AN ANALYSIS OF A SET OF MEDICAL DATA USING THE BOOTSTRAP PROCEDURE

AN ANALYSIS OF A SET OF MEDICAL DATA USING THE BOOTSTRAP PROCEDURE

BY

LORRAINE MIGNON TAWFIK, B.Sc.

A Project
Submitted to the School of Graduate Studies
in Partial Fulfillment of the Requirements
for the Degree
Master of Science

McMaster University
June 1984

MASTER OF SCIENCE (1984)
(Mathematical Science)

McMASTER UNIVERSITY
Hamilton, Ontario.

TITLE:	An Analysis of a Set of Medical Data Using the Bootstrap Procedure
AUTHOR:	Lorraine Mignon Tawfik, B.Sc. (McMaster University)
SUPERVISOR:	Professor C.W. Dunnett
NUMBER OF PAGES:	v, 115

ABSTRACT

The efficacies of two anti-inflammatory drugs in ankylosing spondylitis and related complaints were studied at a single medical clinic over a period of twenty-eight weeks.

The purposes of this project were:

(1) To determine any significant differences within and between the two drug groups using well-known nonparametric procedures, and

(2) To illustrate the use of the bootstrap method and determine whether it is appropriate and useful for this data set.

Some statistically significant changes indicative of improvement occurred among both groups of patients for primary efficacy variables. No definite trend was found for most of the laboratory variables.

Both drugs demonstrated effective pain relief. Regarding the variables of day and night pain relief as well as pulse, the Experimental Drug proved to be clinically but not statistically superior to the other commonly used drug. Analyses of safety data indicated some statistically significant changes in both drug groups. There was a statistically significant difference between drug groups at baseline.

ACKNOWLEDGEMENT

I would like to express my sincere appreciation to my supervisor, Dr. C.W. Dunnett, and to Dr. C.H. Goldsmith for their suggestions, patience and guidance during the final stages of this project.

Special thanks must go to Mrs. Maria Wong of CSU for her valuable help and assistance in the use of the HP3000 with the MINITAB and GRAPHICS program.

Further, I am privileged to have gained experience at Wyeth Ltd. which is pertinent to this work.

TABLE OF CONTENTS

I. INTRODUCTION

Ankylosing spondylitis is a form of arthritis which "principally affects the spinal column, reducing its mobility. It is progressive, and tends to affect young adults". (The Reader's Digest, 1964)

The following project deals with two different drugs used to treat this disease and their subsequent effects. A double blind trial was conducted by a doctor in Quebec in 1981 to assess the safety and efficacy of indomethacin versus an experimental new drug, which will be known thereafter as Experimental Drug #1.

Indomethacin (Indocid ®) is an anti-inflammatory, analgesic drug widely used in the symptomatic treatment of rheumatoid arthritis, osteoarthritis and degenerative hip joint disease as well as ankylosing (rheumatoid) spondylitis. It may replace other commonly used agents such as corticosteroids, salicylates, phenylbutazone and colchicine.

In order to assess the clinical efficacy of two drugs, parametric procedures have traditionally been used in studies involving human subjects if normality can be assumed and a large sample size (greater than 30) can be obtained. However, in this clinical trial only a very small sample (10 subjects) was obtainable and also the usual normality assumption was violated. Thus, nonparametric procedures were substituted for parametric ones in order to analyse the drugs' efficacies.

However, the question arises: Is the bootstrap procedure as rigorous as a corresponding nonparametric procedure (i.e. the Wilcoxon signed rank test) in being able to detect differences between groups based on such a small sample size? One approach to this question would be to compare these nonparametric methods with another statistically acceptable method, if one were available.

The relatively new statistical tool called the bootstrap procedure would appear to fill this role, since it does not assume normality nor does it require a large sample size. In the following, the bootstrap procedure is explored in order to determine its usefulness in assessing the efficacies of the two drugs used in this study, and its results will be compared to those obtained using the nonparametric methods only.

II. DATA DESCRIPTION

1. Sample

The sample consisted of ten out-patients with ankylosing spondylitis who were treated at a single medical center. These patients were randomly assigned to one of the two treatments in equal numbers; namely, five to each group. Thus, five were on the standard medication (Indocid) and five were on an experimental new drug. Neither the physician nor the patients were aware of which drug was used on an individual.

Of the ten patients who initially entered the study, only seven completed the required twenty-eight weeks, hereafter known as six-month completers. There were three people who dropped out of the study prior to their week 28 visit and were subsequently labelled as discontinuations. Patients who did not have acceptable data at the baseline (pretherapy) visit and/or the last visit (week 28) were labelled incomplete. The three patients who discontinued were also labelled incomplete as they had some data missing at the last assigned visit (week 28).

2. Evaluation Method

For a patient visit to be considered acceptable for efficacy analysis, certain pre-established ground rules have to be satisfied. Most important of these is that the patient not have excessive missed drug, not have doses consistently below the minimum allowed (100 mg for indomethacin and 1200 mg for Experimental Drug #1) nor be taking unacceptable amounts of concomitant medications. Unacceptable levels of medication were 250 mg of indomethacin daily or more, and 2000 mg of experimental Drug #1 or more daily.

All patients who, according to the above criteria, entered the study and had acceptable efficacy data at the baseline and at least one visit during therapy, were classified as efficacy patients. Efficacy assessment data were collected at the baseline visit and at each of the seven on-therapy visits. The first visit (baseline) preceded the start of treatment with the study drugs. This was followed by seven during therapy visits scheduled at 2,4,8,12,16,20 and 28 weeks after the baseline visit.

Regarding dosage, the initial dose of Experimental Drug #1 was 1200 mg daily which could be increased to 1800 mg. The initial dose of indomethacine was 150 mg daily which could be increased to 200 mg daily. Dose information was also recorded at each visit. In addition to complete information regarding the dosing regimen for the study drugs, all concomitant medications taken during the study were to be listed along with the start and stop dates for their use and the total quantity taken. Concomitant medications considered unacceptable were compounds containing aspirin, all nonsteroidal anti-inflammatory agents and steroids.

The data collected were classified into nine categories: demography, history and baseline diagnosis, efficacy, dose, patient attrition, adverse experiences, laboratory determinations, additional safety data and termination summary.

The demographic record, taken prior to the start of therapy, consists of age, sex, height, weight and duration of illness. The history and diagnosis data, providing a screening record for study eligibility, include confirmation of ankylosing spondylitis, results of physical examination and any secondary concomitant diagnosis and treatment.

The following seven variables have been considered the primary efficacy measurements: observer's and patient's assessments of the disease, day and night evaluations of sacroiliac pain intensity, chest expansion, fingertips to floor test, and erythrocyte sedimentation rate (ESR). The ESR variable has been included as a primary efficacy variable because it is well known that the ESR of patients with inflammatory diseases is high (Schulak 1982). However, it is expected that by the use of the drugs a decreasing trend will be observed. Secondary efficacy assessments were activity impairment, duration of morning stiffness, time to walk 50 feet, spinal motion flexion (anterior, left lateral and right lateral), occiput to wall test and intermalleor straddle distance.

For analysis purposes, numerical values have been associated with the measurement scales used for the subjective efficacy evaluations. The following scores were assigned to the patient's and physician's assessment of disease: 1 = asymptomatic, 2 = mild, 3 = moderate, 4 = severe, and 5 = very severe. Pain intensity has been evaluated by the patient on a scale of 0 = none, 1 = slight, 2 = moderate, 3 = severe, and 4 = extreme. If the intensity level was intermediate, the higher one of the two values was recorded. As an example, for an interval of 2-3, 3 was recorded. Analyses of the activity impairment data (a measure of how much pain interferes with activity) have been based on a scale of 1 = not at all, 2 = slightly, 3 = moderately, 4 = to a great extent, and 5 = completely.

The scores were assigned by the nurses who were measuring all the laboratory variables of the patients. If a patient dropped out of the study, a reason for discontinuation was to be provided. Attrition rate totals were compiled from this information. Adverse experiences were recorded as they occurred during therapy.

The laboratory evaluations were done at the pretreatment baseline and at weeks 2,8,16 and 28. These data were classified into three main groups: hematology, blood chemistry and urinalysis.

Also, safety data collected included pulse and sitting blood pressure, both systolic and diastolic as well as weight data.

The termination summary consists of the reasons for early termination of patients as well as the physician's evaluation of the patient's response to treatment.

In order to assess the efficacy of the two treatments, answers to the following questions should be inferred from the data:

(1) Do the patients manifest a decreasing trend of pain intensity while on therapy?

(2) Has morning stiffness of the joints decreased significantly, either, clinically or statistically?

(3) Are there significant changes in weight and other safety data; either clinically or statistically?

(4) Has the physician's final evaluation of patients shown therapeutic results?

(5) Is there any significant difference between the two drugs either clinically or statistically?

The answers to the above questions will be discussed in Section V, and are formulated into statistical terms as follows:

The null hypothesis is based on the prior assumption that one treatment is not any better or worse than the other, thus:

H_o: $\theta = 0$, versus the alternate H_a: $\theta \neq 0$ where θ = the difference within treatment groups while on therapy: that is, we let $Z_i = Y_i - X_i$ and take as our model

$$Z_i = \theta + e_i, \quad i = 1, \ldots, n$$

where n is the number of subjects (patients) and the e's are unobservable random variables. The parameter of interest is θ, the unknown "treatment" effect. The Y_i's are the values of the variables while on therapy; the X_i's are the values of the variables at baseline (pretherapy). The X_i's and Y_i's are paired.

The assumptions for this test include:

(1) The e's are mutually independent, and

(2) Each e comes from a continuous population with $E(e) = 0$.

Thus, the steps involved in testing the foregoing null hypothesis are as follows: 1) form the absolute differences $Z_1, \ldots, Z_n$. Let R_i denote the rank of Z_i in the joint ranking from least to greatest of $Z_1, \ldots, Z_n$. Step 2) is to define the indicator variables ψ_i, $i = 1, \ldots, n$, where

$$\psi_i = \begin{cases} 1 \text{ if } Z_i > 0, \\ 0 \text{ if } Z_i < 0. \end{cases}$$

Step 3) is to form the n products $R_i \psi_1, \ldots, R_n \psi_n$, and set

$$T^+ = \sum_{i=1}^{n} R_i \psi_i .$$

The product $R_i \psi_i$ is known as the positive signed rank of Z_i. It takes on the value zero if Z_i is negative and is equal to the rank of Z_i when Z_i is positive. The statistic T^+ is the sum of the positive signed ranks.

For a two-sided test of H_o versus the alternative $\theta \neq 0$, at the α level of significance, (as in our case),

reject H_o if $T^+ \geq t(\alpha_2, n)$ or $T^+ \leq \frac{n(n+1)}{2} - t(\alpha_1, n)$

accept H_o if $\frac{n(n+1)}{2} - t(\alpha_1, n) < T^+ < t(\alpha_2, n)$

where $\alpha = \alpha_1 + \alpha_2$, and the constant $t(\alpha_2, n)$ satisfies the the equation

$P_o\{T^+ \geq t(\alpha_2,n)\} = \alpha_2$. The constant $t(\alpha_2,n)$ is obtained from a table of upper tail probabilities for the null distribution of the Wilcoxon signed rank T^+ statistic (Table A.4, Hollander and Wolfe, 1973).

The procedures outlined above describe the distribution-free Wilcoxon signed rank test. When this test is performed on the HP3000 using the MINITAB program, the data are first recalled from the EDITOR and brought out onto the workspace; the next step is to subtract corresponding columns and store the results in a new column in the file. Finally, one applies the Wilcoxon test to the differences by entering the command "Wtest of MU = 0 ON DATA IN (the appropriate) COLUMNS". Thereafter the program produces the Wilcoxon statistic T^+ mentioned previously and also the p-value to test for statistical significance. This method will be illustrated in Section III2. (pg 20)

To test for statistically significant differences between treatments, a distribution-free rank sum test is utilized with the following null hypothesis:

H_o: $\Delta = 0$ versus the alternate H_a: $\Delta \neq 0$

where Δ , the parameter of interest, is the unknown shift in location due to the 'treatment'; that is, the difference between treatment groups while on therapy. We take as our model

$$x_i = e_i, \qquad i = 1, \ldots, m \qquad \text{and}$$

$$Y_j = e_{m+j} + \Delta, \qquad j = 1, \ldots, n; \; m + n = N$$

where m is the number of patients in the first drug group and n is the number in the second drug group. The X's and Y's are the values of the variables while on therapy for each group, respectively. The e's are unobservable random variables.

The assumptions for this test include:

(1) The e's are mutally independent,

(2) Each e comes from the same continuous population.

Thus, the steps involved in testing the foregoing null hypothesis are as follows: 1) order the N observations from least to greatest and let R_j denote the rank of Y_j in this ordering. Step 2) is to set

$$W = \sum_{j=1}^{n} R_j$$

The statistic W is the sum of the ranks assigned to the Y's. Finally, step 3) is to test the null hypothesis against its alternative in a two-sided test (as is our case) at the α level of significance:

reject H_o if $W \geq w(\alpha_2,m,n)$ or $W \leq [n(m+n+1) - w(\alpha_1,m,n)]$

accept H_o if $[n(m+n+1) - w(\alpha_1,m,n)] < W < w(\alpha_2,m,n)$

where $\alpha = \alpha_1 + \alpha_2$ and the constants $w(\alpha_2,m,n)$ and $w(\alpha_1,m,n)$ satisfy the equations $P_o\{W \geq w(\alpha_2,m,n\}$ or $P_o\{W \geq w(\alpha_1,m,n\}$. Values of $w(\alpha_1 m,n)$ and $w(\alpha_2,m,n)$ are given in tables of upper tail probabilities for the null distribution of Wilcoxon's rank sum W statistic. (Table A.5, Hollander and Wolfe, 1973).

The procedures outlined above describe the distribution-free Wilcoxon rank sum test. When this test is performed on the HP3000 using the MINITAB program, the data are first recalled from the EDITOR and brought out onto the file workspace; the next step is to apply the Wilcoxon rank sum test, which is also called the Mann-Whitney-Wilcoxon test in the MINITAB program as can be seen from the following output:

```
MANN-WHITNEY [ALT,=K][PERCENT CONFIDENCE=K] FOR DATA IN C,C MORE? YES
DOES A TWO-SAMPLE RANK TEST (ALSO CALLED MANN-WHITNEY-WILCOXON TEST FOR
THE DIFFERENCE BETWEEN THE 2 POPULATIONS.  IT ALSO CALCULATES THE
CORRESPONDING POINT AND CONFIDENCE INTERVAL ESTIMATES.  THE ALT. IS
GIVEN AS -1, 0, OR +1 FOR <, NOT 'EQUAL' AND >, RESPECTIVELY.
```

Thus one enters the command "MANN-WHITNEY TEST OF MU = 0 DATA IN C1, C6". Thereafter the program produces the statistic W mentioned previously and also the p-value to test for significance. It should also be noted here that the p-values are calculated using a normal approximation with continuity correction. Both the p-values for the Wilcoxon signed rank and rank sum test are computed using the symmetric normal distribution; the p-value is the largest value for which the null hypothesis is rejected. That is, for any level of significance α less than the p-value given by the MINITAB program, the null hypothesis will be rejected. In contrast, any α level larger than that given by the computer output will result in the decision to fail to reject the null hypothesis.

When the null hypothesis is true, there is a large sample approximation to both the Wilcoxon signed rank and rank sum tests. By large sample it is meant that n, the sample size, is larger than 15 for the signed rank test and that m as well as n (the sample sizes of the two populations) are larger than 10 for the rank sum test. This large sample approximation has an asymptomatic (n tending to infinity) normal distribution with mean equal to zero and variance equal to one.

For the signed rank test, the normal theory approximation to test the null hypothesis is:

reject H_o if $T^* \geq z\ (\alpha_2)$

accept H_o if $T^* < z\ (\alpha_1)$

where z represents the standard normal distribution tabled probability values for a two-sided test at the $\alpha = \alpha_1 + \alpha_2$ level of significance. As well,

$$T^* = \frac{T^+ - [n(n + 1)/4)]}{[n(n + 1)(2n + 1)/24]^{1/2}} \cdot$$

For the rank sum test, the normal theory approximation to test the null hypothesis is:

reject H_o if $W^* \geq z_{(\alpha_2)}$

accept H_o if $W^* < z_{(\alpha_1)}$

where z again represents the standard normal distribution probability values for a two-sided test at the $\alpha = \alpha_1 + \alpha_2$ level of significance. This statistic W^* has an aymptotic normal distribution as the lesser of either m or n tends to infinity for the rank sum test. The statistic W^* is calculated as follows:

$$W^{*} = \frac{W - [n(m + n + 1)/2]}{[mn(m + n + 1)/12]^{1/2}} .$$

The accuracy of the normal approximation with continuity correction for the sample sizes in this trial should be questioned: The larger the n (or sample size) the better the approximation. Since our sample sizes are both less than five, it is doubtful that the normal approximation was a good basis in calculating the p-values generated by the MINITAB program as it is recommended that at least ten observations be in each group before the large approximation be used. (Hollander and Wolfe, 1973)

In order to assess the clinical efficacy of the two drugs, nonparametric procedures were utilized due to their relaxed distributional assumptions and the small sample size available. As well, nonparametric procedures are appropriate when only comparative rather than absolute magnitudes are .available, such as in our case where patients can only be classified as better, unchanged or worse. In fact, theoretical investigations have established that the rank sum procedure has power only slightly less than that of the t test. (Remington and Schork, 1970) Thus these procedures are quite adequate.

The infrequency of this illness natrually limited the practical number of patients available for study. As the final sample size was seven; four in the indomethacin group and three in the Experimental Drug #1 group, any of several nonparametric procedures could have been used. However, the following were applied: Wilcoxon signed rank and rank sum tests, (Hollander and Wolfe, 1973) and the bootstrap (Diaconis and Efron, 1981). Several parametric procedures were also performed such as the one-way analysis of variance, Fisher's exact test and a paired and two-sample t-test as a check for the Wilcoxon tests.

A discussion of the power of some of the above tests can be found in Section IV.

III. METHODS OF STATISTICAL ANALYSIS

1. Nonparametric Procedures

All of the statistical tests have been based on two-sided alternative hypotheses since no prior assumption was made that one treatment would perform better or worse than the other, nor that improvement relative to baseline would occur rather than a worsening of patient condition. In each statistical test, a difference between means was declared significant if it indicated the probability of random occurrence of the difference was 0.05 or less.

A. Demographic Data

Descriptive statistics were computed for these data as mentioned below. Subsequent to the random allocation of patients to either the indomethacin group or the Experimental Drug #1 group at the beginning of the study, a series of statistical tests were performed to confirm the pre-therapy equivalence of these groups. Age, height and weight were variables with continuous distributions, and comparisons between therapy groups were made using the one-way analysis of variance. (Steel and Torrie, 1960 and Snedecor and Cochran, 1967) The distribution of sex was checked for eqivalence between groups by means of Fisher's exact test (Fleiss, 1981) as well as with a χ^2 (chi-squared) contingency table; and, a nonparametric procedure, the Wilcoxon rank sum test (Lehmann, 1975) was used for the duration of illness. Although the one-way analysis of variance is the same as the two-sample t-test when testing two treatment groups, the ANOVA test was performed as a check of the two-sample t-test. The two-sample t-test considers if there is any difference between group means, similar to the ANOVA test. Thus the ANOVA was performed on the age, height and weight variables. After using the bootstrap method to generate repeated artificial random samples of size 5, two-sample t-tests were performed on the variables of weight, height, age, blood pressure, resting pulse and duration of disease. However, chi-squared contingency tables were produced to check if there were any significant differences between groups for the variable of sex.

