of them in the first course. Four patients were withdrawn during the first course and two during later courses on this account. Two of the eighteen patients also had gastro-intestinal upsets from aspirin; five others in the trial had gastro-intestinal disturbances on aspirin but not on NAPAP. Gastro-intestinal disturbances on one drug only were significantly commoner with NAPAP ($p < 0.03$).

One patient complained of short episodes of tinnitus and had an attack of vertigo and vomiting on NAPAP during the first course but completed a second course without trouble, and another who received NAPAP first noted mild deafness on this which became much worse with aspirin, regressed during the second week on NAPAP, and became so severe after changing back to aspirin that the trial regime was discontinued 2 days later.

Thus, in the dosages employed the incidence of toxicity was much the same with NAPAP and aspirin.

“Spontaneous Improvement.”—Several factors probably contribute to the spontaneous improvement with time which is usually observed after admission to hospital regardless of the medication. These may include the effect of rest in bed, increasing familiarity with hospital environment, and, in the case of grip, practice in the use of the apparatus. The natural history of rheumatoid arthritis is usually one of relapses and remissions; the tendency for patients to seek advice and to be admitted to hospital at the height of a relapse and therefore about to pass into some degree of spontaneous remission in the natural course of the disease may be important, but is probably less so in short-term trials of this type than in studies designed to assess the long-term effects of treatment on the disease.

Table IX (overleaf) shows estimates of spontaneous improvement. The method by which these are obtained is explained in the Appendix. Briefly, the figures in Table IX measure the improvement which went on between the first and second trial weeks, whatever the order in which the drugs were given. The figures represent the increases in mean grip and the decreases in the mean pain and stiffness ratings between the first and second weeks.

It will be seen that the spontaneous improvement is substantial in relation to the drug effects. In a clinical trial not designed to allow for it, an ineffective treatment could well be credited with substantial therapeutic value simply on account of the spontaneous improvement.
In view of the size of the standard errors, all the estimates for spontaneous improvement from the four trials are in moderately good agreement, except that the spontaneous improvement in grip seems definitely larger for the prednisone trial than the other trials. On a priori grounds it would be expected that, even apart from variations due to the small number of patients, the estimates of spontaneous improvement would not be exactly the same with different pairs of drugs owing to such factors as withdrawal effects.

The unusually large spontaneous improvement in grip in the prednisone trial is puzzling. If it were particularly small, explanations would readily suggest themselves, for example, that a "rebound" of symptoms on withdrawal of prednisone had partially masked the effects of aspirin or that the prednisone had resulted in such complete remission of symptoms that the effects of spontaneous improvement were obscured.

It is possible to obtain a more detailed insight by following the successive assessments separately rather than merely the changes from the mean of one week to the mean of the next. For grip the situation may be seen in Table X.

In the prednisone trial it will be seen that the grip strength of those who started on prednisone fell upon changing to aspirin, but recovered to the final level of the previous week which was very much above the level at the beginning of the trial. Too much emphasis should not be placed on these changes between successive assessments: the numbers of patients in this trial were relatively small and the chance variation was difficult to analyse since successive changes in the same patient could not be assumed to be independent. Comparison with the other trials, however, suggested that the fall in grip strength on changing from prednisone to aspirin could not be due entirely to a "steroid withdrawal rebound" since a fall was also observed on changing from aspirin in each of the other trials and also on changing from phenylbutazone to aspirin.

**Discussion**

Pope's dictum that "the proper study of mankind is man" is peculiarly apposite to the study of the effectiveness of drugs in relieving pain arising from human disease. Pain is a complex personal experience and its intensity and effects are influenced by many factors both within and without the individual experiencing it (Hardy, Wolff, and Goodell, 1952; Beecher, 1956). The effects of drugs noted in experiments on animals and healthy human beings subjected to painful stimuli may thus have little bearing on their value in relieving pain in clinical

<table>
<thead>
<tr>
<th>Assessment</th>
<th>HIPA</th>
<th>Prednisone</th>
<th>Phenylbutazone</th>
<th>NAPAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grip</td>
<td>130</td>
<td>111</td>
<td>106</td>
<td>98</td>
</tr>
<tr>
<td>Pain</td>
<td>130</td>
<td>111</td>
<td>106</td>
<td>98</td>
</tr>
<tr>
<td>Stiffness</td>
<td>130</td>
<td>111</td>
<td>106</td>
<td>98</td>
</tr>
</tbody>
</table>

**Table X**

**SEQUENCES OF ASSESSMENTS**

**MEAN GRIP STRENGTH (mm. Hg)**

<table>
<thead>
<tr>
<th>Trial</th>
<th>HIPA/Aspirin</th>
<th>Prednisone/Aspirin</th>
<th>Phenylbutazone/Aspirin</th>
<th>NAPAP/Aspirin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug given First</td>
<td>HIPA</td>
<td>Aspirin</td>
<td>Prednisone</td>
<td>Aspirin</td>
</tr>
<tr>
<td>1</td>
<td>148</td>
<td>78</td>
<td>130</td>
<td>111</td>
</tr>
<tr>
<td>2</td>
<td>142</td>
<td>91</td>
<td>147</td>
<td>116</td>
</tr>
<tr>
<td>3</td>
<td>144</td>
<td>107</td>
<td>158</td>
<td>112</td>
</tr>
<tr>
<td>4</td>
<td>180</td>
<td>64</td>
<td>141</td>
<td>136</td>
</tr>
<tr>
<td>5</td>
<td>149</td>
<td>70</td>
<td>148</td>
<td>150</td>
</tr>
<tr>
<td>6</td>
<td>174</td>
<td>82</td>
<td>160</td>
<td>161</td>
</tr>
</tbody>
</table>