B. History and Diagnosis

Descriptive data by patient are provided. No analysis was performed.

C. Efficacy

The seven efficacy variables (primary) mentioned in the previous section were checked for pretreatment equivalence between treatment groups using a Wilcoxon rank sum test (Lehmann, 1975). Although four variables are continuous, nonparametric procedures were employed because of suspected non-normality and small sample size; and with respect to morning stiffness, due to frequent reports of stiffness lasting throughout the day. Nonparametric tests with their relaxed distributional assumptions were considered to be more appropriate for analysis of these data.

On treatment, changes from baseline in these efficacy variables were tested within either treatment group by means of the Wilcoxon signed rank test. To compare, a paired t-test was performed on this data. Generally, Friedman's test is used to test the differences from baseline as it takes into account the number of patients with no change from baseline. Friedman's test is based on the hypothesis of no treatment differences when the data consist of nk observations, where n is the number of blocks and k is the number of treatments. It is assumed that one has at least two or more treatments. Friedman's test was not used for this analysis due to its low sensitivity when only two treatments are compared, that is, paired treatments. For paired treatments, the Wilcoxon signed rank is a better alternative to detect within group differences.

In order to illustrate the bootstrap procedure, the primary efficacy variable of erythrocyte sedimentation rate would have been analysed; however, there were only two data points for most patients so that the large number of missing data were insufficient in generating a regression line. The individual patients' data for the other primary and secondary efficacy variables were not available at all so that the bootstrap procedure could not be applied. However, a visual representation of the ESR for the two groups of patients can be seen in Figures 9-10. To test for differences between treatment groups, at each visit differences from baseline were calculated. These were compared using the Mann-Whitney rank sum test (Lehmann, 1975) and compared with a two-tailed t-test.

Both within-group and between-group analyses were carried out separately for each of the seven on-therapy visits.

D. Dose

Descriptive data by patient of concomitant medication was provided. No analysis was performed on this data.

E. Patient Attrition

No analysis of this data was done since only three patients failed to complete the study.

F. Adverse Experiences

Descriptive data by patient are provided. The adverse experiences that occurred during the seven month study have been summarized and tabulated.

G. Laboratory Determinations

Laboratory data for the following variables considered to be of most clinical interest were analysed statistically: white blood cells (WBC), erythrocyte sedimentation rate (ESR), phosphorous, chloride, and pH of urine. Although all of these variables were measured on a continuous scale, non-parametric methods were used for their analysis; as well, non-normality of the population distribution from which the sample was chosen was suspected. At each visit, the Wilcoxon signed rank test with the baseline values and the final values forming the pairs was used to test for significant changes over time with in treatment groups.

At each on-therapy visit, differences from baseline were calculated. Between-group comparisons of the differences were performed by the Mann-Whitney test (Wilcoxon rank sum test). These within-group and between-group analyses were carried out separately for each of the seven on-therapy visits.

The bootstrap procedure was also applied to the above laboratory variables for reasons of comparison of results. Most of the individual patient data were available for this purpose. The regression subroutine in the MINITAB program (Ryan et al., 1982) was utilized to estimate the curve of best fit in order to obtain the missing data for the variables white blood cells (WBC), phosphorous and pH of urine for the individual patients. Specifically, the program "MTBPLRG1.HELUVA.CEB" was used in order to fit regression curves for within-patient analysis (Stitt, L., 1984).

Regression analysis uses the method of least squares in order to fit the 'best' straight line to given data. The resulting regression curve yields the expected value of y for a given x-value, and thus is useful in obtaining an average estimate of a missing dependent variable for a given independent observation.

A straight-line dependence of laboratory determinations on time (weeks) was assumed, and thus we have the model:

$$Y_i = \beta_o + \beta_1 X + \varepsilon_i, \qquad i = 1,2, \ldots, n$$

where β_o is the y-intercept of the resulting regression line, and β_1 is the slope of the curve; ε is the residual.

The assumptions for the above model include:

(1) ϵ_i is a random variable with mean zero and variance σ^2 (unknown)

(2) ϵ_i and ϵ_j are uncorrelated, $i \neq j$, so that $\mathrm{cov}(\epsilon_i,\epsilon_j)=0$.

(3) ϵ_i is a normally distributed random variable, with mean zero and the variance σ^2 by (1), that is,

$\epsilon_i \sim N(0,\sigma^2)$

H. Safety Data (Vital Signs)

Although continuous scales of measurement were used for pulse and diastolic blood pressure, these variables were not analysed using parametric methods again because of suspected non-normality and small sample size. These variables were analysed by the same non-parametric methods as the laboratory evaluations; for each of the seven on-therapy visits. The individual pulse data were unavailable thus the bootstrap procedure could not be used for vital signs analysis.

I. Termination Summary

Descriptive data by patients and a summary of the physician's evaluation of therapeutic effect are provided.

2. The Bootstrap Procedure

The name 'bootstrap', which is derived from the old saying about pulling yourself up by your own bootstraps, reflects the fact that one available sample gives rise to many others. (Diaconis and Efron, 1981) These samples are generated from the data in the original sample as follows: First, the data for each patient is copied onto another file 50 times for 50 bootstrap samples. Thereafter, samples of size five are then selected at random with replacement and the corresponding nonparametric tests are applied susbsequently. The reason for selecting only five patients is that in this study there are five patients in each drug group. On the HP3000 computer the steps of copying, mixing and selecting new data samples are all carried out by a procedure that is much faster but mathematically equivalent; the computer assigns a number to each patient and then generates the samples by matching a string of random numbers to the rows that correspond to the patients. The samples generated in this way are called bootstrap samples.

This technique is now practical as it requires the use of a computer which produces quick and inexpensive computations using Monte Carlo approximations. The advantages of the bootstrap procedure include: 1) the fact that it can be applied to any statistic and 2) it does not rely on Gaussian assumptions while facilitating statistically sophisticated computations. Inherently, this procedure can be applied to nonparametric testing and also has the benefit of being able to deal with small sample sizes and thus serves as an alternate approach to the pertinent data analysis. The bootstrap can estimate the amount of variability that would be shown by all the samples on the basis of 1 sample. The bootstrap procedure was considered to be appropriate for the data of this project for several reasons: 1) the sample size of only five patients in each treatment group was very small; it could thus serve the very necessary purpose of generating more random samples from which one could make more accurate statistical conclusions; 2) since the normality assumption was not necessarily met, the bootstrap is an excellent choice for estimating statistical variables since it can be applied to non-normal data. Thus, nonparametric tests performed on the data could also be applied after the bootstrap method had been utilized to generate more random samples of size five each, since there were five patients in each drug group. Another reason for employing the bootstrap procedure was to compare results obtained using it in combination with non-parametric tests with those obtained using non-parametric tests alone.

How does the bootstrap work? The bootstrap procedure is a method of obtaining the actual variability of a statistic from its variability over many sets of randomly generated data (Diaconis and Efron, 1981). It may be applied to any parametric or nonparametric statistic such as the correlation coefficient (a parametric statistic) or the Wilcoxon signed rank or rank sum test statistic (a non-parametric statistic) such as in this clinical trial. The advantage of this bootstrap procedure is that it can quickly give an estimate of variability using the original sample data without assuming the data are normally distributed.

It has been recommended that between 50 to 1000 bootstrap samples be generated before a reasonably accurate frequency distribution for the bootstrap samples can be determined (Diaconis and Efron, 1981). For our present case, fifty random bootstrap samples were considered an adequate number due to time limitations; it took ten minutes on the HP3000 to set up one bootstrap sample using the random number generator and subsequently using MINITAB to calculate the Wilcoxon signed rank statistic for one variable of one drug group at one time point in the trial.

The distribution of this Wilcoxon signed rank statistic (T^+) can be treated as if it were a distribution constructed from true data samples; it gives an estimate of the statistical accuracy of the value of T^+ that was calculated for the original sample. The statistical accuracy in this case is not the difference between the estimate and the true value of the Wilcoxon signed rank statistic T^+, since the true value of T^+ is not known. Rather, the statistical accuracy refers to the average magnitude of the deviation of the estimate from the true value.

The bootstrap procedure was used on the following demographic variables: weight, height, age, blood pressure, resting pulse and duration of disease. It should be noted here that the bootstrap procedure could only be performed when the individual patient data were available, as they were for the previously mentioned demographic variables. The bootstrap procedure could not be performed on any efficacy variables where the individual patients' data were unavailable. However, the bootstrap was used for the following laboratory variables: white blood cells (WBC), phosphorous, and pH of urine. It should also be mentioned here that the bootstrap procedure was performed on each treatment group separately rather than on the entire sample of patients, since the objective of this project is to test whether there are any differences between treatment groups.

Finally, the bootstrap was not utilized in the analysis of safety (vital signs) data since once again the individual patient's data were unavailable.

From the foregoing it can be seen that the nonparametric estimation of statistical error is the objective of the bootstrap procedure. By error is meant the bias and standard error of an estimator, such as the Wilcoxon signed rank statistic. The data set under consideration consists of a random sample of size ten from an unknown population distribution, say F. The sample empirical distribution puts probability mass 1/n on each x (where x is the value of the variable under consideration), and then lets X_1^*, X_2^*, $\ldots, X_n^*$ be a random sample from F, such that

$$X_1^*,\ X_2^*, \ldots, X_2^* \sim F. \qquad (1)$$

In other words, each X_1^* is drawn independently with replacement and with equal probability from the set $\{x_1, x_2, \ldots, x_n\}$.

Then $\bar{X}^* = \Sigma_i^n = 1\ X_1^*/n$ has variance

$$\text{var. } \bar{X}^* = \frac{1}{n}\, 2 \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

var. indicating variance under sampling scheme in (1) above. The bootstrap estimate of standard error for an estimator $\hat{\theta}(X_1, X_2, \ldots, X_n)$ is

$$\hat{\sigma}_B = [\text{var. } \hat{\theta}(X_1^*, X_2^*, \ldots, X_n^*)]^{1/2}.$$

This standard error is very similar to the standard error of a sample average, as it only differs by a factor of $[n(n-1)]^{1/2}$.

For our present case, the empirical distribution of F (Wilcoxon signed rank or rank sum test statistic) is known. But the reason we are using the bootstrap is to generate more random samples from which we can estimate the accuracy of the test statistic, which is the standard error, as well as compare the results of this method with simply using nonparametric procedures alone.

The approximation to $\hat{\sigma}_B$ in this case is given as follows:

$$\hat{\sigma}_B = [(\sum_{b=1}^{B} (\hat{T}^{+*} - \hat{T}^{+*\cdot})^2)/(B - 1)]^{1/2}, \quad T^{+*\cdot} = \frac{\Sigma\hat{T}^{+*}}{B} \qquad (2)$$

As B —> ∞, (1) approaches the original definition of

$$\hat{\sigma}_B = \sigma\ (\hat{F})$$

The bootstrap sampling procedure was done fifty times for the phosphorous laboratory variable at 2 weeks into the trial. Thus, the B (the number of times this was done) is equal to 50. The value of $T^{+*\cdot}$ is simply the fifty values of the bootstrap Wilcoxon signed rank statistic summed together and divided by 50, which is 255/50 = 5.1.

Thus

$$\hat{\sigma}_B = [\frac{447.991}{50 - 1}]^{1/2} = 9.1426^{1/2} = 3.02.$$

The value of 447.991 is equal to

$$\sum_{b=1}^{B} (\hat{T}^{+*} - \hat{T}^{+*\cdot})^2$$

which is the numerator in formula (2) above. When the value of 5.1 is compared to the value under the null distribution of the Wilcoxon signed rank statistic given in Selected Tables in Mathematics and Statistics (1970) which is T=5 for a sample size n of 5 in each group, the probability that this rank is equal to or less than 5 is 2(.3125) = .625. Therefore, there is not sufficient evidence at the 5 per cent level to reject the hypothesis that the two samples came from the same population.

The fifty values of T^{+*} were subsequently plotted in order to generate a frequency distribution (see Graph 8, Appendix B) of our bootstrapped sample data. The expected frequency distribution of the Wilcoxon signed rank statisticsare plotted according to the probability of their occurrence on Graph 9, Appendix B.

Also, to compare the value of the standard error σ_B (which is 3.02), we need to calculate the expected value of the standard error of our test statistic under the alternate hypothesis, that is, $u_1 \neq u_2$. The standard error of the Wilcoxon signed rank statistic is given in Lehmann (1975)(pg.128):

$$\hat{\sigma} = \left[\frac{N(N + 1) \ (2N + 1)}{24}\right]^{1/2} \qquad (3)$$

where N equals the total number of paired subjects, which is 5 patients in each drug group for this trial. This formula uses the normal approximation to the Wilcoxon signed rank test, which says that the sum T of a large number of independent random variables is approximately normally distributed. Thus σ equals $13.75^{1/2} = 3.71$.

To check the significance probability $P_H(T \leq 5.1)$ when N=5, using the continuity correction for our bootstrap sample estimate of the Wilcoxon signed rank statistic, we use the following formula given in Lehmann, 1970:

$$P\left[\frac{T - E(T)}{\sqrt{\text{Var.}(T)}} \leq a\right] \simeq \Phi\,(a)$$

where T is the value of the calculated Wilcoxon test statistic, E(T) is the expectation of T and Var(t) is the variance of T. The expectation of T is given by the following formula:

$$E(T) = \frac{N(N + 1)}{4}$$

and the variance is given by (3). Thus E(T) = 7.5 and

$$P_H\ (T \leq 5.1) \simeq \phi\left[\frac{5.1 - 7.5}{3.71}\right] = \phi\ (-0.65) = .2578 \times 2 = 0.5156$$

which agrees quite nicely with the value calculated earlier; namely, 0.6250.

To illustrate this method using one of the variables in this study, let us take phosphorous, one of the laboratory variables for which the individual patient data are available. After the bootstrap procedure had been applied in order to generate fifty phosphorous samples to test, the following file was produced after the MINITAB program had been used to recall the data which are for the Experimental Drug #1:

```
                  MTB > READ 'PHOSB002' C1-C10
                        5 Rows Read

                     Original Sample

     ROW     C1      C2      C3      C4      C5
      1      2.6     2.2     2.7     2.9     2.7
      2      3.0     2.5     2.9     3.0     3.1
      3      3.4     3.4     3.2     3.2     2.9
      4      2.3     2.9     2.6     3.2     2.4
             ...


                Bootstrap Selected Sample

     ROW     C6      C7      C8      C9      C10

      1      3.0     2.5     2.9     3.0     3.1
      2      2.6     2.2     2.7     2.9     2.7
      3      2.6     2.2     2.7     2.9     2.7
      4      2.3     2.9     2.6     3.2     2.4
             ...
```

As one can see by observing the numbers in columns C6 and C1, the numbers in column C6 often are simply the same numbers as in column C1. This is because the random number generator in the BASIC program chose row 2 (columns C1 to C5, patient 2) to be the first patients' data in the new bootstrap sample. Thus, row 2 (column C1 to C5) can be found in exact replicate as the first row in columns C6 to C10, which makes up the first patient's data for this sample obtained using the bootstrap procedure. Continuing in this manner, we see that rows 2 and 3 in columns C6 to C10 are exactly the same; this is due to the fact that the random number generator in the BASIC program (using RANDOM.BASIC.LIB) chose the first patient's data twice and thus it became the second and third rows in this bootstrap sample. Similarly, row 4 (columns C6 to C10) happens to be by chance the same patient (number 4) chosen by RANDOM.BASIC.LIB as was in the original sample. Therefore, in row 4 columns C1 to C5 have the same numbers in sequential order as columns C6 to C10. The following computer output illustrates how the program RANDOM.BASIC.LIB was used to generate random numbers, allowing for repetition of patients' data in rows.

```
RUN
RANDOM
DO YOU NEED RANDOM NUMBERS WHICH DO NOT REPEAT OR CAN THE SAME NUMBER
BE USED MORE THAN ONCE?  0 = NO REPEATS 1 = CAN REPEAT?1
WHAT IS THE SAMPLE DESIRED AND POPULATION SIZE??5,5

SAMPLE NO.          RANDOM NO.

   1                   2
   2                   1
   3                   1
   4                   4
   5                   3
```

From the above one can observe how the patients were chosen for the bootstrap sample shown in columns C6 to C10 (rows 1 to 4) on the previous page. The next step, when applying a nonparametric procedure such as the Wilcoxon signed rank test, is to subtract corresponding columns. By this it is meant that phosphorous levels at corresponding time points in the trial be subtracted, one from the other, and the resulting number stored in a new column. The commands necessary for this step are seen in the following MINITAB output:

```
            MTB > SUBT C6 FRM Cl,Cll

            MTB > SUBT C7 FRM C6,C12

            MTB > SUBT C8 FRM C6,C13

            MTB > SUBT C9 FRM C6,C14

            MTB > SUBT C10 FRM C6,C15
```

Columns Cll to C15 have the resulting differences. Next, one asks the program to execute the Wilcoxon test on the data in columns Cll through to C15, as can be seen in the following output:

```
     MTB > WTEST OF MU=0 ON DATA IN Cll,C12,C13,C14,C15

  TEST OF CENTER = 0 VERSUS CENTER N.E. 0

                     N FOR       WILCOXON                   ESTIMATED
                  N   TEST       STATISTIC     P-VALUE       CENTER

   Cll            5    4           5.5          1.000               0
   C12            5    4           6.0          0.855          0.2000
   C13            5    5           7.0          1.000        -0.05000
   C14            5    4           1.0          0.201         -0.3000
   C15            5    5           5.0          0.590         -0.1000
     MTB > STOP
```

As can be clearly seen from the above output, all the p-values are greater than 0.05 and thus the Wilcoxon signed rank test is not statistically significant for any of the ranked differences present in columns Cll to C15.

It should be noted here that the data in corresponding columns were the phosphorous levels recorded at weeks 0,2,8,16 and 28. Thus the total number of time points for which laboratory variables were recorded was five; and so including the five time points for the bootstrap sample one obtains the necessary ten columns found previously.

Thus, in the manner described above, ten similar bootstrap samples were chosen using the random number of generator for each of the laboratory variables of pulse, white blood cells and pH of urine. As well, ten bootstrap samples were chosen for each of the following demographic variables: weight, height, age, blood pressure, resting pulse and duration of disease.

Only ten bootstrap samples were performed for each of the above variables due to the reasons mentioned earlier on page 23. The ten bootstrapped samples were then analysed using several nonparametric and parametric procedures to see whether there were any differences between groups between the original sample and the bootstrapped samples.

IV. RESULTS OF ANALYSIS

A. Demographic Data

The demographic data by patient are presented in Table 2, and a summary with the results of the statistical analysis is presented in Table 3. No significant differences (p>0.07) in the demographic characteristics were found when traditional nonparametric procedures were conducted (see Table 3).

All but two patients were male with an overall average age of 37.2 (interval 20-57), average weight 57.54 kilograms (interval 38-76), average height 161.55 centimetres (interval 151-178), and average duration of illness 11.6 years (interval 1-20). The resting pulse rates were normal for all patients varying from 62 to 108 beats per minute; however, there was a statistically significant difference in pulse between the two drug groups (p=0.0216) at baseline. Blood pressure was abnormal (>140/90 mm Hg) for a number of patients (#'s 2,5,8). Two patients (#'s 2,8) had systolic blood pressure greater than 140 mm Hg and one patient (#5) had diastolic blood pressure of 105 mm Hg. This same patient (#5) was being treated for hypertension and had normal blood pressure at baseline. All of the above results are shown in Tables 2 and 3.

For reasons of comparison the pretherapy equivalence of the patients was again tested by using the bootstrap procedure on the two groups. Ten bootstrap samples were chosen in the following manner: first, the BASIC program on the HP3000 was used to generate many random samples each of size five with allowable repetitions. Next, the selected patients' data were copied onto the original sample data file and differences between similar columns were computed using the MINITAB program. (Ryan et al., 1982) Then the Wilcoxon signed rank test was used to test for any significant differences within the drug groups for the variable duration of illness. One-way analysis of variance tables were computed for the continous variables age, height and weight. When the above steps were completed for both drug groups, the two newly selected bootstrap samples (one for each drug group) were put together into one file and, using the MINITAB program, the Mann-Whitney (Wilcoxon rank sum test) test was performed between similar columns to check for any differences between groups.

The results of analysis using the bootstrap procedure are as follows: no significant changes within groups for the variable duration of illness were found when the Wilcoxon signed rank test was applied. When the Mann-Whitney (Wilcoxon rank sum test) was performed, no statistically significant differences between groups were found after the bootstrap had been applied ten times. These results are not tabulated due to the lack of statistical significance.

When the one-way analysis of variance (ANOVA) was computed, no differences between age, weight or height were found for the two drug groups. The paired t-test was also done as a check against the results obtained using nonparametric tests. The same results were obtained for the variables of weight, height, age, systolic and diastolic blood pressure and duration of illness; namely, no significant differences between groups. However, a statistically significant difference between groups was found for sitting pulse; the indomethacin patients demonstrated higher values than the experimental patients pretherapy. The calculated p-value was $p = 0.019$.

The chi-squared (χ^2) contingency table was constructed for the sex variable; however, the χ^2 random variable could not be computed due to two cells with expected frequency less than one and four cells with expected frequency less than five. No conclusion was made. The Fisher's exact test showed no differences between groups for sex. The tabulated values of the tests mentioned above can be seen in Tables 19 and 20. The F values for the ANOVA tables are one-tailed, while the χ^2 test is an upper one-tailed test.

The reason for both the F values and the χ^2 values being upper one-sided is that we expect the two drug groups to be different from each other; the more different they are from each other, the larger the corresponding values of χ^2 and F become. Thus the chi-square frequency tests and the F tests are inherently upper one-sided tests under the alternate hypothesis.

B. History and Diagnosis Data

A detailed description of the history and previous treatment of the present diagnosis as well as any secondary concomitant diagnosis and therapy is presented in Table 4. All patients had been previously treated for the present diagnosis.

Three patients had a secondary concomitant diagnosis including convulsions, high blood pressure and hypertension.

C. Efficacy

The results of the analysis within and between therapy groups for all six primary efficacy variables are presented in Table 5. A visual presentation of the first five variables are presented in Figures 1-10 (Appendix A) showing trends over time.

Analyses based on the 5 Experimental Drug and 5 indomethacin patients showed some statistically significant changes, indicative of improvement, among both groups of patients. Among the Experimental Drug group, significant changes were found in observer's opinion at 8 weeks and patient's opinion at 8 weeks also. The indomethacin group showed no significant changes from baseline for any of the primary efficacy variables, using the Wilcoxon signed rank test.

Comparisons between the two treatment groups resulted in a significant difference detectable for day pain intensity at 2 weeks, using the Mann-Whitney (Wilcoxon rank sum test).

The erythrocyte sedimentation rate was previously mentioned as a measure of the efficacy of the two drugs; however, due to the number of missing data points, not even a regression line could be estimated so that the bootstrap procedure was not performed for this or any other primary or secondary efficacy variable. All of the secondary efficacy variables were analyzed using the Wilcoxon signed rank and sum tests with the following results: Intermalleor straddle distance for the indomethacin group was statistically significant at 4,8,12,20 and 28 weeks from baseline. Also for the Experimental Drug group, significant changes from baseline occurred at 2,4,8, and 12 weeks. Within this same group there was a change from baseline at 12 and 16 weeks for the variable lateral right spinal motion flexion. Finally, for the Indocid group there was a significant change from baseline at 4 weeks for anterior spinal motion flexion. These results are shown in Tables 6-7.

Comparisons between the two treatment groups resulted in no significant differences detectable. All of the secondary efficacy variables were analysed using the Wilcoxon signed rank and rank sum tests. Visual presentations of activity impairment and intermalleor straddle distance are shown in Figures 11-14 (Appendix A).

D. Dose

Those patients taking concomitant medication during the study are listed in Table 8 along with the start date and duration for the use of each concomitant medication. Only two patients, both males and in the Experimental Drug group, were on any secondary drugs.

E. Patient Attrition

Two Experimental Drug patients (both males) discontinued from the study (Table 9). One left due to unsatisfactory response to treatment as well as requiring additional medication. The other one dropped out due to severe pain in the neck. Only one indomethacin patient dropped out because of severe headaches.

F. Adverse Experiences

Table 10 provides the adverse experiences reported by patient. Five patients reported adverse experiences. The most common adverse reaction was that of abdominal cramps, which was not attributed to treatment while on trial. A summary of adverse reactions can be seen in Table 11.

G. Laboratory Determinations

The analyses of the laboratory data are presented in Tables 13-15. These within-group and between-group analyses show scattered instances of statistical significance, but significant differences were not found consistently across time intervals for most variables. Normal lab values are found in Table 12.

There appears to be no definite trend at baseline for most of the hematology variables; only two of them are illustrated in Table 13. There is one significant change in WBC from baseline within the indomethacin group at 2 weeks. There is also one significant difference between the two drug groups at 2 weeks for the variable White Blood Cells (WBC).

For the blood chemistry data, there are no trends in the baseline data as there should be; the values remain very consistent. There was a significant change within the indomethacin group for the phosphorous variable at weeks 2 and 28. For the chloride variable there was a detectable change from baseline at 2 weeks for the Experimental Drug patients as well as a similar change within the indomethacin group at 16 weeks. There appears to be very little difference between the experimental patients and the indomethacin patients; these results are in Table 14.

For the urinalysis variables, there appears to be a very slight decreasing trend in the indomethacin group from baseline for pH, as well as for the experimental patients. There were no detectable changes within therapy groups nor any significant differences between groups, as can be seen in Table 15, for pH.

All of the above results were obtained using the traditional nonparametric procedures mentioned previously, namely the Wilcoxon signed rank and rank sum tests. However, white blood cells, phosphorous and pH were also analysed using the bootstrap procedure. Before doing this, regression methods were used to estimate the missing data for individual patients. The resulting graphs for the laboratory variables mentioned previously can be found in Appendix B.

When the bootstrap procedure was applied to the individual patients' white blood cell (WBC) data, no significant changes were found when the Wilxocon signed rank test was performed within groups. As well, no significant differences were obtained when the Mann-Whitney (Wilcoxon rank sum) test was applied between groups. The two-sample t-test resulted in the same conclusion.

For the phosphorous variable, barely (i.e. p slightly larger than 0.05) statistically significant changes were observed within the indomethacin group at 8,16 and 28 weeks ($p = 0.059$ in all cases) when the Wilcoxon signed rank test was applied after many random samples had been generated using the bootstrap. No significant changes were found within the experimental group; however, the Mann-Whitney (Wilcoxon rank sum) test was able to detect differences between groups at 2 weeks (p = 0.037), 8 weeks (p = 0.12) and 28 weeks (p = 0.012). The two-sample t-test picked up a significant change within the indomethacin group at 16 weeks (p = 0.0004).

When pH was analysed using the bootstrap procedure, no significant changes were detected within either group using the Wilcoxon signed rank test. The Mann-Whitney test also showed no differences between groups, and the two-sample t-test agreed with this conclusion.

All of the above results may be seen in Table 18.

H. Safety Data (Vital Signs)

Analysis of the vital sign variables of diastolic blood pressure and pulse are presented in Table 16. Baseline comparisons indicated no significant difference between therapy groups with respect to these variables. Diastolic blood pressure seems to illustrate a decreasing trend at baseline over time for the Experimental Drug patients. However, the pulse variable shows an opposite effect of slight increase over time for the experimental patients. There does not seem to be a definite trend within therapy groups.

Analysis of pulse data resulted in significant changes within the indomethacin group at 4 and 16 weeks. As well, there was a significant difference between the two therapy groups at baseline (pretrial). Graphs illustrating trends over time can be found in Figures 15-18 (Appendix A).

I. Termination Summary

Descriptive data by patient and the physician's evaluation of therapeutic effect are provided in Table 17. Evaluation of both indomethacin and Experimental Drug #1 was satisfactory for all patients who did not drop out of the trial. Only those patients who discontinued had unsatisfactory evaluations. None of the patients were worse at termination.

J. Power of Statistical Tests

The power of the statistical tests conducted in this study is discussed as follows as indicated in Section II, Part 2.

The definition of power of a test is the probability of rejecting the null hypothesis when the alternate hypothesis is true. This is the probability of correctly rejecting the null hypothesis and thus it is desirable that this probability be as large as possible. The power against specified alternative is equal to the quantity $1 - \beta$, where β is the probability of making a type II error. A type II error means failing to reject the null hypothesis when in fact the alternate is true. If the power, $1 - \beta$, is large, then β, the type II error will be small, which is highly desirable.

As the sample size increases to infinity, the alternative hypothesis test will have the power tending to 1, where 1 is the highest value power can have. In this trial, the final sample size of seven was often not large enough to make an adequate distinction between the hypothesis and its alternative. This is especially true when normality is assumed for the parametric tests.

As well, we would like α to be as small as possible; in this study α was chosen before sampling began to be 0.05. In order to reduce β for a fixed α (0.05 in this case), the sample size must be increased. This in turn increases the power so that it is easier to detect bias.

In summary, for a fixed level of significance α, as the alternative hypothesis deviates by a greater amount from the null hypothesis, the power increases and the type II error probability decreases as desired.

It is difficult to state exactly what sample sizes should be used in future studies of these two drugs. Firstly, the population variances are not known. Secondly, the sample sizes of the present study groups were too small to obtain good estimates of the sample variances. Further, the acceptable levels of the laboratory variables were given as intervals rather than specific values; thus, population means could not be established for statistical testing (see Table 12).

However, it is recommended that sample sizes of at least thirty patients be employed in future studies before parametric tests such as the t-test be performed (Remington and Schork, 1970). For a visual presentation of the differences in power when two different sample sizes are compared, see Figure 20. This figure assumes the population is normally distributed, which should be the case when a future study is done on the two drugs if there are at least 15 patients in each group. The shaded areas under the distributions correspond to the power (Remington and Schork, 1970).

The power of the nonparametric tests used in this project is very difficult to calculate as it involves the summation of all the possible permutations of rankings of the patients' data, which is beyond the scope of this project. This could be simulated on the computer; however, this was not the intention of this discourse.

If, however, normality is assumed and the sample variances were good estimates of the population variances, one could calculate the power of various tests performed after first computing the necessary sample sizes needed.

For the two-sample t-test when the population variances are not assumed equal, the power of this test for our data of five patients in each drug group can be derived in the following manner (Cohen, 1969):

The first step is to decide on the degree of departure from the null hypothesis we wish to detect. This is known as the effect size, hereafter symbolized by the letter d. For a two-sample case such as is present here, the effect size is:

$$d = \frac{|m_1 - m_2|}{\sigma'}$$

where $\sigma' = \sqrt{\dfrac{\sigma_1^2 + \sigma_2^2}{2}}$; that is, σ' is the root mean square of σ_1 and σ_2. The mean of the Experimental Drug group variable day pain, for example, at pretherapy (see Table 5) is denoted by m_1 and the mean of day pain for the indomethacin Drug group at pretherapy is denoted by m_2. Thus $m_1 = 0.80$ and $m_2 = 2.00$ in this case. The Experimental Drug group's variance is represented by σ_1^2 and the indomethacin drug group's variance is represented by σ_2^2, repsectively. These variances can be easily calculated by the formula:

$$S.E._{\bar{x}} = \sqrt{\frac{s^2}{n}}$$

where $S.E._{\bar{x}}$ is the standard error of the mean of the variable in question, namely day pain. The sample size is denoted by n, and s^2 is the unbiased estimate of the population variance of x, day pain. So in this case the standard errorsof the two drug groups (see Table 5) are 0.37 and 0.63 respectively, while the sample size is 5 for each group.

From the above formula it can be seen that $\hat{\sigma}_1 = 0.83 = s_1$ and $\hat{\sigma}_2 = 1.41 = s_2$. Thus $\sigma' = \sqrt{\frac{0.68 + 1.99}{2}} = 1.15$. From this we see that $d = \frac{|0.80 - 2.00|}{1.15} = 1.04$. For our level of significance $\alpha = 0.05$, from the table on page 53 of Cohen (1969), we see that the power of our t-test for five patients in each group was 0.25. This also assumes a two-tailed test, since we did not assume that one drug would be better than the other. Graph 7 in Appendix B portrays power versus sample size for two different effect sizes d.

Thus, for future studies, if we decided that we wanted to be able to detect an effect size d of 1.04 (or 1.00 for the purposes of using the tables), and one also assumed a two-tailed t-test with a level of significance equal to 0.05, the sample size required to yield power of 0.80 (highly desirable) would be 17 patients in each group (see Cohen 1969; pg.53), thus a total of 34 patients would be required for the day pain variable.

We can see from the above that the power of 0.25 taken from the tables for our available sample size of only five patients in each treatment group is quite undesirable; this means that one only has one chance in four of rejecting a null hypothesis. The probability of making a Type II error (β) would be quite high, since β would equal (1-power) = 1-0.25 = 0.75.

For future studies, it is desirable that the power be quite high and the probability of making a type II error be low; increasing the sample size accomplishes both of these goals. In fact, as an example, if a power of 0.90 were desired, then the β would equal 1-power = 1-0.90 = 0.10. If one still wanted to be able to detect the same effect size d of 1.00, 22 patients would be required in each drug group (Cohen, 1969).

Finally, if one wished to reduce the amount of departure d from the null hypothesis, to say half the original amount, that is 0.50, and kept the power at 0.90, β at 0.10, the level of significance α at 0.05;

the necessary sample size would be greatly increased to 85 patients in each group. This also assumes a two-tailed test. It might be very difficult to obtain so many patients with this particular disease in Canada.

V. DISCUSSION AND CONCLUSION

The following discussion addresses the questions raised in Section II, Part 2.

A statistically significant difference in day pain was found at baseline (pretherapy) between the two drug groups. A general decreasing trend in day pain was found for both drug groups while on therapy, although not statistically significant. The variable of night pain demonstrated a decreasing trend for the Experimental Drug patients during therapy. However, in the indomethacin group fluctuations in pain level occurred. Thus, the decreasing trends in pain relief for both day and night pain make the new drug appear therapeutically superior.

For the variable of morning stiffness, fluctuations occurred for both groups indicating that neither drug is effective in alleviating this problem.

With regards to the safety (vital signs) data, no significant changes in weight were detected for either drug group. Also, blood pressure seemed to remain relatively constant while on therapy for both groups. However, for sitting pulse, statistically significant differences were detected between drug groups at 2 weeks, and within the indomethacin group at 4 and 16 weeks. Thus, the Experimental Drug appears to be superior in keeping pulse normal and lower than the indomethacin.

In the physician's opinion, the severity of the experimental patients' disease seemed to decrease clinically, with a statistically significant change occurring after 8 weeks on trial. However, for the indomethacin group only fluctuations were observed with no statistically significant changes.

In conclusion, the physician's evaluation indicates that the new drug is therapeutically better regarding day and night pain relief and pulse clinically, but not statistically.

Having discussed power limitations of tests performed, we shall now focus attention on the comparison of results between nonparametric tests and nonparametric tests applied after the bootstrap was performed. A summary of demographic and laboratory variables analysed is shown in Table 18. The differences in results obtained may be attributed to the following reasons:

(1) The estimated regression curves were obtained by using the available data which sometimes included only four (4) time points (x-values). This may be an insufficient number of points in order to obtain a good 'fit' of the data due to the low number of degrees of freedom used in order to estimate σ^2. There would only be n-2, equal 2 degrees of freedom that would be used in estimating σ^2, which is very small.

(2) The bootstrap procedure can only be used if individual patient data are available. This may not always be practical as in this study where individual efficacy and

safety (vital signs) data were unattainable for analysis. Thus, grouped data cannot always be analysed using the bootstrap method. As well, the generation of many artificial random samples requires a large enough computer system that can handle massive calculations quickly and inexpensively.

(3) Further, the bootstrap does not always guarantee a true picture of the statistical accuracy of a sample estimate. This limitation is not so much a failure of the procedure as it is a restatement of the conditions of uncertainty under which all statistical analyses must proceed.

(4) The two-sample t-test works best when the assumption of population normality is met; however in this study non-normality was suspected, thus this assumption may be violated.

(5) The one-way analysis of variance (ANOVA) also assumes two normally distributed populations; this assumption is again violated.

The following addresses the usefulness of the bootstrap procedure for this data and whether its application is appropriate.

The first observation to be made is that the bootstrap procedure was only applied when individual patient data were available; thus it could not be applied to efficacy data, either primary or secondary. It would have been desirable to conduct the bootstrap method on the efficacy data if they were available so that comparisons and contrasts could have been made of the results obtained. This would have also provided more insight as to whether one drug is more, or less, efficacious than the other or whether both were similar.

The bootstrap was a useful tool in generating more random samples from which both nonparametric and parmetric test statistics were calculated. In this way one could obtain a general idea of the statistical accuracy of a test statistic from the frequency distribution of the samples so generated.

Another advantage of the bootstrap procedure is that it makes no distributional assumptions which means that it can be applied to any statistic. Also, the small sample size as well as the suspicion that the data are not from a normal distribution indicate that the bootstrap was an excellent choice for estimating the actual variability of test statistics caluculated.

The following points may be recommended for similar studies in the future:

(1) Counteract the dropout rate (40% for the experimental group and 20% for the Indocid) by telling future patients that both drugs are proven pain relievers so that they are more inclined to stay on therapy.

(2) Continue to use both drugs because there is evidence indicating that both relieve pain.

(3) Further, if there are few volunteers due to the

infrequency of the disease, combine the data in this study with that of a new study if patients are not significantly different at baseline (pretherapy).

(4) Conduct a multicentre double-blind trial to obtain a larger population from which more patients may be sampled for analysis. If necessary, combine present data with those of other provinces or countries such as the U.S.

(5) It is advisable to conduct an investigation to discover why so many undesirable side effects occur for patients while on trial. These should then be monitored and the dosing regimen altered in order to reduce the prevalence of adverse reactions. In this way patients will be encouraged to stay on therapy. Also, nurses or administrative personnel should continually remind patients of their next scheduled appointment. Perhaps the measurement scales for subjective efficacy evaluation should be simplified (see Section II, Part 2).

(6) Conduct another double-blind trial to test the efficacy of the Experimental Drug versus a different clinically accepted drug.