**Note:** Assessments made during week on aspirin are in bold type.
practise. Assessment of the value of an analgesic agent in painful disease states is further complicated by the finding that a substantial proportion of those suffering from them may derive complete or partial relief from an inert placebo (Beecher, 1955; Lasagna Mosteller, von Felsinger, and Beecher, 1954; Traut and Passarelli, 1957; Orkin, Joseph, and Helrich, 1957) and that this “placebo effect” may have properties such as time relationship to administration similar to those of agents known to have definite analgesic potency (Lasagna, Laties, and Dohan, 1958). It seems probable that some at least of the spontaneous improvement noted between the first and second weeks in these trials may have been a “placebo effect” of admission to hospital. In the absence of a common yardstick, the problem of communication by the person experiencing it of an account of the pain such that an observer can form an accurate estimate of its severity in any absolute terms is insuperable and the best that can be achieved is an estimate of how the pain has changed during a given interval. Furthermore, when the effects of different drugs on the pain are being compared, unconscious bias on the part of the observers may seriously influence the results unless a “double blind” procedure is adopted so that both patient and observer are unaware which drug is being given at any one time.

Beecher (1957) has recently reviewed the many methods which have been devised for the clinical testing of analgesics, and some justification is necessary for the introduction of still another one. The method described is simple and can be readily incorporated in the routine management of patients admitted to hospital for treatment of inflammatory polyarthritis and is therefore suited to the screening of new analgesics which appear so frequently for use in rheumatic diseases. Administration of the drugs for a week as part of the treatment is more likely to give an idea of their relative effects in clinical practice and to induce less “trial consciousness” than methods employing frequent questioning of the patient after administration of a single dose. Restriction of the trial to new in-patients has the advantage that environmental variables can to some extent be controlled, certainty that a patient has taken the correct medication at the right times can be reasonably assured and the early signs of toxicity can be recognized. While change in grip strength over short periods of time is undoubtedly influenced by such factors as mood in addition to severity of pain; it is less influenced by these and also by similar factors operating on the observer than is the latter’s interpretation of the severity of the pain from the patient’s statements, and it thus provides a more objective criterion of change in pain intensity. Alternation in the order of drug administration permits estimation of the spontaneous improvement which may be appreciable in relation to the effect of medication. These trials have also shown that the method can achieve discrimination between drugs with a reasonably small number of patients.

In view of the high dosage of prednisone employed, 30 mg. per day, its superior analgesic effects as judged by this method are not surprising. The result with phenylbutazone is at variance with that obtained by Fjellström, Goldberg, Lindgren, and Nilsson (1957) in the only comparable controlled trial of which we are aware; they found a greater decrease in the number of joints tender to pressure and painful on passive motion in patients receiving phenylbutazone in addition to aspirin *ad lib.* than in those receiving aspirin *ad lib.* alone. The discrepancy may be due to their patients having received a total of 5 g. phenylbutazone in the first week as compared with 2·1 g. in this study, and to their control patients having taken a substantially lower dose of aspirin. The finding by this method that NAPAP has analgesic properties inferior to those of aspirin is in accordance with that of Orkin and others (1957), who compared the effects of 0·6 g. aspirin with those of 0·5 g. NAPAP in relieving postpartum cramps. Wallenstein and Houde (1954), whose patients were suffering from pain due to cancer, found little or no difference between equal dosages of NAPAP and aspirin. Batterman and Grossman (1955) concluded that, in patients with painful musculo-skeletal disorders of various types, with the exception of those suffering from “exudative forms of arthritis”, NAPAP in comparable dosage was superior to aspirin as an analgesic. This unresponsive category of patients was not further defined, but would presumably include most patients with active inflammatory polyarthritis such as took part in the present trials.

Much publicity has attended the recent reports of the finding of occult gastro-intestinal bleeding in a high proportion of patients receiving aspirin (Stubbé, 1958; Alvarez and Summerskill, 1958; Porter, Lewis, and Dixon, 1958), but, while in occasional individuals this may result in serious anaemia (Summerskill and Alvarez, 1958), in the majority it is not a serious problem (Duthie, 1958).

The low incidence of toxicity observed with prednisone was related to the short duration of therapy in this trial, because such high dosage if given for prolonged periods, would undoubtedly be attended by serious side-effects. The low incidence of toxicity observed with phenylbutazone is also misleading, partly on account of the short duration of
therapy, but also because patients who had previously had toxic reactions or who were predisposed to toxic reactions by the nature of their disease were excluded. With NAPAP, toxic effects were observed with approximately the same frequency as with aspirin. The results of these trials would therefore support the contention that, of the agents tested, aspirin is still the most satisfactory analgesic for routine use in patients with polyarthritis of the rheumatoid type.