(7) Develop a computer program which greatly speeds up the time necessary to generate the random samples needed using the bootstrap method. Lack of sufficient computer programming knowledge as well as the time necessary to

write such a program prevented this step from being carried out.

(8) Finally, after this program is produced, apply the bootstrap procedure from 500 to 1000 times in order to generate the necessary random samples from which accurate test statistics are calculated and frequency distributions are generated. This can be done in any clinical trial, regardless if small or large sample is available.

TABLES

TABLE 1: SUMMARY OF THE FINAL STATUS OF PATIENTS

	SAMPLE	NUMBER	NUMBER WITH ACCEPTABLE
THERAPY	SIZE	DISCONTINUED	DATA IN COMPLETION INTERVAL
EXPT.* DRUG #1	5	2	3
INDOMETHACIN	5	1	4
TOTAL	10	3	7

* - EXPT is the abbreviation for Experimental

TABLE 2: Demographic Record

PT #	STUDY DRUG	ILLNESS DURATION (YRS)	SEX	AGE (YRS)	HEIGHT (cm)	WEIGHT (kg)	B.P. (systolic/ diastolic) (mm Hg)	PULSE (beats/ min.)
1	Indomethacin	11	F	39	151	38	130/85	108
2	Expt. Drug #1	20	M	57	157	65.6	160/90	78
3	Indomethacin	15	M	44	153	52	105/80	102
4	Expt. Drug #1	5	F	27	162	59.7	110/70	76
5	Indomethacin	20	M	38	159	63.1	140/105	78
6	Expt. Drug #1	4	M	41	164	53	120/80	70
7	Expt. Drug #1	1	M	20	170.5	54	110/70	80
8	Indomethacin	15	M	31	154	52.5	145/85	72
9	Expt. Drug #1	20	M	50	178	76	140/80	62
10	Indomethacin	5	M	25	167	61.5	120/70	100

TABLE 3: SUMMARY OF DEMOGRAPHIC ATTRIBUTES

ATTRIBUTE	EXPT. DRUG #1	INDOMETHACIN	p-VALUE
TOTAL NUMBER OF PATIENTS	5	5	
SEX			
MALE	4(80%)	4(80%)	1
FEMALE	1(20%)	1(20%)	
AGE (YEARS)			
MEAN	39.0	35.4	0.6512
(MIN.-MAX.)	20-57	25-44	
WEIGHT (kg)			
MEAN	61.66	53.42	0.2176
(MIN.-MAX.)	53-76	38-63.1	
HEIGHT (cms)			
MEAN	166.30	156.80	0.0745
(MIN.-MAX.)	157-178	151-167	
DURATION OF ILLNESS (YEARS)	10 1-20	13.2 5-20	0.5264

HISTORY AND DIAGNOSIS

TABLE 4: Previous Therapy of Ankylosing Spondylitis

							2nd Concomitant Diag.	
PT#	DRUGS	FREQUENCY OF ADMINISTRATION	SINGLE TABLET STRENGTH	START DATE	STOP DATE	RESPONSE	ILLNESS	MEDICATION
1	Indocid	T.I.D.	25 mg	1974	1976	good	none	none
	Naprosyn	B.I.D.	250 mg	1976	1978	good		
	Alka-Butazolidin	T.I.D.	100 mg	Feb/78	Ap/79	good		
	Motrin	Q.I.D.	400 mg	Ap/79	Jul/79	poor		
	Clinoril	B.I.D.	200 mg	Jul/79	Ap/81	good		
2	Butazolidin	T.I.D.	100 mg	1960	1977	good	none	none
	Indocid	Q.I.D.	25 mg	1977	Jun/79	good		
	Gold Salts	once/week	50 mg	Nov/78	Feb/79	good		
	Clinoril	B.I.D.	200 mg	Jun/79	Sep/81	good		
3	Butazolidin	T.I.D.	100 mg	1965	1970	good	convulsions	Dilantin Phenobarbitol
	Indocid(supp.)	h.s.	100 mg	1970		good		
	Naprosyn	3 cap/B.I.D.	125 mg	1975	1979	poor		
	Indocid	T.I.D	25 mg	1979	1979	good		
	Clinoril	B.I.D.	200 mg	Nov/79	Dec/80	poor		
	Orudis(supp.)	h.s.	100 mg	Aug/80	Dec/80	poor		
	Nalfon	Q.I.D.	600 mg	Dec/80	Sep/81	poor		
4	Indocid	T.I.D.	25 mg	1976		good	none	none
	Entrophen	Q.I.D.	10 gm	1978	1979	poor		
	Indocid(supp.)	h.s.	100 mg	Dec/80	Sep/81	good		
5	Alka-Butazolidin	Q.I.D.	100 mg	1977	Sep/81	good	High Blood Pressure	Dyazide
6	Entrophen	4 - 6	10 gm	1958	1978	none	none	none
	Naprosyn	3 B.I.D.	125 mg	Sep/78	Oct/78	poor		
	Indocid(supp.)	h.s.	100 mg	Sep/78	Oct/78	poor		
	Valteren	2 B.I.D.	50 mg	Aug/81	Aug/81	good		
	Indocid	T.I.D.	25 mg	Jul/78	Aug/78	poor		
7	Naprosyn	2 B.I.D.	125 mg	Sep/81	Sep/81	good	none	none
8	Butazolidin	T.I.D.	100 mg	1964		poor	none	none
	Naprosyn	3 B.I.D.	125 mg	Jun/81		none		
	Clinoril	B.I.D	200 mg	May/81		none		
9	ASA	T.I.D.	5 gm	Feb/79	Aug/81		Hypertension	Hydrodiuril
10	Naprosyn	2-3 cap/B.I.D.	125 mg		Aug/81	good	none	none

TABLE 5: Results of Analysis Within and Between Treatment Groups for Primary Efficacy Data

		Experimental Drug #1				Indomethacin			
EFFICACY PARAMETER	WEEK	SAMPLE SIZE	PRETHERAPY MEAN(S.E.)	ON THERAPY MEAN(S.E.)	MEAN CHANGE(S.E.)	SAMPLE SIZE	PRETHERAPY MEAN(S.E.)	ON THERAPY MEAN(S.E.)	MEAN CHANGE(S.E.)
Observer's Opinion	2	5	3.40(0.51)	2.80(0.37)	-0.60(0.51)	5	3.60(0.24)	2.80(0.37)	-0.80(0.20)
	4	5	3.40(0.51)	2.60(0.51)	-0.80(0.24)	4	3.50(0.29)	2.75(0.25)	-0.75(0.25)
	8	5	3.40(0.51)	2.40(0.51)	-1.00(0.20)*	4	3.50(0.29)	2.50(0.50)	-1.00(0.41)
	12	5	3.40(0.51)	2.60(0.68)	-0.80(0.00)	4	3.50(0.29)	2.50(0.50)	-1.00(0.41)
	16	5	3.40(0.51)	2.40(0.75)	-1.00(0.49)	3	3.67(0.33)	3.33(0.33)	-0.33(0.33)
	20	3	3.00(0.58)	2.00(0.57)	-1.00(0.84)	4	3.50(0.29)	2.75(0.63)	-0.75(0.48)
	28	3	3.00(0.58)	2.00(0.57)	-1.00(0.00)	3	3.33(0.33)	2.33(0.67)	-1.00(0.58)
Patient's Opinion	2	5	3.40(0.51)	2.80(0.37)	-0.60(0.24)	5	3.60(0.24)	2.80(0.37)	-0.80(0.20)
	4	5	3.40(0.51)	2.60(0.51)	-0.80(0.20)	4	3.50(0.29)	2.75(0.25)	-0.75(0.25)
	8	5	3.40(0.51)	2.40(0.51)	-1.00(0.00)*	4	3.50(0.29)	2.00(0.41)	-1.50(0.29)
	12	5	3.40(0.51)	2.60(0.68)	-0.80(0.49)	4	3.50(0.29)	2.25(0.48)	-1.25(0.25)
	16	5	3.40(0.51)	2.40(0.75)	-1.00(0.84)	3	3.67(0.33)	3.00(0.58)	-0.67(0.33)
	20	3	3.00(0.58)	2.00(0.58)	-1.00(0.00)	4	3.50(0.29)	2.50(0.65)	-1.00(0.41)
	28	3	3.00(0.58)	2.00(0.58)	-1.00(0.00)	3	3.33(0.33)	2.00(0.58)	-1.33(0.33)
Day Pain	2	5	1.40(0.25)	0.80(0.37)	-0.60(0.24)	5	2.40(0.25)#	2.00(0.63)	-0.40(0.40)
	4	5	1.40(0.25)	0.60(0.40)	-0.80(0.20)	4	2.25(0.25)	1.50(0.87)	-0.75(0.63)
	8	5	1.40(0.25)	0.60(0.40)	-0.80(0.37)	4	2.25(0.25)	0.25(0.25)	-2.00(0.00)
	12	5	1.40(0.25)	0.40(0.25)	-1.00(0.32)	4	2.25(0.25)	0.25(0.25)	-2.00(0.00)
	16	5	1.40(0.25)	0.40(0.25)	-1.00(0.32)	3	2.33(0.33)	0.67(0.67)	-1.67(0.33)
	20	3	1.67(0.33)	0.33(0.33)	-1.33(0.33)	4	2.25(0.25)	0.50(0.50)	-1.75(0.25)
	28	3	1.67(0.33)	0.67(0.67)	-1.00(0.58)	4	2.25(0.25)	1.00(0.71)	-1.25(0.48)
Night Pain	2	5	1.80(0.58)	1.40(0.40)	-0.40(0.24)	5	2.00(0.55)	0.80(0.37)	-1.20(0.37)
	4	5	1.80(0.58)	0.60(0.40)	-1.20(0.58)	4	2.50(0.29)	1.00(0.41)	-1.50(0.29)
	8	5	1.80(0.58)	0.60(0.40)	-1.20(0.58)	4	2.50(0.29)	0.50(0.30)	-2.00(0.00)
	12	5	1.80(0.58)	0.60(0.40)	-1.20(0.58)	4	2.50(0.29)	0.50(0.30)	-2.00(0.00)
	16	5	1.80(0.58)	0.40(0.24)	-1.40(0.51)	3	2.67(0.33)	1.33(0.90)	-1.33(0.67;
	20	3	2.00(1.00)	0.33(0.33)	-1.67(0.88)	4	2.50(0.29)	1.00(0.71)	-1.50(0.50)
	28	3	2.00(1.00)	0.67(0.33)	-1.33(0.67)	4	2.50(0.29)	1.00(0.71)	-1.50(0.50)
Chest Expansion	2	5	3.20(0.58)	3.40(0.60)	0.20(0.37)	5	2.40(0.75)	2.70(0.60)	0.30(0.41)
	4	5	3.20(0.58)	2.90(0.48)	-0.30(0.25)	4	1.75(0.48)	2.25(0.72)	0.50(0.35)
	8	5	3.20(0.58)	3.70(0.80)	0.50(0.39)	4	1.75(0.48)	2.63(0.69)	0.87(0.31)
(cm)	12	5	3.20(0.58)	3.40(0.80)	0.20(0.25)	4	1.75(0.48)	2.38(0.55)	0.63(0.63)
	16	5	3.20(0.58)	3.50(0.74)	0.30(0.30)	3	1.67(0.67)	1.67(0.33)	0.00(0.58)
	20	3	3.65(0.88)	3.83(1.48)	0.17(0.93)	4	1.75(0.48)	2.25(1.01)	0.50(0.84)
	28	3	3.65(0.88)	4.33(1.30)	0.67(0.93)	4	1.75(0.48)	2.75(0.52)	1.00(0.46)
Fingertips To Floor	2	5	17.00(5.96)	19.60(8.61)	2.60(7.81)	5	25.20(6.58)	23.00(7.23)	-2.20(1.28)
	4	5	17.00(5.96)	17.60(7.97)	0.60(7.46)	4	31.00(4.02)	27.89(5.40)	-3.13(1.69)
	8	5	17.00(5.96)	20.30(8.53)	3.30(6.90)	4	31.00(4.02)	25.63(6.06)	-5.38(2.05)
	12	5	17.00(5.96)	18.50(8.70)	1.50(8.50)	4	31.00(4.02)	27.00(5.93)	-4.00(2.38)
	16	5	17.00(5.96)	20.20(8.82)	3.20(7.89)	3	33.67(4.26)	33.00(4.58)	-0.67(0.33)
	20	3	15.00(9.29)	10.67(10.67)	-4.33(2.85)	4	31.00(4.02)	26.25(6.54)	-4.75(2.75)
	28	3	15.00(9.29)	10.33(10.33)	-4.67(2.67)	4	31.00(4.02)	26.50(6.36)	-4.50(2.99)

* - denotes statistically significant changes from baseline within groups at the 0.05 level (Wilcoxon signed rank test)

- denotes statistically significant changes from baseline within groups at the 0.05 level (Mann-Whitney test)

TABLE 6: Results of Analysis Within and Between Treatment Groups for Secondary Efficacy Data

		Experimental Drug #1				Indomethacin			
EFFICACY PARAMETER	WEEK	SAMPLE SIZE	PRETHERAPY MEAN(S.E.)	ON THERAPY MEAN(S.E.)	MEAN CHANGE(S.E.)	SAMPLE SIZE	BASELINE MEAN(S.E.)	THERAPY MEAN(S.E.)	MEAN CHANGE(S.E.)
Activity Impairment	2	5	3.20(0.58)	2.00(0.63)	-1.20(0.37)	5	3.20(0.58)	2.40(0.51)	-0.80(0.37)
	4	5	3.20(0.58)	2.00(0.63)	-1.20(0.37)	4	3.00(0.71)	2.00(0.41)	-1.00(0.41)
	8	5	3.20(0.58)	2.00(0.63)	-1.20(0.37)	4	3.00(0.71)	1.75(0.25)	-1.25(0.63)
	12	5	3.20(0.58)	2.40(0.67)	-0.80(0.66)	4	3.00(0.71)	2.25(0.63)	-0.75(0.25)
	16	5	3.20(0.58)	2.20(0.58)	-1.00(0.63)	3	3.33(0.89)	2.67(0.67)	-0.67(0.33)
	20	3	3.00(1.00)	1.67(0.33)	-1.33(0.88)	4	3.00(0.71)	2.50(0.87)	-0.50(0.29)
	28	3	3.00(1.00)	2.00(0.58)	-1.00(0.58)	4	3.00(0.71)	2.75(0.85)	-0.25(0.25)
Morning Stiffness (hrs.)	2	5	4.40(2.65)	0.65(0.19)	-3.75(2.82)	5	1.80(0.73)	1.00(0.55)	-.80(0.37)
	4	5	4.40(2.65)	0.65(0.19)	-3.75(2.82)	4	2.25(0.75)	1.19(0.64)	-1.06(0.46)
	8	5	4.40(2.65)	0.40(0.19)	-4.00(2.76)	4	2.25(0.75)	0.65(0.22)	-1.60(0.78)
	12	4	1.76(0.14)	0.38(0.24)	-1.38(0.31)	4	2.25(0.75)	0.63(0.24)	-1.63(0.64)
	16	4	1.75(0.14)	0.44(0.26)	-1.31(0.31)	3	2.66(0.88)	0.83(0.17)	-1.83(0.43)
	20	3	1.83(0.17)	0.33(0.33)	-1.50(0.29)	4	2.25(0.75)	0.88(0.43)	-1.38(0.55)
	28	3	1.83(0.17)	0.50(0.50)	-1.33(0.44)	4	2.25(0.75)	1.06(0.53)	-1.19(0.47)
Time to Walk 50 Feet (secs)	2	5	13.30(1.73)	12.80(1.11)	-0.50(0.63)	5	15.30(2.01)	15.90(1.23)	0.60(1.71)
	4	4	11.63(0.55)	11.38(0.38)	-0.25(0.25)	4	16.38(2.19)	15.00(1.47)	-1.38(0.44)
	8	4	13.38(2.23)	13.00(1.08)	-0.38(1.34)	2	12.75(1.25)	11.75(0.25)	-1.00(1.50)
	12	4	13.38(2.23)	12.75(1.18)	-0.63(1.14)	3	14.83(2.20)	13.40(1.23)	-1.43(1.48)
	16	4	13.63(2.19)	13.00(1.08)	-0.63(1.18)	3	18.00(2.08)	14.00(1.76)	-4.00(0.50)
	20	3	14.50(2.84)	12.17(0.93)	-2.33(1.92)	3	15.50(2.84)	11.17(0.90)	-4.33(2.22)
	28	2	16.50(3.50)	13.10(2.10)	-3.40(1.40)	3	15.50(2.84)	12.00(1.15)	-3.50(2.18)
Anterior Spinal Motion Flexion (cm)	2	5	2.70(0.90)	2.40(0.93)	-0.30(0.12)	5	0.60(0.40)	0.70(0.49)	0.10(0.10)
	4	5	2.70(0.90)	2.70(1.01)	0.00(0.45)	4	0.50(0.50)	1.75(0.48)	1.25(0.48)*
	8	5	2.70(0.90)	2.90(0.90)	0.20(1.57)	4	0.50(0.50)	1.25(0.32)	-0.75(0.32)
	12	5	2.70(0.90)	2.90(0.91)	0.20(1.59)	4	0.50(0.50)	1.13(0.38)	0.63(0.31)
	16	5	2.70(0.90)	2.90(0.84)	0.20(1.42)	3	0.00(0.00)	0.83(0.33)	0.83(0.33)
	20	3	3.84(1.01)	2.67(1.20)	-1.17(1.70)	4	0.50(0.50)	0.75(0.48)	0.25(0.25)
	28	3	3.83(1.01)	3.00(1.15)	-0.83(1.83)	4	0.50(0.50)	0.75(0.48)	0.25(0.25)
Lateral Left Spinal Motion Flexion (cm)	2	5	10.50(3.71)	11.10(3.71)	0.60(0.75)	5	3.50(1.50)	5.20(1.07)	1.70(0.86)
	4	5	10.50(3.71)	10.90(4.20)	0.40(1.39)	4	3.12(1.87)	5.50(0.87)	2.38(1.25)
	8	5	10.50(3.71)	11.90(3.89)	1.40(1.18)	4	3.12(1.87)	4.75(1.93)	1.63(1.46)
	12	5	10.50(3.71)	11.10(3.80)	0.60(1.04)	4	3.13(1.87)	4.63(2.25)	1.50(1.67)
	16	5	10.50(3.71)	11.90(4.04)	1.40(2.04)	3	1.67(1.67)	3.50(1.50)	1.83(2.74)
	20	3	15.66(3.48)	14.83(6.22)	-0.83(2.74)	4	3.12(1.87)	5.25(2.02)	2.13(1.76)
	28	3	15.67(3.48)	16.67(6.64)	1.00(3.21)	4	3.13(1.87)	4.38(2.61)	1.25(2.25)

* - denotes statistically significant changes from baseline within groups at the 0.05 level (Wilcoxon signed rank test)

TABLE 7: Results of Analysis Within and Between Treatment Groups for Secondary Efficacy Data

		Experimental Drug #1				Indomethacin			
EFFICACY PARAMETER	WEEK	SAMPLE SIZE	PRETHERAPY MEAN(S.E.)	ON THERAPY MEAN(S.E.)	MEAN CHANGE(S.E.)	SAMPLE SIZE	BASELINE MEAN(S.E.)	ON THERAPY MEAN(S.E.)	CHANGE(S.E.)
Lateral	2	5	10.80(4.11)	11.40(4.87)	0.60(1.06)	5	4.50(1.97)	4.60(1.44)	0.10(2.48)
Right	4	5	10.80(4.11)	12.10(3.99)	1.30(0.98)	4	3.88(2.42)	5.63(1.43)	1.75(3.03)
Spinal	8	5	10.80(4.11)	12.40(4.78)	1.60(0.99)	4	3.88(2.42)	3.88(1.48)	0.00(2.39)
Motion	12	5	10.80(4.11)	12.50(4.18)	1.70(0.30)*	4	3.88(2.42)	5.38(1.80)	1.50(2.67)
Flexion	16	5	10.80(4.11)	13.10(4.74)	2.30(0.80)*	3	3.33(3.33)	8.50(6.50)	5.17(8.66)
(cm)	20	3	15.50(5.27)	17.00(5.69)	1.50(1.04)	4	3.88(2.42)	4.88(1.91)	1.00(2.88)
	28	3	15.50(5.27)	1.700(6.35)	1.50(2.75)	4	3.88(2.42)	3.88(2.26)	0.00(3.11)
Occiput	2	5	8.60(2.40)	9.40(2.54)	0.80(0.73)	5	13.60(3.22)	13.20(3.51)	-0.40(0.58)
To Wall	4	5	8.60(2.40)	7.90(2.27)	-0.70(0.58)	4	11.75(3.40)	10.75(3.12)	-1.00(1.08)
Test	8	5	8.60(2.40)	7.90(2.38)	-0.70(0.37)	4	11.76(3.40)	11.13(2.85)	-0.63(0.63)
(cm)	12	5	8.60(2.40)	7.50(2.42)	-1.10(0.33)	4	11.75(3.40)	10.50(3.20)	-1.25(0.48)
	16	5	8.60(2.40)	8.70(2.52)	0.10(0.60)	3	13.33(4.26)	12.33(3.70)	-1.00(0.58)
	20	3	6.33(3.53)	6.00(3.46)	-0.33(0.67)	4	11.76(3.40)	10.88(2.96)	-0.88(0.88)
	28	3	6.33(3.53)	6.00(3.46)	-0.33(0.67)	4	11.75(3.40)	11.75(3.04)	0.00(0.91)
Intermalleor	2	4	72.25(12.12)	82.25(10.84)	10.00(2.20)*	5	69.60(6.01)	83.60(9.36)	14.00(5.07)
Straddle	4	4	72.25(12.12)	84.00(8.88)	11.75(3.57)*	4	70.25(7.71)	86.50(6.55)	16.25(3.75)*
Distance	8	4	72.25(12.12)	85.25(10.16)	13.00(4.88)*	4	70.25(7.71)	88.25(6.24)	18.00(4.60)*
(cm)	12	4	72.25(12.12)	87.13(10.84)	14.88(6.42)*	4	70.25(7.71)	91.25(5.58)	21.00(3.72)*
	16	4	72.25(12.12)	87.75(7.64)	15.50(8.53)	3	68.66(10.67)	91.33(6.17)	22.67(5.24)
	20	3	72.33(17.14)	86.00(14.64)	13.67(9.70)	4	70.25(7.71)	91.75(4.97)	21.50(5.24)*
	28	3	72.34(17.14)	84.67(15.34)	12.33(12.41)	4	70.25(7.71)	91.25(6.76)	21.00(4.45)*

* - denotes statistically significant changes from baseline within groups at the 0.05 level (Wilcoxon signed rank test)

TABLE 8: LISTING BY PATIENT OF CONCOMITANT MEDICATIONS

PATIENT #	STUDY DRUG	AGE	SEX	WEEK TAKEN (# OF DAYS)	CONCOMITANT THERAPY (CATEGORY)
6	Expt. Drug #1	41	M	2(1)	Instantin & Instantine R-Plus (Analgesic)
				3(4)	Instantin & Instantine Plus (Analgesic)
				4(1)	Empracet 30 [Acetominophen - codeine] (Analgesic)
				8(28)	Empracet 30 (Analgesic)
				13(1)	Indocid (Anti-inflammatory - Analgesic)
9	Expt. Drug #1	50	M	14(1)	Nalfon (Anti-inflammatory - Analgesic) Orudis Suppository Orudis (Both Anti-inflammatory - Analgesic)

TABLE 9: Listing of Patients of Reasons for Discontinuation

PT #	THERAPY	AGE	SEX	ILLNESS DURATION (YRS)	REASON FOR DISCONTINUATION	DOSE AT TIME OF DISCONTINUATION	TOTAL WEEKS IN STUDY
6	Experimental Drug #1	41	M	4	Unsatisfactory response-requires additional medication	1500 mg	16
8	Indomethacin	31	M	15	Severe headaches (adverse experiences)	150 mg	2
9	Experimental Drug #1	50	M	20	Severe pain in neck (cervical spine)	1500 mg	16

TABLE 10: Adverse Reactions By Patients

PT #	STUDY DRUG	ADVERSE REACTION	PRESENT PRETHERAPY	WEEK OF ONSET	WEEK OF DISAPPEARANCE	SEVERITY	TREATMENT RELATED	ACTION/OUTCOME
1	Indocid	Abdominal Cramps	yes	3	10	mild	no	Tolerated with continued therapy
		Nervousness	yes	9	9	moderate	possible	Tolerated with continued therapy
		Memory loss	yes	9	15	moderate	possible	Tolerated with continued therapy
		Nodule-left breast	no	18	19	mild	no	Therapy discontinued due to adverse experience
2	Expt. Drug #1	Dizziness	yes	2	31	moderate	no	Tolerated with continued therapy
		Memory loss	no	2	31	moderate	probable	Tolerated with continued therapy
		Headaches	no	2	2	moderate	probable	Disappeared with continued therapy
		Lassitude	no	2	4	moderate	possible	Tolerated with continued therapy
		Aggresivity	no	2	4	mild	possible	Tolerated with continued therapy
3	Indocid	Abdominal Cramps	no	2	6	mild	possible	Tolerated with continued therapy
		Drowsiness	no	10	24	mild	possible	Tolerated with continued therapy
		Dizziness	no	18	20	mild	possible	Tolerated with continued therapy
4	Expt. Drug #1	Drowsiness	no	3	15	mild	possible	Tolerated with continued therapy
		Constipation	no			moderate	possible	Requires symptomatic therapy - Doxidan h.s. p.r.n
10	Indocid	Psychomotor Episodes	no	11		mild	no	Tolerated with continued therapy

TABLE 11: SUMMARY OF ADVERSE REACTIONS

ADVERSE REACTION	INCIDENCE (%) EXPT. DRUG #1 N = 5	 INDOMETHACIN N = 5
DROWSINESS	1 (20)	1 (20)
DIZZINESS	1 (20)*	1 (20)
ABDOMINAL CRAMPS	0 (0)	2 (40)*
MEMORY LOSS	1 (20)	1 (20)
HEADACHES	1 (20)	0 (0)
NERVOUSNESS	0 (0)	1 (20)
CONSTIPATION	1 (20)	0 (0)
LASSITUDE	1 (20)	0 (0)
AGGRESSIVITY	1 (20)	0 (0)
NODULE-LEFT BREAST	0 (0)	1 (20)*
PSYCHO-MOTOR EPISODES	0 (0)	1 (20)*

* - not treatment related

TABLE 12: NORMAL LABORATORY VALUES - ANKYLOSING SPONDYLITIS

Hemoglobin (Males)* (gm/100 ml)	12 - 15
Hematocrit (Males)* (%)	37 - 47
White Blood Cells (1000's/cu mm)	4.8 - 10.8
Segmented neutrophiles (%)	40 - 65
Band neutrophiles (%)	0 - 7
Lymphocytes (%)	24 - 43
Monocytes (%)	0 - 8
Eosinophiles (%)	0 - 4
Basophiles (%)	0 - 0.5
Platelets (estimate)	150 - 400
Erythrocyte sedimentation rate (mm/hr)	0 - 20
Total protein (gm%)	6 - 8
Albumin (gm%)	3.5 - 5
Calcium (mg%)	8.5 - 10.5
Phosphorous (mg%)	2.5 - 4.5
Cholesterol (mg%)	140 - 260
Uric Acid (mg%)	2.5 - 7.8
Creatine (mg%)	0.4 - 1.3
Total bilirubin (mg%)	0 - 1.2
Alkaline Phosphatase (I.U.)	30 - 120
SGOT (I.U. ml)	0 - 24
Chloride (mEq/L)	85 - 110
Potassium (mEq/L)	3.5 - 5

Table 12 continued

Sodium (mEq/L)	135 - 150
Bicarbonate (mEq/L)	23 - 29
BUN (Blood Urea Nitrogen)(mg%)	7 - 24
Glucose (mg%, random)	65 - 110

* - Note: Hemoglobin and Hematocrit female values are not presented due to only 2 females being in the trial.

TABLE 13: Results of Analysis Within and Between Treatment Groups for Two Laboratory Determinations - Hematology Data

		Experimental Drug #1				Indomethacin			
PARAMETER	WEEK	SAMPLE SIZE	PRETHERAPY MEAN(S.E.)	ON THERAPY MEAN(S.E.	MEAN CHANGE(S.E.)	SAMPLE SIZE	PRETHERAPY MEAN(S.E.)	ON THERAPY MEAN(S.E.)	MEAN CHANGE(S.E.)
White	2	5	7.54(0.64)	6.94(0.59)	-0.60(0.51)	5	8.52(1.37)	9.56(1.37)*	1.04(0.43)#
Blood Cells	8	5	7.54(0.64)	7.78(0.47)	0.24(0.25)	4	8.70(1.75)	9.08(2.14)	0.38(0.63)
(thousands/cu mm)	16	4	7.43(0.81)	7.30(0.76)	-0.13(0.37)	4	8.70(1.75)	8.80(2.32)	0.10(0.74)
	28	3	7.47(1.15)	7.67(1.52)	0.20(0.40)	4	8.70(1.75)	9.55(2.29)	0.85(0.76)
Erythrocyte	2	5	20.00(4.40)	14.60(2.54)	-5.40(3.03)	5	21.00(2.37)	15.40(1.75)	-5.60(1.12)
Sedimentation	8	4	18.75(5.54)	16.75(3.97)	-2.00(6.05)	2	24.00(2.00)	13.00(7.00)	-11.00(5.00)
Rate (ESR)	16	4	23.25(3.84)	19.00(4.60)	-4.25(7.60)	4	23.25(0.95)	14.25(3.90)	-9.00(3.49)
(mm/hr)	28	3	27.00(1.15)	15.67(5.24)	-11.33(6.01)	4	23.25(0.95)	12.50(1.55)	-10.75(1.60)

* - denotes statistically significant changes from baseline within therapy groups at the 0.05 level

- denotes statistically significant differences between therapy groups at the 0.05 level

TABLE 14: Results of Analysis Within and Between Treatment Groups for Two Laboratory Determinations - Blood Chemistry Data

		Experimental Drug #1				Indomethacin			
PARAMETER	WEEK	SAMPLE SIZE	PRETHERAPY MEAN(S.E.)	ON THERAPY MEAN(S.E.)	MEAN CHANGE(S.E.)	SAMPLE SIZE	PRETHERAPY MEAN(S.E.)	ON THERAPY MEAN(S.E.)	MEAN CHANGE(S.E.)
Phosphorous (mg%)	2	5	2.80(0.19)	2.82(0.21)	0.02(0.22)	5	3.28(0.19)	3.38(0.23)*	0.10(0.20)
	8	5	2.80(0.19)	2.80(0.11)	0.00(0.09)	4	3.35(0.23)	3.45(0.09)	0.10(0.17)
	16	4	2.83(0.24)	3.08(0.08)	0.25(0.24)	4	3.35(0.23)	3.30(0.17)	-0.05(0.10)
	28	3	2.63(0.20)	2.73(0.20)	0.10(0.00)	4	3.35(0.23)	3.36(0.25)*	0.28(0.22)
Chloride (mEq/L)	2	5	100.60(1.33)	102.00(0.95)*	1.40(0.51)	5	99.20(0.92)	99.60(1.17)	0.40(2.04)
	8	5	100.60(1.33)	83.20(18.57)	-17.40(18.41)	4	98.75(1.03)	99.00(1.68)	0.25(2.39)
	16	4	101.75(0.85)	103.75(0.48)	2.00(1.08)	4	98.75(1.03)	103.25(1.11)*	4.50(1.26)
	28	3	102.33(0.88)	103.67(0.88)	1.33(1.76)	4	98.75(1.03)	101.25(1.49)	2.50(2.47)

* - denotes statistically significant changes from baseline within therapy groups at the 0.05 level

TABLE 15: Results of Analysis Within and Between Therapy Groups for Laboratory Determinations - Urinalysis Data

		Experimental Drug #1				Indomethacin			
PARAMETER	WEEK	SAMPLE SIZE	PRETHERAPY MEAN(S.E.)	ON THERAPY MEAN(S.E.)	MEAN CHANGE(S.E.)	SAMPLE SIZE	PRETHERAPY MEAN(S.E.)	ON THERAPY MEAN(S.E.)	MEAN CHANGE(S.E.)
pH	2	5	6.20(0.41)	5.80(0.34)	-0.40(0.48)	5	6.80(0.46)	6.40(0.58)	-0.40(0.58)
	8	5	6.20(0.41)	6.60(0.43)	0.40(0.37)	4	6.50(0.46)	6.50(0.61)	0.00(0.71)
	16	4	6.00(0.46)	6.25(0.48)	0.25(0.32)	4	6.50(0.46)	6.50(0.54)	0.00(0.41)
	28	3	5.67(0.44)	6.00(0.58)	0.33(0.83)	4	6.50(0.46)	6.13(0.43)	-0.38(0.55)

TABLE 16: Results of Analysis Within and Between Treatment Groups for Vital Signs (Safety) Data

		Experimental Drug #1				Indomethacin			
PARAMETER	WEEK	SAMPLE SIZE	PRETHERAPY MEAN(S.E.)	ON THERAPY MEAN(S.E.)	MEAN CHANGE(S.E.)	SAMPLE SIZE	PRETHERAPY MEAN(S.E.)	ON THERAPY MEAN(S.E.)	MEAN CHANGE(S.E.)
Diastolic Blood Pressure (mm Hg)	2	5	77.00(4.90)	76.00(4.00)	-1.00(3.31)	5	87.00(3.74)	83.00(2.00)	-4.00(1.87)
	4	5	77.00(4.90)	79.00(3.67)	2.00(2.55)	4	88.75(4.27)	82.50(2.50)	-6.25(4.73)
	8	5	77.00(4.90)	78.00(3.74)	1.00(2.45)	4	88.75(4.27)	88.75(5.91)	0.00(5.40)
	12	5	77.00(4.90)	77.00(3.00)	0.00(5.24)	4	88.75(4.27)	87.50(4.79)	-1.25(4.27)
	16	4	76.25(6.25)	77.50(4.79)	1.25(6.57)	3	91.67(4.41)	80.00(5.77)	-11.67(6.01)
	20	3	75.00(8.66)	81.67(4.41)	6.67(4.41)	4	88.75(4.27)	85.00(2.89)	-3.75(2.39)
	28	3	75.00(8.66)	78.33(4.41)	3.33(4.41)	4	88.75(4.27)	81.25(1.25)	-7.50(4.33)
Pulse (beats.minute)	2	5	74.75(1.89)	77.00(5.50)	2.25(4.66)	5	91.20(6.18)	90.80(7.53)#	-0.40(3.96)
	4	5	74.75(1.89)	78.00(2.58)	3.25(2.14)	4	95.25(6.02)	91.50(7.14)*	-3.75(3.75)
	8	5	74.75(1.89)	81.00(5.20)	6.25(4.80)	4	95.25(6.02)	93.50(9.91)	-1.75(4.13)
	12	5	74.75(1.89)	79.50(5.68)	4.75(3.82)	4	95.25(6.02)	87.50(8.38)	-7.75(6.28)
	16	4	74.75(1.89)	80.75(3.25)	6.00(4.06)	3	95.33(8.51)	98.00(11.13)*	2.67(2.67)
	20	3	75.67(2.33)	84.00(0.00)	8.33(2.33)	4	95.25(6.02)	92.00(4.69)	-3.25(7.97)
	28	3	75.67(2.33)	80.00(2.00)	4.33(0.88)	4	95.25(6.02)	92.00(9.70)	-3.25(4.84)

* - denotes statistically significant changes from baseline within groups at the 0.05 level

- denotes statistically significant differences between therapy groups at the 0.05 level

TABLE 17: Termination Summary

PT #	STUDY DRUG	CONCOMITANT THERAPY	REASON FOR EARLY TERMINATION	PHYSICIAN'S EVALUATION
1	Indomethacin	none	- - - -	satisfactory response
2	Expt. Drug #1	none	- - - -	satisfactory response
3	Indomethacin	none	- - - -	satisfactory response
4	Expt. Drug #1	none	- - - -	satisfactory response
5	Indomethacin	none	- - - -	satisfactory response
6	Expt. Drug #1	Instantin & Instantin Plus Empracet 30 Indocid	requires additional medication	unsatisfactory response
7	Expt. Drug #1	none	error in timing of in final visit	satisfactory response
8	Indomethacin	none	adverse experiences	unsatisfactory response
9	Expt. Drug #1	Nalfon Orudis	marked pain in neck (cervical spine)	unsatisfactory response
10	Indomethacin	none	- - - -	satisfactory response

TABLE 18: SUMMARY OF RESULTS USING BOOTSTRAP PROCEDURE

VARIABLE:	PULSE		WBC		PHOSPHOROUS		PH	
	EXPT. DRUG #1	INDO -CID	EXPT. DRUG #1	INDO -CID	EXPT. DRUG #1	INDO -CID	EXPT. DRUG #1	INDO -CID
USING GROUPED DATA:								
Wilcoxon Signed Rank Test		$p<0.05$ at 4,16 wks.		$p<0.05$ at 2 wks.		$p<0.05$ at 2, 28 wks.		
Mann-Whitney Test		$p<0.05$ at 2 wks.		$p<0.05$ at 2 wks.				
Two-Sample T-Test	$p=0.0216$ at 0 wks.	$p=0.0216$						
BOOTSTRAP PROCEDURE WITH INDIVIDUAL DATA:								
Wilcoxon Signed Rank Statistic			$p=0.059$ at 0,8, 28 wks.	$p=0.059$ at 0 wks.		$p=0.059$ at 8,16, 28.	$p=0.106$ at 16 wks.	$p=0.59$ at 0 wks.
Mann-Whitney Test			$p=0.095$ at 2 wks.	$p=0.095$ at 2 wks.		$0=0.04$, 0.01 at 2,8,28		$p=0.0601$ at 0 wks.
Two-Sample T-Test	$p=0.019$ at 0 wks.	$p=0.019$	$p=0.091$ at 2 wks.	$p=0.063$ at 2 wks.		$p=0.0004$ at 16 wks.	$p=0.44$ at 8 wks.	$p=0.44$ at 8 wks.

NOTE: 'BLANK' means that corresponding tests agree, and in cases where no statistical significance was found, the lowest value of α is recorded for whichever time point corresponded in the trial.