The principal advantage of sequential analysis over the conventional procedure, by which the number of patients in a trial is fixed independently of the results, is a saving in the number of patients required (Armitage, 1954). If a conventional fixed-sample trial is to be as effective in discriminating between drugs as the sequential procedure chosen, about forty patients would be needed (twenty receiving the medications in each order). It is assumed, of course, that apart from the sequential aspect, the trial arrangements would be the same.

In a sequential procedure, the number of patients varies from trial to trial. The average length of trials conducted according to the chosen sequential procedure cannot be precisely computed. It seems likely that the saving in patients in comparison with the conventional method will often be rather small. If the difference between the two drugs compared is roughly equal to the target difference of the sequential test, the average number of pairs of patients required would probably be about fifteen. However, data for a greater number will probably be collected, because some patients would have begun the trial when it is decided to stop further intake—and indeed there are advantages in collecting the additional data, as explained above. Under these circumstances the average saving in patients would be slight. If the difference between the drugs compared were about half the target difference, there would probably be no saving at all.

The main argument for sequential analysis as used in the present trials must, therefore, be the likelihood that a trial would stop early if the difference between the drugs being compared is large—considerably larger than the target difference. This advantage would be more appreciable in circumstances where the length of the corresponding fixed sample size trial is greater, i.e. in trials where a finer discrimination between drugs is attempted so that a greater number of patients will usually be required (this means either that more stringent significance levels are used, or a smaller target difference is chosen). For example, in relation to a fixed sample trial involving 100 patients, the advantage of a sequential method which might detect a large difference between drugs with only thirty patients is readily apparent.

Against this must be set the risk of an occasional long run. This is a more serious disadvantage than appears at first sight. In a trial of fixed size, the duration of which can be predicted with reasonable accuracy, it is possible to maintain the quality of the data. We found that the longer the trial continued the greater was the proportion of inadequate records.

Statistical methods are much better developed for conventional trials of fixed size. Hence much greater flexibility is possible both in planning and in the analysis of the results. A special disadvantage of sequential analysis is the need to use a single index. An awkward form of this difficulty is presented by toxic reactions. Patients suffering from toxic reactions serious enough to cause their withdrawal from the trial are, in the present scheme, simply left out of the sequential decision procedure. Such toxicity ought to weigh heavily against the medication presumed responsible, but it is not easy to see how to take it into account. Again, a procedure based on a single index might conceivably give misleading results if one of the drugs tested caused a withdrawal rebound in symptoms. In the worst case a drug might have little analgesic effect, though its withdrawal caused a considerable increase in symptoms. The present scheme of sequential analysis might thus make it appear a good analgesic. To some extent this difficulty is common to all methods involving the administration of drugs for short periods in succession, however they may be analysed; but, whereas detailed examination of the data might well reveal the withdrawal effect, analysis based on a single criterion might prove unsatisfactory.

The saving in the number of patients which sequential methods are designed to achieve could be more usefully exploited if two modifications could be introduced. These modifications are not within the scope of existing statistical techniques, but as statistical theory develops, both modifications may soon become feasible. In the first place the present procedure combines results for pairs of patients who received the drugs in opposite order, and ignores their separate contributions. Thus the observations collected are not fully exploited in the estimation of variability. It would be desirable to have a procedure not subject to this criticism. An approximate method devised by Rushton (1952) may supply what is needed here. This problem is part of the reason why the saving in the number of patients secured by the sequential analysis is in the present instance relatively small. Secondly, the risk of occasional long runs would be avoided...
if a procedure were devised by which a fixed number* of patients is included in a trial, but with the possibility of stopping early if large differences appear. Armitage (1957) has worked out such schemes for qualitative data (i.e. data of the type where patients are classified as “improved” or “not improved”); he has called them restricted sequential schemes. So far no such schemes are available for trials where the observations are measurements.

The sequential analysis employed in the present trials has proved entirely practicable. However, it has serious drawbacks. In the absence of new developments in statistical theory, its advantages are outstanding only where the corresponding conventional trial would require a large number of patients and it is important to be able to stop early if the difference between the drugs considerably exceeds the “target” difference.

**Summary**

A method for comparing the short-term symptom-relieving properties of two drugs in patients with rheumatoid arthritis is described. By this method patients on admission to hospital receive either aspirin or the drug being compared, for one week, and the other drug in the second week. The patient is questioned at three specified times in each week regarding the severity of pain and stiffness by an observer who is unaware which medication he is receiving, and his strength of grip is recorded.

The differences between drugs were found to be largest in relation to other fluctuations when measured by grip strength, and this was used as the “index” in a sequential procedure which indicated when sufficient observations had been accumulated to determine that there was either a difference of predetermined magnitude, or no significant difference, between the effects of the drug being tested and those of aspirin. It was found that the strength of grip tended to be better in the second week in hospital than in the first week, independent of the medication, while pain and stiffness tended to be less. The effects of this spontaneous improvement may be large enough to obscure important differences between drugs. Adjustments were made for these effects in the analysis of the data.

Four substances were compared with aspirin 5 g. daily.

(1) 4-hydroxy-isophthalic acid (HIPA) 6.7 g. daily.
(2) Prednisone 30 mg. daily.
(3) Phenylbutazone 300 mg. daily.
(4) N-acetyl-p-aminophenol (NAPAP) 7.5 g. daily.