TABLE 19: ANOVA TABLES OF AGE, WEIGHT AND HEIGHT AFTER BOOTSTRAP METHOD APPLIED TO TEST FOR DIFFERENCES BETWEEN DRUG GROUPS

```
                         ANALYSIS OF VARIANCE (AGE)

SOURCE          DF          SS          MS            F
FACTOR           1         1.6         1.6         0.02
ERROR            8       642.8        80.4

TOTAL            9       644.4

LEVEL            N        MEAN       STDEV
C5               5       38.00        7.78
C17              5       37.20       10.01

POOLED STDEV = 8.96
   MTB > AOVONEWAY ON WEIGHT OF EXPT. DRUG #1, WEIGHT OF INDOCID GRPS.
```

```
                         ANALYSIS OF VARIANCE (WEIGHT)

SOURCE          DF          SS          MS            F
FACTOR           1       122.5       122.5         1.29
ERROR            8       760.4        95.1

TOTAL            9       882.9

LEVEL            N        MEAN       STDEV
C3               5       53.40       10.09
C15              5       60.40        9.40

POOLED STDEV = 9.75
   MTB > AOVONEWAY ON HEIGHT OF EXPT, DRUG#1, HEIGHT OF INDOCID GRPS.
```

```
                         ANALYSIS OF VARIANCE (HEIGHT)

SOURCE          DF          SS          MS            F
FACTOR           1       220.9       220.9         4.98
ERROR            8       355.2        44.4

TOTAL            9       576.1

LEVEL            N        MEAN       STDEV
C4               5       156.6         6.5
C16              5       166.6         6.8

POOLED STDEV = 6.7
   MTB > HEK
         LP TWOT
```

TABLE 20: χ^2 CONTINGENCY TABLE FOR SEX

EXPECTED FREQUENCIES ARE PRINTED BELOW OBSERVED FREQUENCIES

ROW CLASSIFICATION - SEX OF EXPT. DRUG #1 GROUP
COLUMN CLASSIFICATION - SEX OF INDOCID GROUP

	1	2	TOTALS
1	0 .4	1 .6	1
2	2 1.6	2 2.4	4
TOTALS	2	3	5

TOTAL CHI SQUARE = Could not be computed due to insufficient data

.40 +	.27 +
.10 +	.07 +

NOTE 2 CELLS WITH EXPECTED FREQUENCIES LESS THAN 1 DEGREES OF FREEDOM = (2 - 1) x (2 - 1) = 1

NOTE 4 CELLS WITH EXPECTED FREQUENCIES LESS THAN 5

TABLE 21: ANOVA TABLES OF WBC FOR INDOMETHACIN GROUP

ANALYSIS OF VARIANCE FOR COMMON SLOPE

SOURCE	DF	SS	MS	F
POOL REG	1.00000	0.61098	0.61098	0.75434
B SLOPES	4.00000	4.75974	1.18993	1.46914
SEP REGR	5.00000	5.37072	1.07414	1.32618
RESIDUAL	15.0000	12.1493	0.8100	
TOTAL	20.000	17.520	0.876	

ANALYSIS OF VARIANCE FOR COINCIDENT LINES

SOURCE	DF	SS	MS	F
REGRESS	1.00000	0.61096	0.61096	0.75432
DIFF POS	4.000	244.178	61.045	75.368*
RESIDUAL	19.00000	16.9090	0.8100	
TOTAL	24.000	261.698	10.904	

* - Statistically significant difference ($\alpha = 0.05$)

TABLE 22: ANOVA TABLES OF WBC FOR EXPERIMENTAL GROUP

ANALYSIS OF VARIANCE FOR COMMON SLOPE

SOURCE	DF	SS	MS	F
POOL REG	1.00000	0.20939	0.20939	0.43170
B SLOPES	4.00000	3.25909	0.81477	1.67982
SEP REGR	5.00000	3.46848	0.69370	1.43020
RESIDUAL	15.0000	7.2755	0.4850	
TOTAL	20.0000	10.7440	0.5372	

ANALYSIS OF VARIANCE FOR COINCIDENT LINES

SOURCE	DF	SS	MS	F
REGRESS	1.00000	0.20936	0.20936	0.43164
DIFF POS	4.0000	32.9504	8.2376	16.9835*
RESIDUAL	19.0000	10.5346	0.4850	
TOTAL	24.0000	43.6944	1.8206	

* - Statistically significant differences ($\alpha = 0.05$)

TABLE 23: ANOVA TABLES OF PHOSPHOROUS FOR INDOMETHACIN GROUP

ANALYSIS OF VARIANCE FOR COMMON SLOPE

SOURCE	DF	SS	MS	F
POOL REG	1.00000	0.22245	0.22245	3.84003
B SLOPES	4.00000	0.74062	0.18515	3.19624*
SEP REGR	5.00000	0.96307	0.19261	3.32500 *
RESIDUAL	15.0000	0.8689	0.0579	
TOTAL	20.0000	1.8320	0.0916	

ANALYSIS OF VARIANCE FOR COINCIDENT LINES

SOURCE	DF	SS	MS	F
REGRESS	1.00000	0.22245	0.22245	3.84004
DIFF POS	4.00000	1.47440	0.36860	6.36298*
RESIDUAL	19.0000	1.6096	0.0579	
TOTAL	24.0000	3.3064	0.1378	

* - Statistically significant difference ($\alpha = 0.05$)

TABLE 24: ANOVA TABLES OF PHOSPHOROUS FOR EXPERIMENTAL GROUP

ANALYSIS OF VARIANCE FOR COMMON SLOPE

	SOURCE	DF	SS	MS	F
POOL REG		1.00000	0.00893	0.00893	0.12942
B SLOPES		4.00000	0.30836	0.07709	1.11757
SEP REGR		5.00000	0.31729	0.06346	0.91994
RESIDUAL		15.0000	1.0347	0.0690	
TOTAL		20.0000	1.3520	0.0676	

ANALYSIS OF VARIANCE FOR COINCIDENT LINES

	SOURCE	DF	SS	MS	F
REGRESS		1.00000	0.00893	0.00893	0.12942
DIFF POS		4.00000	1.11040	0.27760	4.02431*
RESIDUAL		19.0000	1.3431	0.0690	
TOTAL		24.0000	2.4624	0.1026	

* - Statistically significant difference ($\alpha = 0.05$)

TABLE 25: ANOVA TABLES OF pH OF URINE FOR INDOMETHACIN GROUP

ANALYSIS OF VARIANCE FOR COMMON SLOPE

SOURCE	DF	SS	MS	F
POOL REG	1.00000	0.71386	0.71386	1.20359
B SLOPES	4.00000	2.53753	0.63438	1.06959
SEP REGR	5.00000	3.25138	0.65028	1.09639
RESIDUAL	15.0000	8.8966	0.5931	
TOTAL	20.0000	12.1480	0.6074	

ANALYSIS OF VARIANCE FOR COINCIDENT LINES

SOURCE	DF	SS	MS	F
REGRESS	1.00000	0.71386	0.71386	1.20359
DIFF POS	4.0000	10.1696	2.5424	4.2866*
RESIDUAL	19.0000	11.4341	0.5931	
TOTAL	24.0000	22.3176	0.9299	

* - Statistically significant difference ($\alpha = 0.05$)

TABLE 26: ANOVA TABLES OF pH OF URINE FOR EXPERIMENTAL GROUP

ANALYSIS OF VARIANCE FOR COMMON SLOPE

SOURCE	DF	SS	MS	F
POOL REG	1.00000	0.00108	0.00108	0.00180
B SLOPES	4.00000	2.88350	0.72088	1.20421
SEP REGR	5.00000	2.88458	0.57692	0.96373
RESIDUAL	15.0000	8.9794	0.5986	
TOTAL	20.0000	11.8640	0.5932	

ANALYSIS OF VARIANCE FOR COINCIDENT LINES

SOURCE	DF	SS	MS	F
REGRESS	1.00000	0.00107	0.00107	0.00178
DIFF POS	4.00000	3.94160	0.98540	1.64610
RESIDUAL	19.0000	11.8629	0.5986	
TOTAL	24.0000	15.8056	0.6586	

REFERENCES

1. Cohen, J. Statistical Power Analysis for the Behavioral Sciences. (1969) Academic Press.

2. Conover, W. J. and Iman, R.L. Rank Transformations as a Bridge Between Parametric and Nonparametrac Statistics. The American Statistician, Aug. 1981, Vol.35(3), p. 87-95.

3. Diaconis, P. and Efron, B. Computer-Intensive Methods in Statistics. Scientific American, May 1981, p. 116-129.

4. Draper, N. and Smith, H. Applied Regression Analysis. (1966) Wiley & Sons, Inc.

5. Efron, B. and Gong, G. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. The American Statistician, Feb. 1983, Vol.37(1), p. 36-48.

6. Fleiss, J.L. Statistical Methods for Rates and Proportions. (1981) Wiley and Sons, Inc.

7. Harter, H.L. and Owen, D.B. (Editors) Selected Tables in Mathematical Statistics. (1970) Published by the Markham Series in Statistics (Chicago). Vol. I.

8. Hollander, M. and Wolfe, D.A. Nonparametric Statistical Methods. (1973) Wiley & Sons.

9. Krogh, C.M.E. (Editor in Chief) Compendium of Pharmaceuticals and Specialties 1983. 18th Edition, Published by Canadian Pharmaceutacal Association.

10. Lehmann, E.L. Testing Statistical Hypothesis. (1975) Wiley & Sons, Ltd.

11. Lehmann, E.L. Nonparametrics: Statistical Methods Based on Ranks. (1975) Holden-Day Inc.

12. Magee, R.A. A Statistical Rank Test for Analysing Biomedical Data. (1976) McMaster Unversity.

13. Rafnsson, V., Bengtsson, C. and Lurie, M. Erythrocyte Sedimentation Rate in Women with Different Manifestations of Joint Disease. Scandinavian Journal of Rheumatology, 1982, Vol.11(2) p. 87-95.

14. Remington, R. and Schork, M.A. Statistics with Applications to the Biological and Health Sciences. (1970) Prentice-Hall.

15. Ryan, T.A., Joiner, B.L. and Ryan, B. Minitab Reference Manual. (1982) Published by Pennsylvania State University.

16. Schulak, D.J. et al. The Erythrocyte Sedimentation Rate in Orthopaedic Patients. Clinical Orthopaedics and Related Research, 1982, Vol.167, p. 197-202.

17. Snedecor, G.W. and Cochran, W.G. Statistical Methods. (1967) Iowa State Unversity Press. (6th Edition).

18. Steel, R.G.D. and Torrie, J.H. Principles and Procedures of Statistics: A Biometric Approach. (1960) McGraw-Hill.

19. Stitt, L. MTBPLRG1.HELUVA.CEB. (1984) McMaster University, Department of Clinical Epidemiology and Biostatistics.

20. The Reader's Digest. The Reader's Digest Great Encyclopaedic Dictionary. (1964) Published by The Reader's Digest Association Ltd.

FIGURES

APPENDIX A

Note: In the following Figures, Therapy 1 refers to the Experimental Drug #1 group and Therapy 2 refers to the indomethacin group

FIGURE 1 :ANKYLOSING SPONDYLITIS

OBSERVER'S OPINION VERSUS TIME

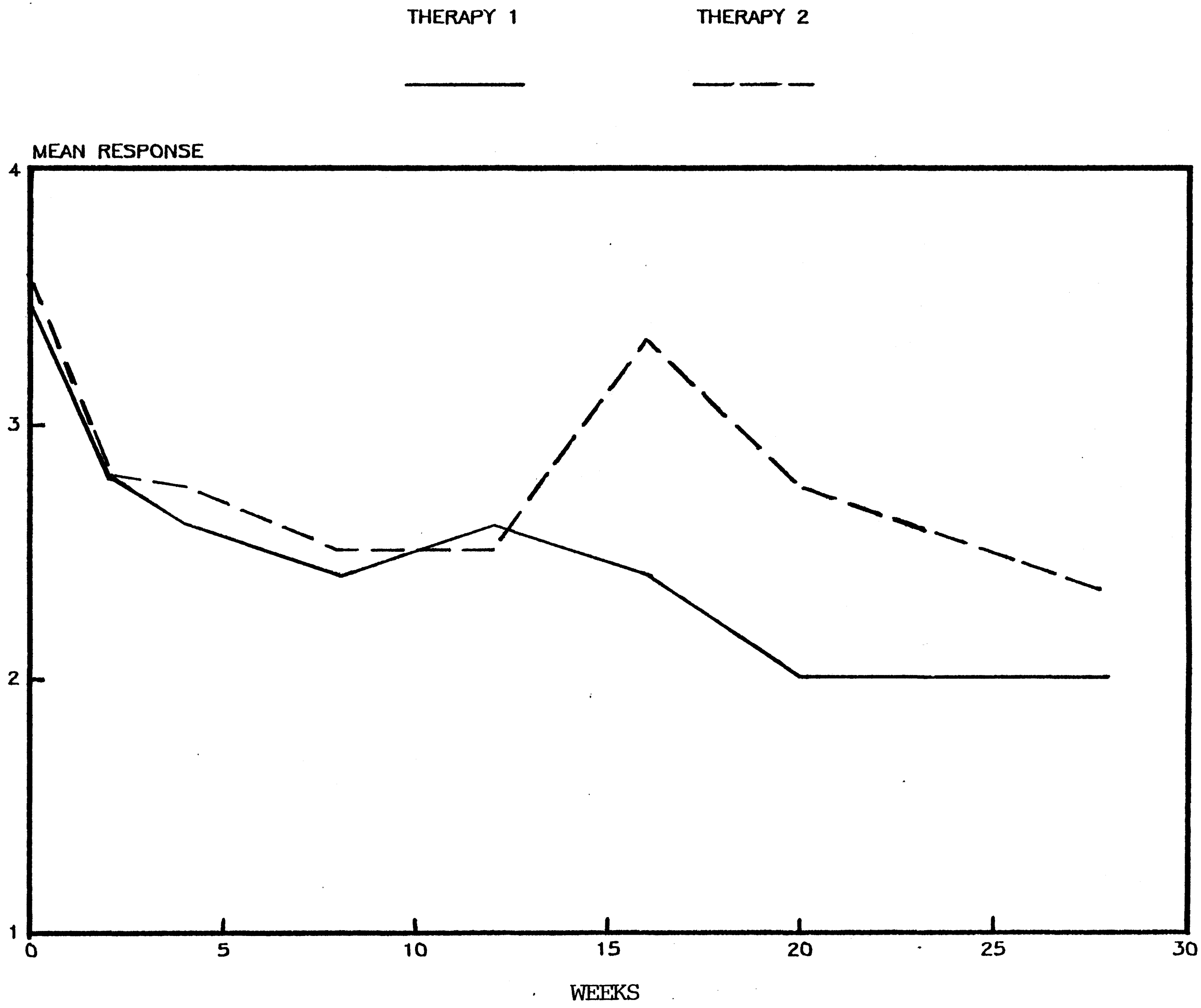
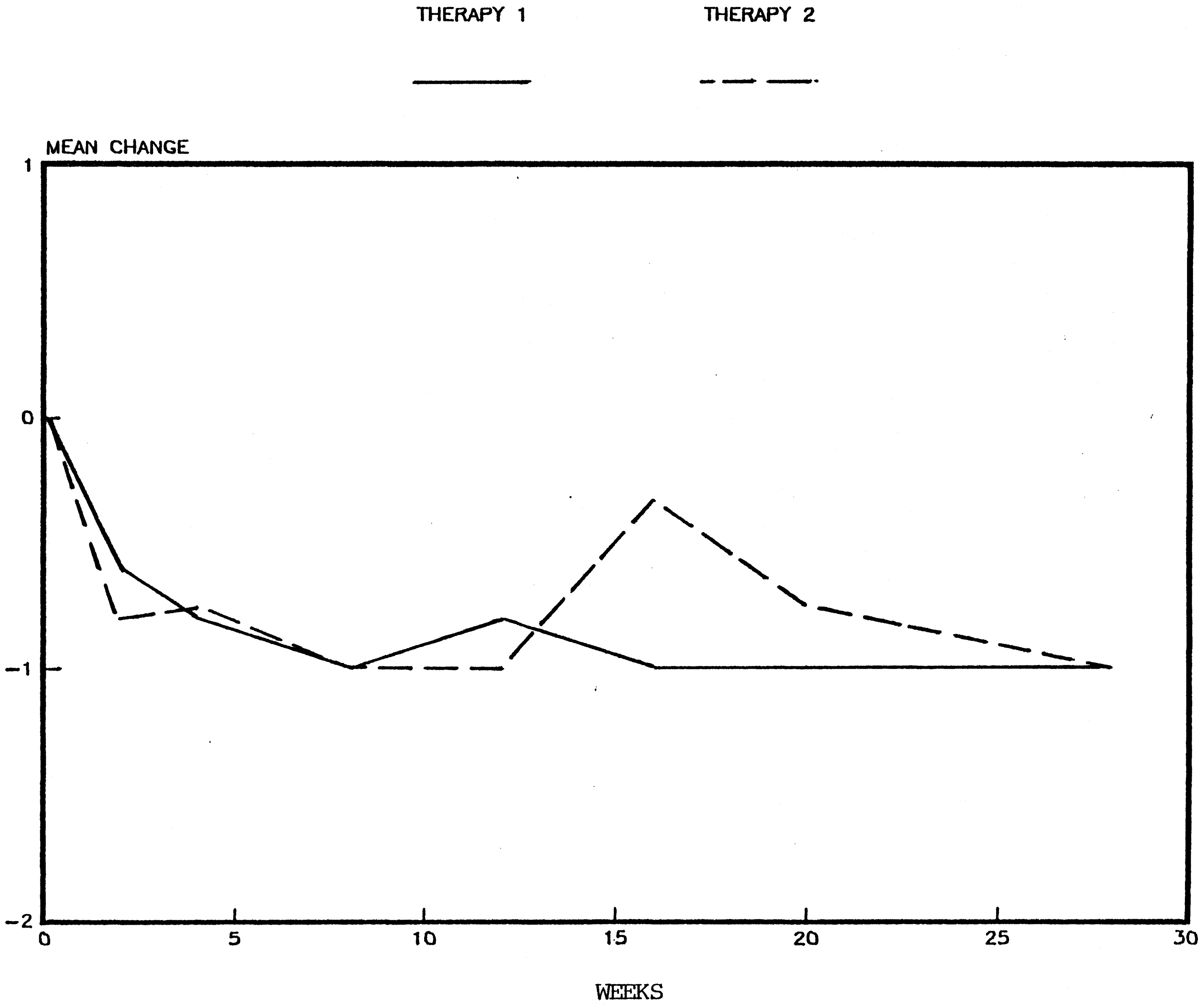


FIGURE 2 :ANKYLOSING SPONDYLITIS

OBSERVER'S OPINION VERSUS TIME



SIG. DIFF WITHIN GROUPS AT 8 WEEKS(P<0.05)