In the dosage given in these trials, prednisone was found to be greatly superior to aspirin. Phenylbutazone and aspirin were similar, while both HIPA and NAPAP were inferior to aspirin.

Data on suspected toxic reactions are reported. The toxicity of aspirin appeared fairly constant throughout the four trials, some 28 per cent. of patients reporting toxic manifestations. The toxicity of prednisone and phenylbutazone was lower than that of aspirin, in the latter case presumably because patients who gave a history of adverse reactions to phenylbutazone or who had conditions predisposing them to serious side-effects had been excluded from the trial. Toxic reactions were recorded at least as frequently for NAPAP as for aspirin.

The trial procedure described can readily be incorporated into the routine management of patients hospitalized for rheumatoid arthritis with little disturbance or additional labour. Trials of this kind seem more appropriate for evaluating the performance of analgesics in clinical practice than methods relying on intensive questioning after administration of a single dose.

The method permits useful discrimination between analgesics with a relatively small number of patients. The degree of discrimination arrived at in the present trials would take some forty patients in a trial of fixed sample size. In fact, the number of patients included varied from trial to trial, since it was determined by the sequential procedure mentioned. This has the advantage of permitting a trial to stop early if large differences appear between the effects of the drugs compared.

The disadvantages of the procedure are also discussed.

The statistical methods used are described in an Appendix.

We are greatly indebted to Prof. J. H. Kellgren for much advice in the planning of these trials which were carried out on patients under his care. These studies would have been impossible without the willing co-operation of Dr. E. J. Epstein, Dr. R. Whitaker, Dr. L. J. Atkinson, and Dr. K. D. Coorey, who were the Medical Residents during the periods of these trials, and of the Nursing Staff of the Manchester Royal Infirmary and of the Devonshire Royal Hospital, Buxton, particularly Sister Whittaker.

We are also grateful to Mr. J. C. Hanbury of Messrs.
Allen and Hanbury, Ltd., who kindly provided the tablets of 4-hydroxy-isophthalic acid, to Dr. R. S. Inch of Bayer Products, Ltd., for supplies of N-acetyl-p-aminoophenol and placebo, and for information regarding other trials of this drug, and to Dr. W. S. Stoddart of Geigy Pharmaceutical Company Ltd., for similar information about phenylbutazone. Mr. J. B. Lloyd and the staff of the Pharmaceutical Department, Manchester Royal Infirmary, prepared the other special tablets used in these trials.

REFERENCES


STATISTICAL APPENDIX

(1) THE SEQUENTIAL TEST USED

The test employed to determine the length of the last three trials was a sequential t-test.

Suppose that we have a sample of patients, each of whom has been treated with two drugs, and that some index of the difference in effectiveness between the two drugs is available for each patient. To apply a conventional t-test, we compute the average of this index and divide it by its standard error. This ratio is then compared with a theoretical ratio from a t-table.

In a sequential t-test we perform such calculations each time that an additional observation has been obtained. At any observation we may either (i) stop, because we already have enough evidence for a decision (average difference very small or very large in relation to fluctuations), or (ii) continue the trial, because there is not yet sufficient evidence for a decision (average difference intermediate in relation to fluctuations).

In practice it is not necessary to compute explicitly the ratio of the average difference to its standard error at every step. It is more convenient to calculate a related quantity, the sequential criterion (z) defined as:

\[ z = \frac{\bar{d}}{\sigma_d / \sqrt{n}} \]

where \( \bar{d} \) is the observed difference between the means of the two groups, \( \sigma_d \) is the standard error of the difference, and \( n \) is the number of observations.

The significance of this criterion can be determined from tables of the standard normal distribution.

(2) SUM OF SQUARES OF OBSERVATIONS

At each stage we compute this criterion using all the observations so far accumulated as illustrated in Tables II and III. The reason why this criterion can be used is that we can find stopping limits in terms of it. There are various ways of doing this. The limits may be computed by means of an approximation developed by Rushton (1952). However, by far the easiest method of obtaining stopping limits is to look them up in a convenient set of Tables published in 1951 by the U.S. National Bureau of Standards.

These Tables list many sets of limits by which we could guide our trial procedure. To choose between them we must decide what properties our procedure is to have. We can use the Tables if we specify:

(a) Significance levels. *
(b) How big a difference (in relation to fluctuations) we want to be fairly sure of detecting; this will be called the “target” difference.

However, in order to make a choice under (b), we need:

(c) A single index to measure the difference between drugs.

The selection of procedure in all these respects involves a substantial arbitrary element.

As for (a) significance levels, we have retained the conventional 5 per cent.

Discussion of points (b) and (c) hinges upon the magnitude of the fluctuations† to which the observations are subject. Drug differences when measured by the chosen index must appear large in relation to the fluctuations in the index caused by other factors. Secondly, the “target” difference must be set smaller than such differences between drugs as experience has shown to be appreciable. The preliminary experiments mentioned before suggested that in terms of grip the difference between aspirin and HIPA was of the order of one standard deviation. In terms of pain and stiffness (as well as other measurements), the difference seemed considerably smaller. A larger number of patients is required to detect a smaller difference. It was, therefore, decided to use grip strength as the index for the sequential analysis, since this would make it possible to ascertain the existence of differences between drugs (or absence of difference) with the smallest number of patients. The fact that on the ratings for pain and stiffness it would be difficult for many patients to record

* In sequential t-tests two significance levels must be chosen:

(i) The probability of concluding falsely, that there is a difference between the drugs when there is in fact none:

(ii) The probability of concluding falsely, that there is no difference when in fact there is a difference of the magnitude specified under (b).