FIGURE 3: ANKYLOSING SPONDYLITIS

PATIENT'S OPINION VERSUS TIME

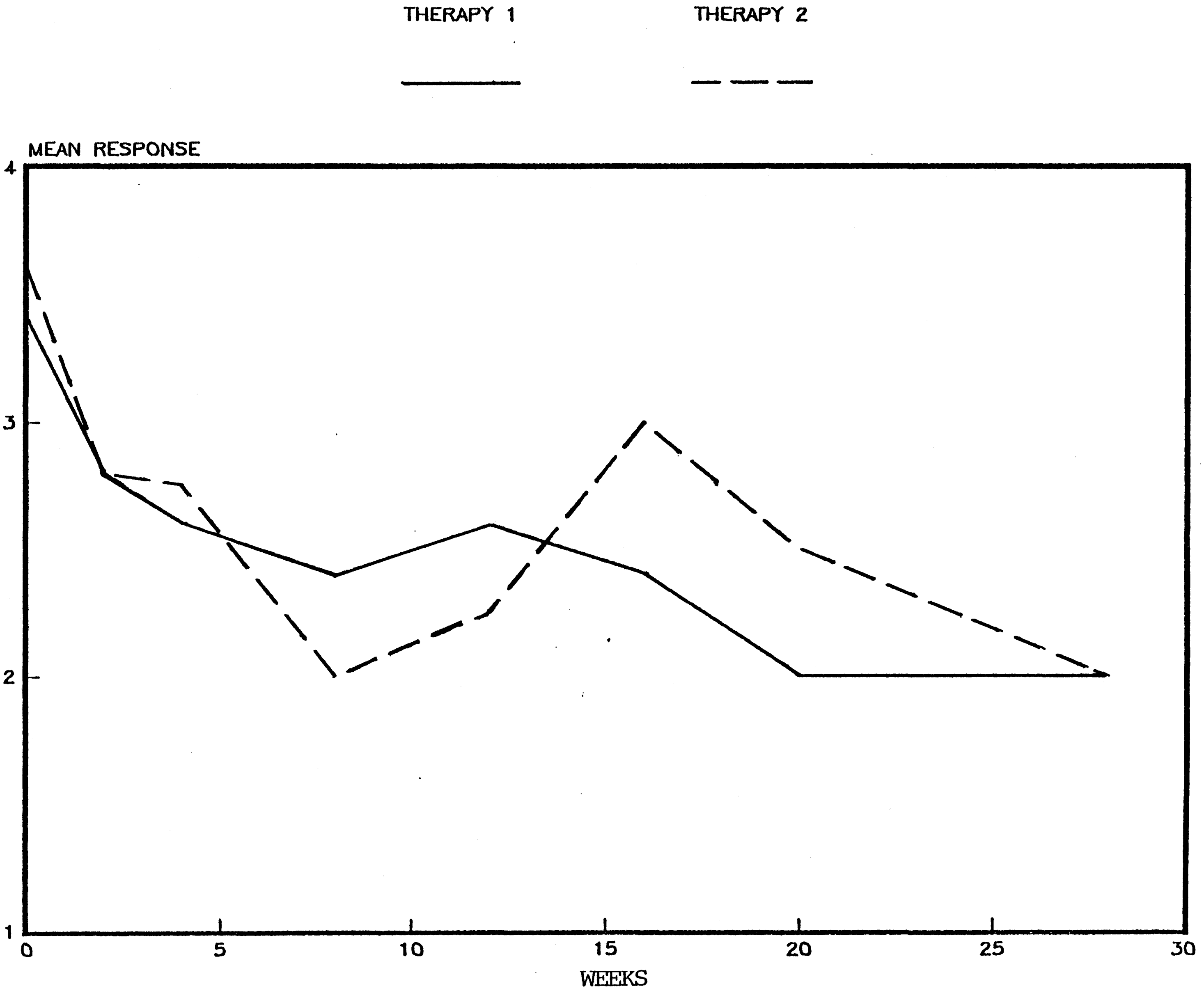


FIGURE4 : ANKYLOSING SPONDYLITIS

PATIENT'S OPINION VERSUS TIME

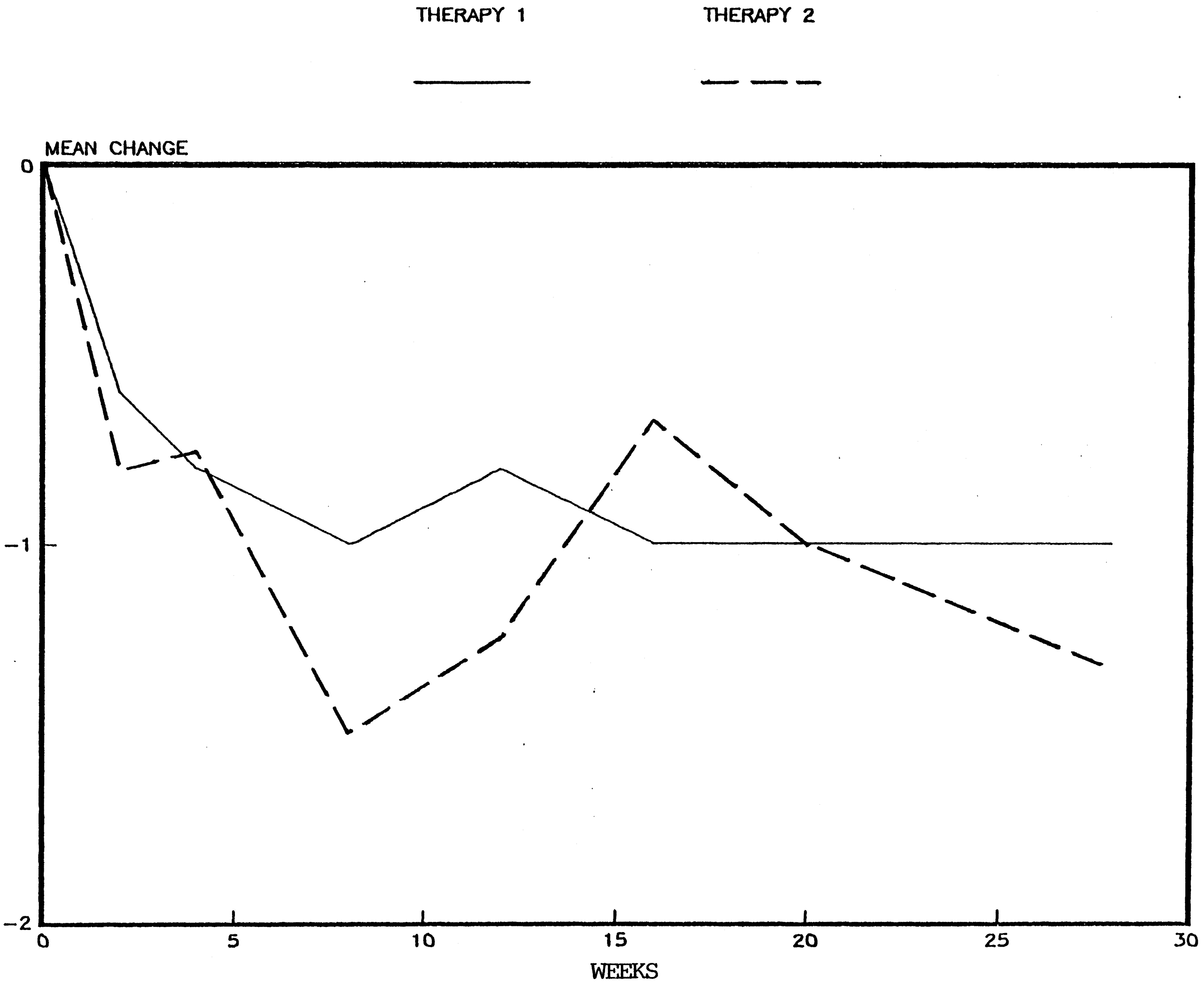
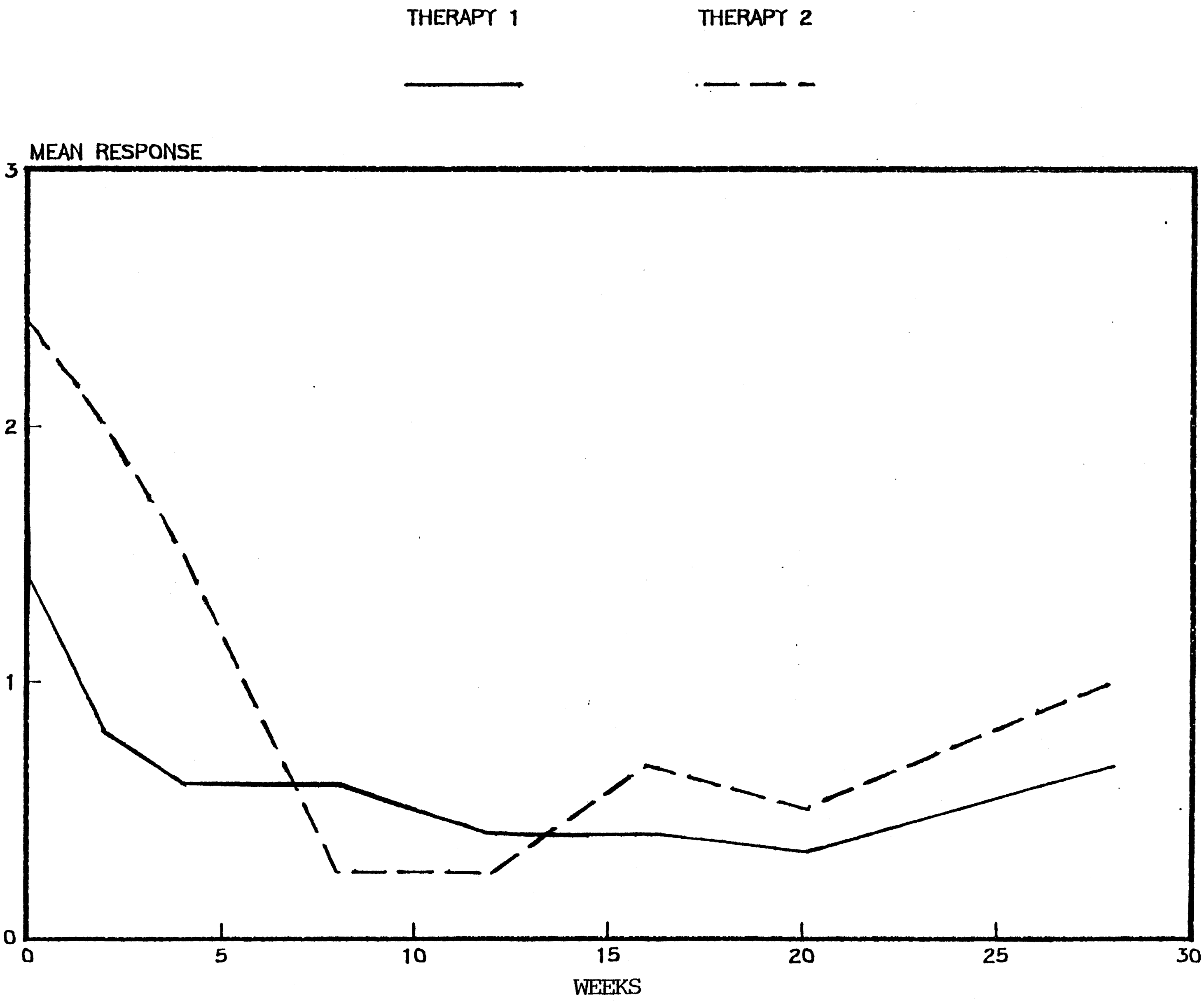