It is not necessary that the same figure should be used for both significance levels, but we have in fact done this in the trials reported here. In a trial with fixed sample size, only the probability of errors of type (i) is explicitly specified in the statistical test.

† Other considerations relevant to the choice of an index are dealt with above, pp. 190 and 200.
a substantial improvement over the level reached when they are treated with aspirin, is also an argument in favour of using grip strength as an index (at least in trials where aspirin is the standard of comparison). An attempt was made (by discriminant analysis) to find a combined index of several criteria which might give better discrimination between drugs than grip strength alone. However, on the data available, the improvement over grip strength alone achieved by the best possible weighted combination of several types of observation was small, and it seemed likely that a combined index with weights fixed in advance would often be no better, if not worse, than grip alone.

Since the difference between HIPA and aspirin appeared to be about one standard deviation in terms of grip, it was felt adequate to set the “target” difference between drugs (the difference which would be detected fairly certainly, i.e. with 95 per cent. probability) at 0·6 of a standard deviation. It seemed from the preliminary experiments that there was little difference between HIPA and a placebo. A trial with a target difference of 0·6 standard deviation should therefore detect any medication which is as much superior to aspirin as aspirin is superior to an inert substance; it would also reveal any medication which is not much better than a placebo.

It would be tempting to set the target difference at a lower figure, i.e. to increase the sensitivity of the trial. However, the number of patients required increases fairly rapidly if this is done.

In accordance with these arguments it was decided to determine the length of the first sequential trial (prednisone/aspirin) by a sequential t-test with:

(a) 95 per cent. significance levels,
(b) A target difference of 0·6 standard deviation,
(c) Using a difference of grip strength as the index of the difference in drug effect.

A snag which was not entirely unexpected appeared almost at the start. A very pronounced “spontaneous improvement” effect became evident, i.e. the results differed strikingly according to the order in which the patient had received the drugs. To allow for this effect (otherwise the trial would have been very prolonged), it was decided to amalgamate the observations for pairs of consecutive patients such that one member of a pair received the drugs in the reverse order from the other member. The observations for the two patients in a pair were added and the sum treated as a unit for the sequential analysis, as explained in the text. The target difference was altered to 0·85 of a standard deviation. This figure for a test of sums of pairs of observations corresponds to a target difference set at 0·6 of a standard deviation for a test using individual observations.*

This procedure was retained unchanged throughout the subsequent trials. The appropriate stopping limits are given in Table III above.* It may be asked how this procedure looks in the light of the evidence on variability in the observations which has thus been accumulated. Data relating to this question are set out in Appendix Tables A and B.

### APPENDIX TABLE A

<table>
<thead>
<tr>
<th>Assessments</th>
<th>HIPA</th>
<th>Prednisone</th>
<th>Phenylbutazone</th>
<th>NAPAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grip (mm. Hg)</td>
<td>17·7</td>
<td>20·9</td>
<td>14·1</td>
<td>13·1</td>
</tr>
<tr>
<td>Pain (Arbitrary rating)</td>
<td>0·79</td>
<td>0·39</td>
<td>0·42</td>
<td>0·50</td>
</tr>
<tr>
<td>Stiffness (Arbitrary rating)</td>
<td>0·41</td>
<td>0·49</td>
<td>0·43</td>
<td>0·46</td>
</tr>
</tbody>
</table>

Note: Figures given are the square roots of variances computed by Formula (2) below.

### APPENDIX TABLE B

<table>
<thead>
<tr>
<th>Assessments</th>
<th>HIPA</th>
<th>Prednisone</th>
<th>Phenylbutazone</th>
<th>NAPAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grip</td>
<td>1·40</td>
<td>0·74</td>
<td>0·11</td>
<td>0·43</td>
</tr>
<tr>
<td>Pain</td>
<td>0·42</td>
<td>0·53</td>
<td>0·07</td>
<td>0·35</td>
</tr>
<tr>
<td>Stiffness</td>
<td>1·03</td>
<td>0·24</td>
<td>0·07</td>
<td>0·18</td>
</tr>
</tbody>
</table>

Note: Figures in Table VI divided by corresponding figures in Table A above. Discrepancies are due to rounding.

Table A shows the standard deviations of the differences between the mean assessments in the two trial weeks. For stiffness the standard deviations in the four trials are practically identical. The figures for grip and pain show variations which are statistically significant.† There may well have been some change in the variability of the observations as the trials proceeded (for example, owing to the changes in admission practices, or differences between the observers at successive periods). Such changes as occurred seem to have been small.

Table B gives the ratios of the drug effects in Table III above to the standard deviations in Table A. The selection of grip for use in the sequential analysis is clearly justified. The drug effects as measured by grip are considerably larger in relation to the standard deviation, than as measured by pain and stiffness. The target difference of 0·6 of a standard deviation;* corre-

* The National Bureau of Standards Tables contain stopping limits for target differences of 0·8 or 0·9 standard deviations. The limits for 0·85 standard deviations can be taken to be halfway between those for 0·8 and 0·9. As it happens, the value of 0·85 was used by Armitage (1954), and the stopping limits are given in his paper.

† Bartlett’s test of the homogeneity of variances, e.g. Quenouille (1950), gives χ² of 9·6, 10·7, and 0·5 for grip, pain, and stiffness respectively. With three degrees of freedom, the probabilities that these figures be exceeded are about 0·03, 0·02, and 0·95.

* The standard deviations involved relate to the population of observations from individual patients, not to sums of observations from pairs. Hence the figures in Table B should be compared with the target difference of 0·6 (not 0·85).
responding to 0.85 for sums of pairs) also seems not unreasonable for detecting drug effects in terms of grip, though for NAPAP the difference seems to be below the target figure. This, and the fact that the ratio for prednisone in Table B is not very greatly in excess of 0-6, are perhaps arguments for choosing a smaller target difference.

(2) ESTIMATION OF SPONTANEOUS IMPROVEMENT

An experiment in which two drugs are given to patients in different orders may be regarded as an example of what is sometimes termed a "cross-over" design. It has been employed in many types of experimentation. Apart from other advantages, it has the merit that it is possible not only to make allowance for the order of administration of the drugs in estimating the drug effects, but also to obtain direct estimates of the effects of the order of administration, i.e., the spontaneous improvement. The analysis of a "cross-over" trial (i.e., computation of both drug effect and spontaneous improvement with standard errors, etc.) is a simple matter and within the capacity of anyone with an elementary knowledge of statistics. The procedure is therefore briefly explained here.* There is no necessary connexion between cross-over designs and sequential methods.

Suppose that two drugs X and Y are compared. If we divide the patients into groups (according to the order in which they receive the drugs) and calculate for each group the mean grip strength for each week, we might obtain the following Table:

<table>
<thead>
<tr>
<th>Week</th>
<th>First</th>
<th>Second</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-Starters</td>
<td>(X) 100</td>
<td>(Y) 150</td>
</tr>
<tr>
<td>Y-Starters</td>
<td>(Y) 170</td>
<td>(X) 160</td>
</tr>
</tbody>
</table>

We make the assumption that the effects of medication and spontaneous improvement add together. We can then regard the figures in the above Table as being made up of the following four factors:

(a) The mean grip of the X-starters over the total trial period (125).
(b) The mean grip of the Y-starters over the total trial period (165).
(c) The drug effect, i.e., the difference between the effects of Drug X and Drug Y. Mean under Drug Y minus mean under Drug X = 30.
(d) Spontaneous improvement, i.e., the difference between the first and second week. Mean of second week minus mean of first week = 20.

The values given in brackets will be found to fit the figures of the Table. For the X-starters, the drug effect and spontaneous improvement reinforce each other and cause a rise of 50 points between the first and second weeks. For the Y-starters, drug effect and spontaneous improvement work in opposite directions and cause a drop of 10 points (i.e., 30 - 20 = 10).

The individual figures in the Table may be expressed in terms of the four factors indicated. For example, the figures for the Y-starters may be represented as follows:

For the first week:

Factor (b) + 1/2 factor (c) - 1/2 factor (d)  
\[ i.e. 165 + 30/2 - 20/2 = 170 \]

For the second week:

Factor (b) - 1/2 factor (c) + 1/2 factor (d)  
\[ i.e. 165 - 30/2 + 20/2 = 160 \]

There are various ways of computing figures for the drug effect and spontaneous improvement from data such as those given in the Table above. One method is to take the difference between the X-week and the Y-week for each group of patients. Suppose we subtract the X-value from the Y-value, as follows:

For X-starters 150 - 100 = 50
For Y-starters 170 - 160 = 10

The drug differences and spontaneous improvement are then respectively half the sum, and half the difference between these two.

Thus:

Drug Effect \[ = \frac{1}{2}(50 + 10) = 30 \]
Spontaneous Improvement \[ = \frac{1}{2}(50 - 10) = 20 \]

The figures given in Table V and VI are drug effects obtained in this manner as already explained in the text.

This way of computing estimates of the drug effect and spontaneous improvement gives a clue to the random fluctuations which we may expect to occur in these estimates as trials are repeated with different groups of patients. It is seen that the level of grip of the patients is irrelevant, only the difference between the values recorded in the two trial weeks affects the estimates. Hence we can obtain standard errors for the estimates in terms of these differences.

This method of computation also indicates that the same standard errors apply for estimates of the drug effect and for estimates of spontaneous improvement. The precise formula is as follows: As above, page 191, let the observations (i.e., differences between the mean assessments in the two weeks for successive patients) be denoted by \( x_1, x_2, \ldots \) for patients who received the drug in one order, and by \( y_1, y_2, \ldots \) for patients for whom the order was reversed. Let \( N_x \) and \( N_y \) be the numbers of the two kinds of patients. We wish to estimate the variance of the estimated drug effect (or of spontaneous improvement).

If we denote the desired variance estimate by \( s^2 \), we have

\[ s^2 = \frac{\sum_{i=1}^{N_x} (x_i - \bar{x})^2 + \sum_{i=1}^{N_y} (y_i - \bar{y})^2}{N_x + N_y - 2} \]  \( (1) \)

The standard error may be written (using the square root \( s \) of this expression):

\[ \text{Standard Error} = \frac{s}{\sqrt{\frac{1}{N_x} + \frac{1}{N_y}}} \]  \( (2) \)

\( \bar{x} \) and \( \bar{y} \) are the means of the \( x_i \) and \( y_i \) respectively.

* For a full technical discussion, see Fieller (1940).
For practical computation it is, of course, more convenient to replace the two expressions in the numerator of (1) by the usual short-cut formulae used in the computation of variances, i.e.

$$
\sum (x_i - \bar{x})^2 = \frac{\sum x_i^2}{N} - \bar{x}^2,
$$

and

$$
\sum (y_i - \bar{y})^2 = \frac{\sum y_i^2}{N} - \bar{y}^2.
$$

It may be noted that the analysis of "spontaneous improvement" can be carried out in the simple form described here only because quantitative variables are involved. The percentages which occur when the observations are qualitative (e.g. "improved" and "not improved") are more difficult to handle in such an analysis.

(3) COMPARISON BETWEEN SEQUENTIAL AND FIXED SAMPLE SIZE TESTS

One difficulty with sequential $t$-tests is that it is not at present possible to compute the average sample size which would be required if trials were repeated many times under identical conditions. From the properties of other sequential tests and from sampling experiments reported in the introduction to the National Bureau of Standards Tables, one may reasonably guess that if the difference between the drugs being compared is, in fact, at the target level a trial by the procedure described would on average require about fifteen pairs of patients. Under other circumstances, the average number of patients required would be different, as explained on pp. 193 and 203 above.

How large a conventional fixed-sample trial would correspond to one using the sequential procedure selected, the trial arrangements being the same except for the sequential aspect? We assume that the trial would comprise two groups of patients (equal, or nearly equal, in number) who are given the drugs in different orders. The analysis of a trial of such design (a "crossover" design) has been described in the previous section. The drug effects would be estimated, as in Table V above, by taking the mean of the results for the two groups. Their standard errors would be given by Formula (2) above. The ratios of the drug effects to their standard errors could be tested by a $t$-test.

If we use a 5 per cent. significance level in such a test, we secure the same chance as in the sequential procedure of wrongly concluding that there is a difference between the drugs when in fact there is none. How many patients need to be included in the trial if there is to be only a 5 per cent. chance, as in the sequential procedure, of concluding falsely that there is no difference between two drugs when this difference is in fact at the target level (i.e. $0.06$ of a standard deviation)? The answer depends upon what is technically known as the "power" of the test used. It turns out upon investigation that about forty patients are needed, roughly twenty receiving the drugs in each order.

We may calculate the approximate precision with which the drug effects could be estimated from a fixed-sample trial involving forty patients. Confidence intervals, at the 95 per cent. level, would have a length of some 10 mm. Hg for grip strength and 0·3 points on the arbitrary rating scale for pain and stiffness. (The width of a 95 per cent. confidence interval is 3·9 times the standard error.)

No similar statement regarding the precision of estimation is possible for the sequential procedure, since the number of patients included would vary from trial to trial. Indeed, methods of estimating and of obtaining confidence intervals or other indicators of precision when the length of a trial has been determined by sequential methods are not well established (Anscombe, 1954; Armitage, 1958). It seems likely that, in the circumstances of the present trials, standard errors and the confidence intervals associated can be interpreted as if the number of patients in the trials had been fixed independently of the results.

Evaluation des analgésiques dans l'arthrite rhumatoïde à l'aide d'un procédé séquentiel

RéSUMé

On décrit une méthode pour comparer l'effet analgésique temporaire de deux médicaments chez des malades atteints d'arthrite rhumatoïde. D'après cette méthode, pendant la première semaine de leur séjour à l'hôpital, un groupe de malades était traité par l'aspirine et un autre groupe par le médicament étudié; pendant la seconde semaine on renversait le traitement dans chaque groupe. Trois fois par semaine, à des temps déterminés, un observateur qui ignorait l'identité du médicament, interrogeait les malades sur la sévérité de la douleur et de l'enraidissement et notait la force de leur poigne.

On a trouvé que la différence entre les médicaments était la plus grande, par rapport aux autres fluctuations, quand on la mesurait par la force de la poigne. On a choisi celle-ci comme "indice" dans le calcul séquentiel qui indiquait quand le nombre d'observations était suffisant pour déterminer soit une différence de magnitude prédéterminée, soit une différence insignifiante entre les effets du médicament étudié et ceux de l'aspirine. On a noté que la force de la poigne tendait à augmenter au cours de la deuxième semaine à l'hôpital, indépendamment du traitement, tandis que la douleur et l'enraidissement tendaient à diminuer. La magnitude de cette amélioration spontanée aurait pu masquer des différences importantes entre les médicaments et on en a tenu compte au moment d'analyser les données.

Quatre substances ont été comparées à 5 g. d'aspirine par jour:

1. Acide 4-hydroxy-isophthalique (HPA) 6,7 g. par jour.
2. Prednisone 30 mg. par jour.
3. Phénylbutazone 300 mg. par jour.
4. N-acétyl-p-aminophénol (NAPAP) 7,5 g. par jour.

Aux doses administrées au cours de ces recherches, la prednisone s'est avérée très supérieure à l'aspirine. La phénylbutazone et l'aspirine étaient similaires, tandis que l'HPA et le NAPAP étaient inférieurs à l'aspirine.

On rapporte les données concernant les réactions toxiques. La toxicité de l'aspirine était assez constante au cours des quatre expériences, et à peu près 28 pour cent des malades a révélé des manifestations toxiques. La toxicité de la prednisone et de la phénylbutazone...
etait inférieure à celle de l'aspirine; ceci, en ce qui concerne la phénylbutazone, peut s'expliquer par le fait que les malades atteints d'arthrite rhumatismale, sans dérangement ou travail excessif. Des expériences de ce genre semblent être plus appropriées pour évaluer l'efficacité des analgésiques en pratique clinique, que des méthodes qui se basent sur une interrogation intensive après l'administration d'une seule dose.

Cette méthode permet une distinction pratique entre des analgésiques avec un nombre relativement petit de malades. Le degré de distinction obtenu dans le travail présent nécessiterait à peu près quarante malades pour une expérience avec des échantillons à grandeur fixe. En réalité, le nombre de malades inclus variait d'une série à l'autre, étant déterminé par le calcul séquentiel déjà mentionné. Celui-ci offre l'avantage de pouvoir terminer rapidement une expérience aussi tôt que la différence entre les effets des médicaments comparés atteigne la magnitude suffisante.

On discute aussi les inconvénients de ce procédé. La description des méthodes statistiques employées se trouve dans l'appendice.

Valoración de analgésicos en la artritis reumatoide utilizando un procedimiento secuencial

**SUMARIO**

Se describe un método para comparar el efecto analgésico temporal de dos medicamentos en enfermos con artritis reumatoide. Según este método, durante la primera semana en el hospital, un grupo de enfermos fue tratado con aspirina y un otro grupo con el medicamento estudiado; a la semana siguiente el tratamiento fue invertido en cada grupo. Tres veces por semana, en tiempos determinados, un observador que desconocía la identidad del medicamento, interrogaba los enfermos acerca de la severidad del dolor y de la rigidez y notaba la fuerza del empuñamiento.

Se notó que la diferencia entre los medicamentos fue mayor, en relación a otras fluctuaciones, cuando se la media por la fuerza del empuñamiento. Esta fue elegida como "indicio" en el cálculo secuencial que indicaba el momento en que el número de observaciones era suficiente para determinar sea una diferencia de magnitud predeterminada, sea una diferencia sin importancia entre los efectos del medicamento estudiado y los de la aspirina. Fué observado, que la fuerza del empuñamiento tendía a aumentar durante la segunda semana de hospitalización, independientemente del tratamiento, mientras que el dolor y la rigidez tendían a disminuir. La magnitud de esta mejora espontánea hubiera podido disfrazar diferencias importantes entre los medicamentos; este factor fué tenido en cuenta al analizar los datos.

Cuatro substancias fueron comparadas a 5 g. de aspirina diaria:

1. Acido 4-hidroxi-isofálico (HIPA) 6.7 g. diarios.
2. Prednisona 30 mg. diarios.
3. Fenilbutazona 300 mg. diarios.
4. N-acetil-p-aminofenol (NAPAP) 7.5 g. diarios.

A las dosis administradas en estas investigaciones, la prednisona se mostró muy superior a la aspirina. Hubo similitud entre los efectos de la aspirina y de la fenilbutazona, mientras que el HIPA y el NAPAP se mostraron inferiores a la aspirina.

Se relatan los datos respecto a las reacciones tóxicas. La toxicidad de la aspirina fue bastante constante en las cuatro investigaciones, y cerca de un 28 por ciento de los enfermos acusó manifestaciones tóxicas. La toxicidad de la prednisona y de la fenilbutazona fue inferior a la de la aspirina; respecto a la fenilbutazona, esto se puede explicar por el hecho de que los enfermos con antecedentes de reacciones secundarias a este medicamento o con disturbios predisponentes a tales reacciones, fueron excluidos de la investigación. Las reacciones tóxicas con el NAPAP fueron al menos tan frecuentes como con la aspirina.

El procedimiento descrito aquí se puede aplicar fácilmente en un hospital con enfermos reumáticos con disturbio o trabajo excesivo. Investigaciones de este tipo parecen más apropiadas para valorar la eficacia de analgésicos clínicamente, que métodos basadas en una interrogación intensiva después de la administración de una sola dosis.

Este método permite un discernimiento práctico entre analgésicos con un número relativamente pequeño de enfermos. El grado de discernimiento obtenido en el trabajo presente necesitaría aproximadamente cuarenta enfermos, en un cálculo con muestras de magnitud fija. En realidad, el número de casos incluidos fué variable en todas las series, siendo determinado por el cálculo secuencial ya mencionado. Este ofrece la ventaja de poder cortar rápidamente una serie tan pronto como las diferencias entre los efectos de los medicamentos comparados adquieran magnitud suficiente.

Se discuten también las desventajas de este procedimiento. La descripción de los métodos estadísticos empleados encuéntrase en el apéndice.