FIGURE 5: ANKYLOSING SPONDYLITIS

DAY PAIN VERSUS TIME



SIG. DIFF. BETWEEN GROUPS AT BASELINE

FIGURE 6 : ANKYLOSING SPONDYLITIS

DAY PAIN VERSUS TIME

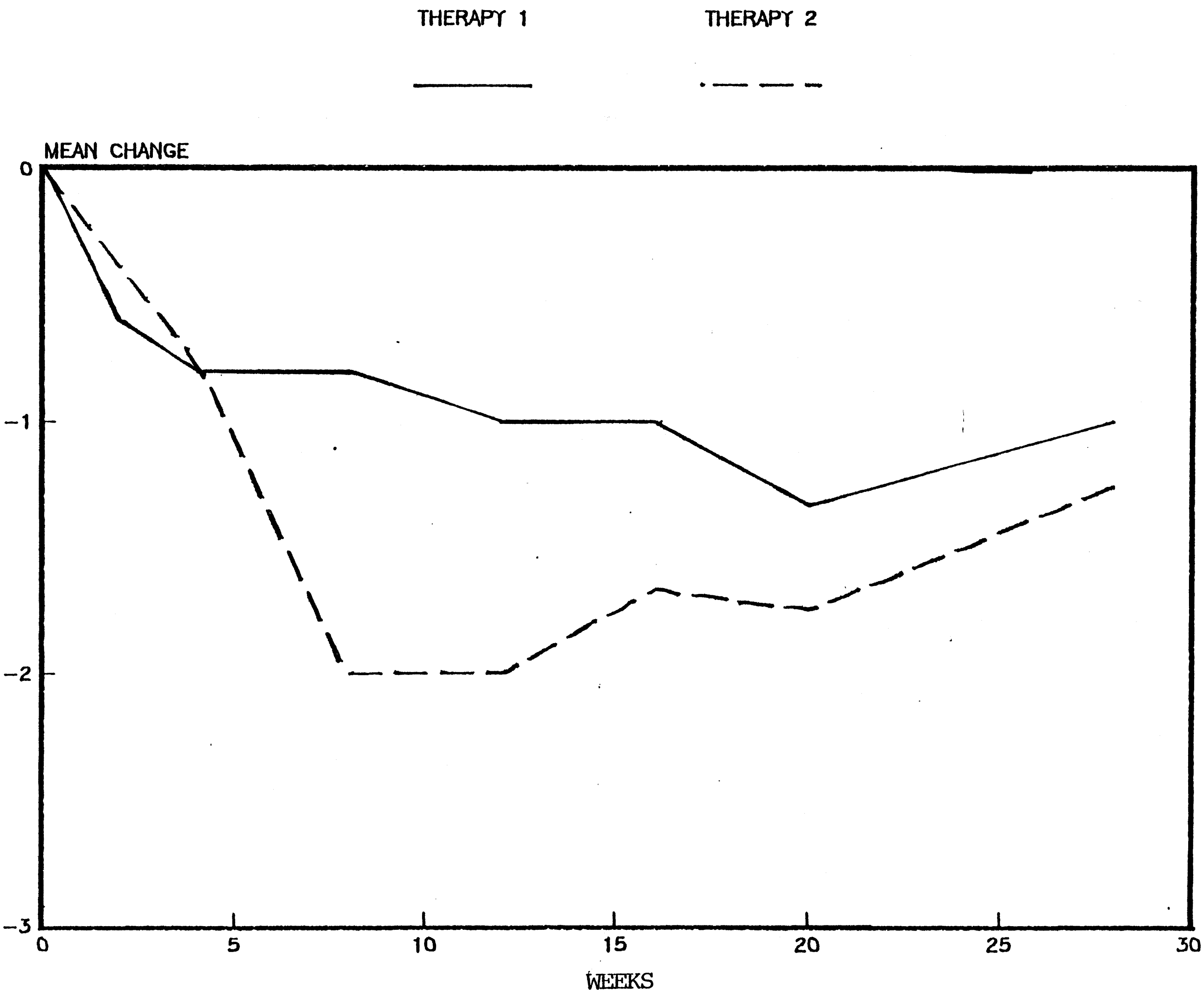


FIGURE 7: ANKYLOSING SPONDYLITIS STUDY

NIGHT PAIN VERSUS TIME

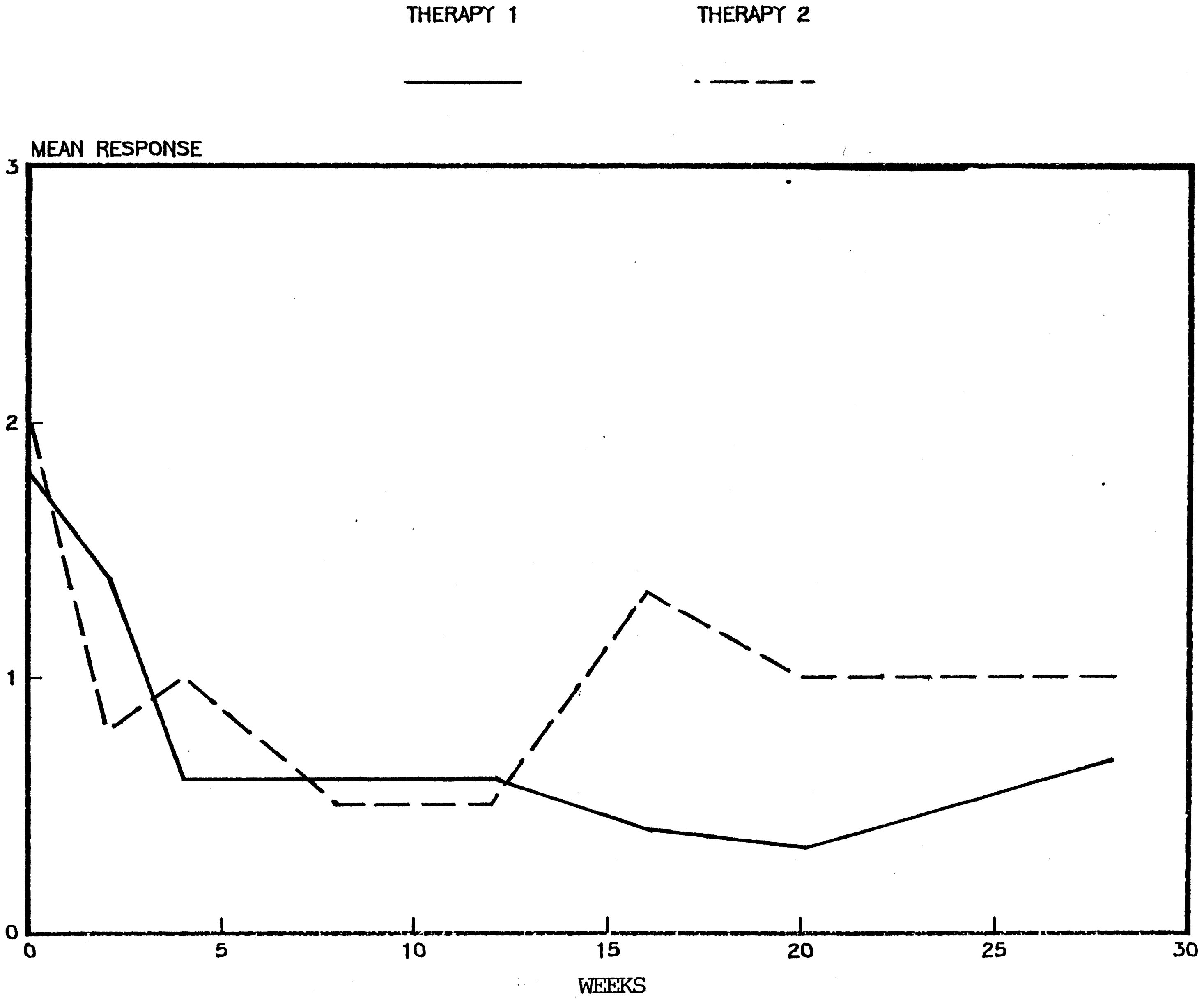


FIGURE 8 : ANKYLOSING SPONDYLITIS STUDY

NIGHT PAIN VERSUS TIME

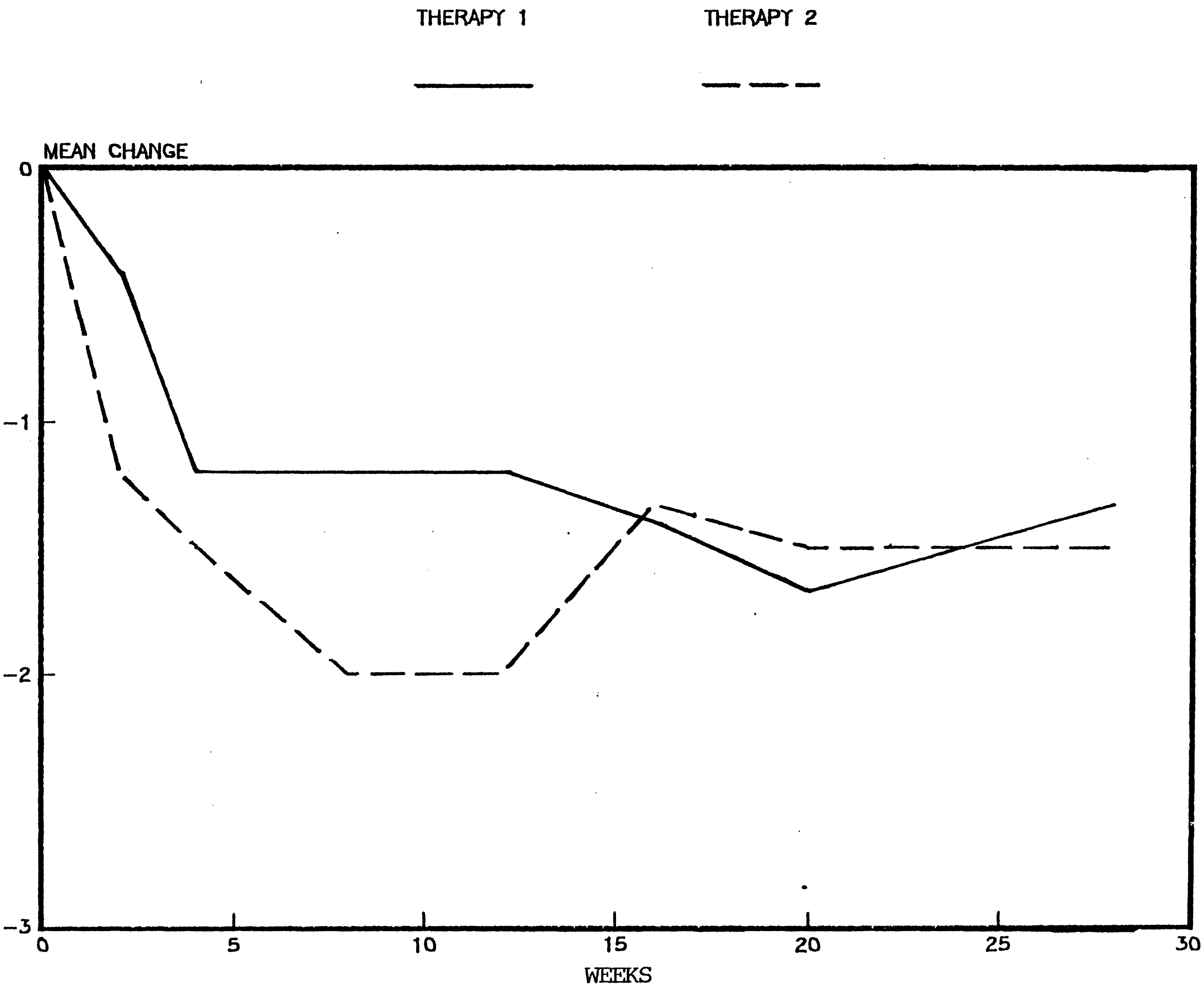


FIGURE 9: ANKYLOSING SPONDYLITIS STUDY

ERYTHROCYTE SEDIMENTATION RATE VS. TIME

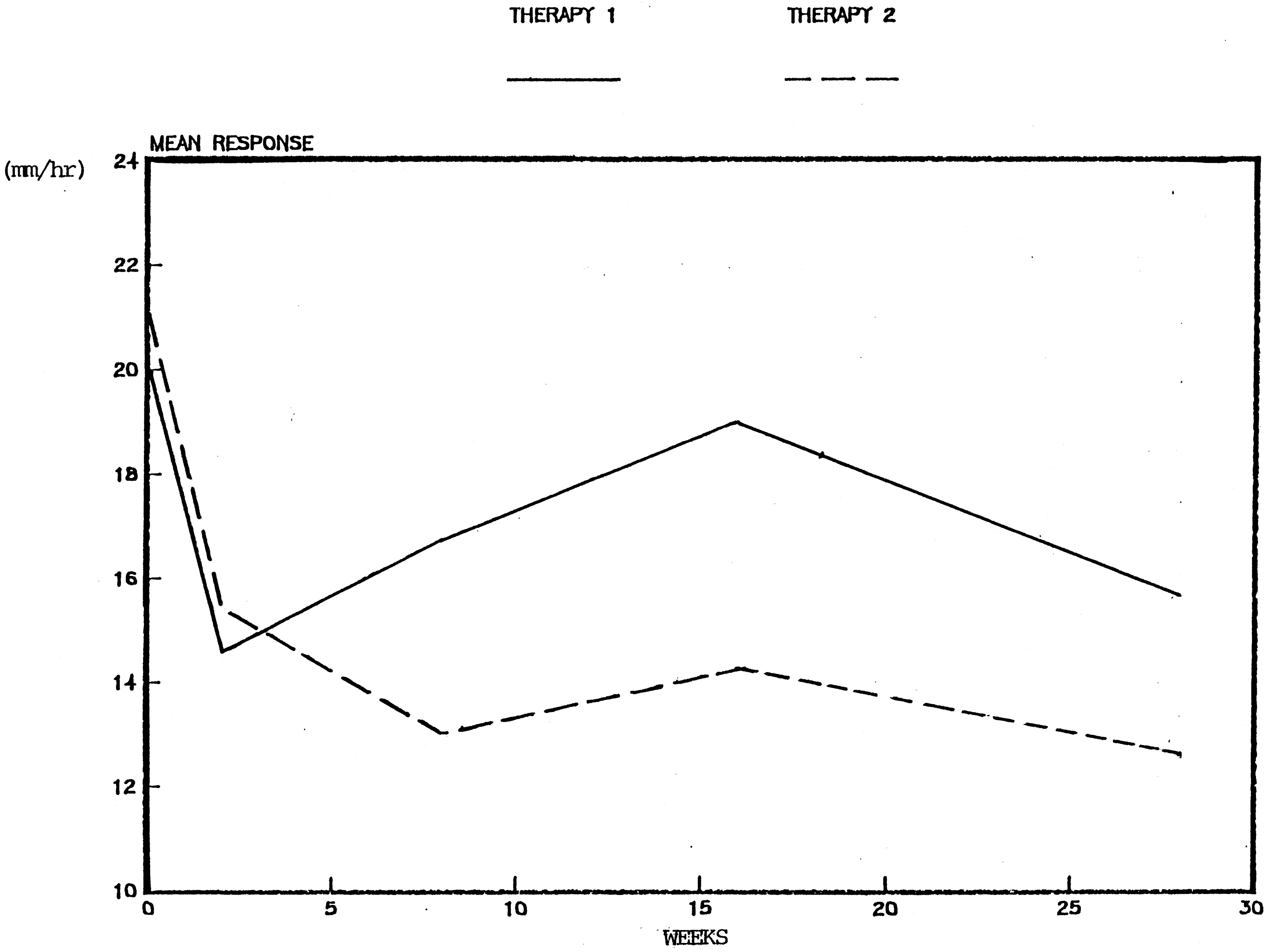


FIGURE 10: ANKYLOSING SPONDYLITIS STUDY

ERYTHROCYTE SEDIMENTATION RATE VS. TIME

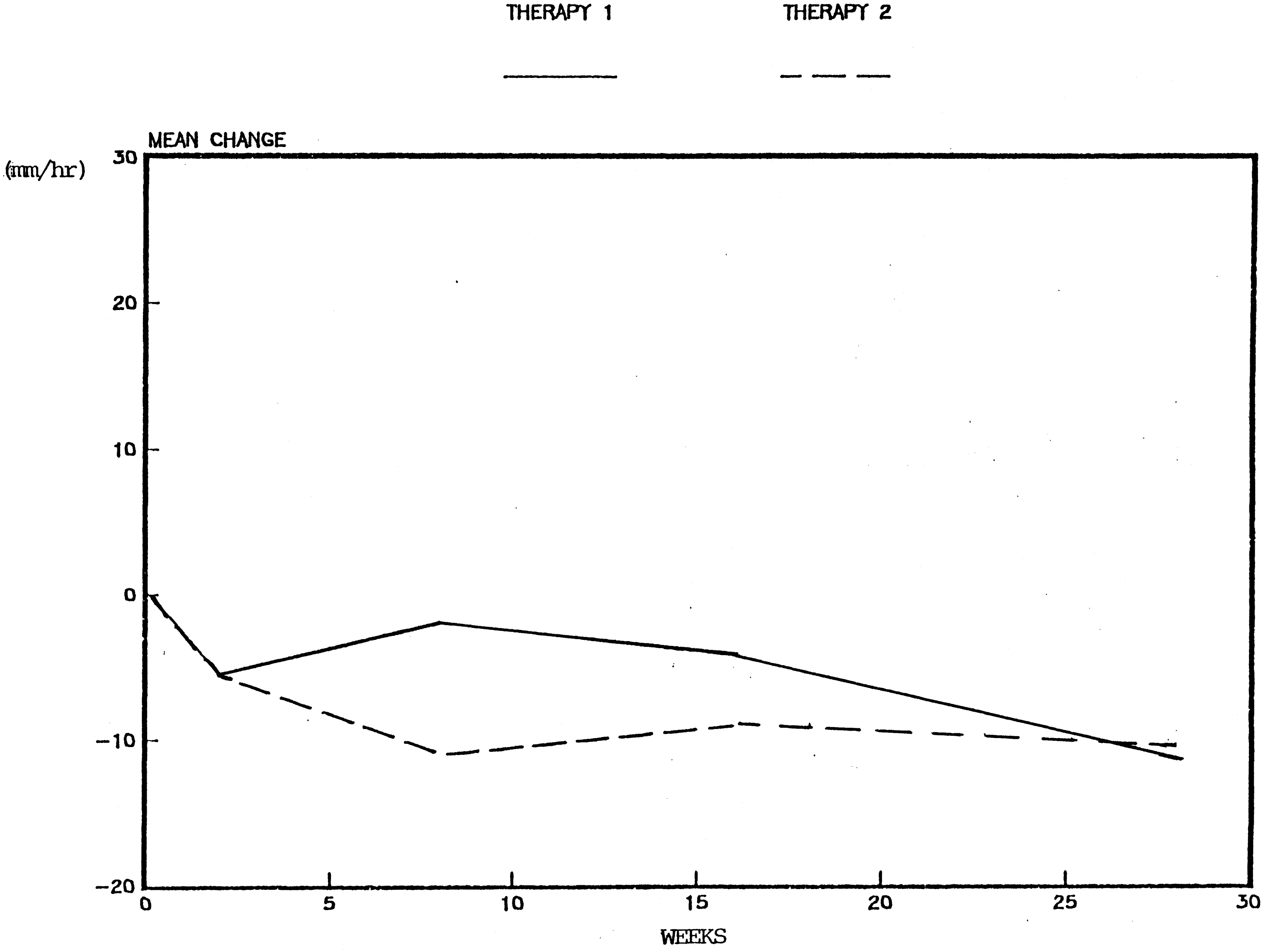
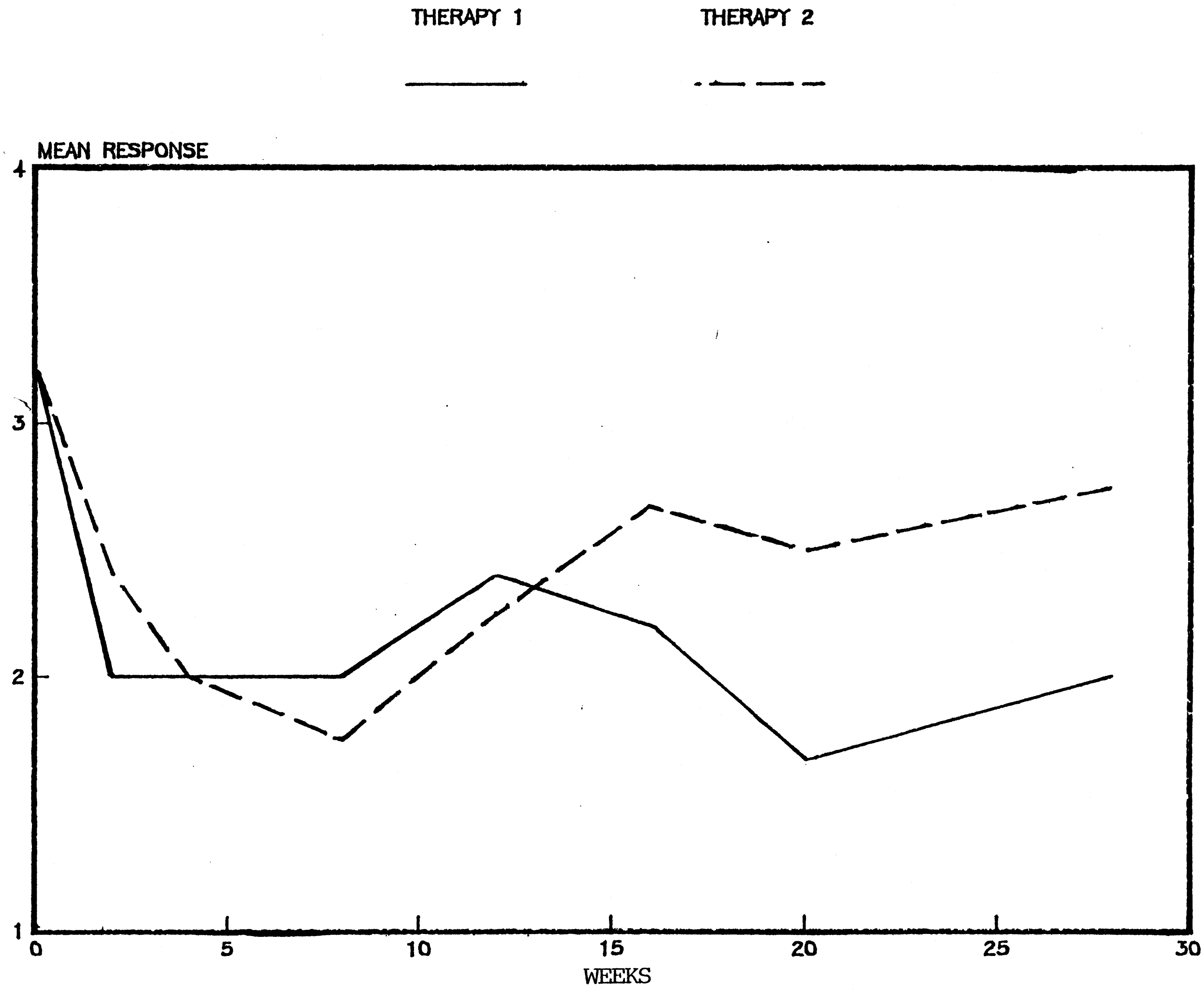


FIGURE 11: ANKYLOSING SPONDYLITIS STUDY

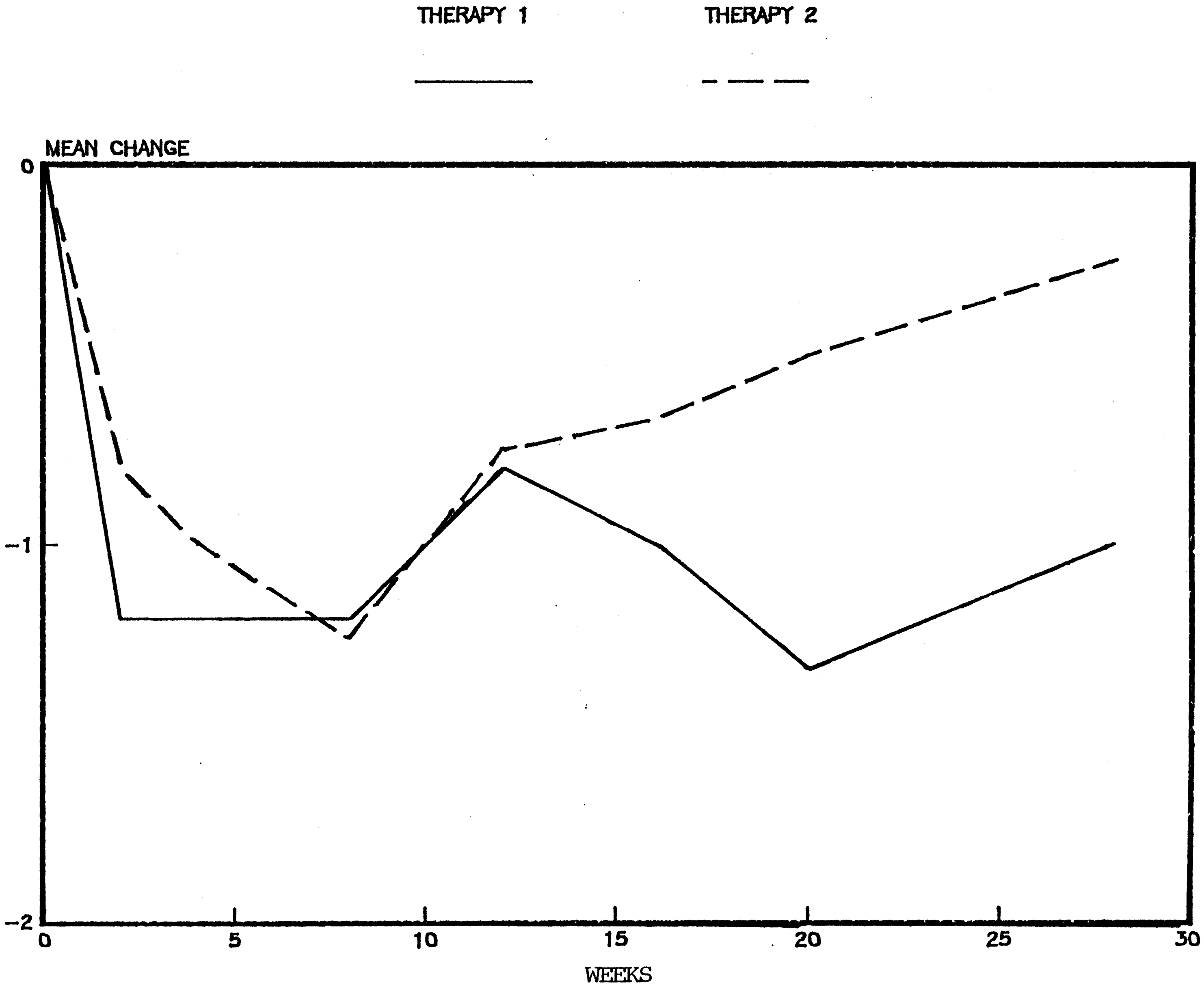
ACTIVITY IMPAIRMENT VS. TIME(2' EFF. PARA.) *



*- refers to Secondary Efficacy Parameter

FIGURE 12: ANKYLOSING SPONDYLITIS STUDY

ACTIVITY IMPAIRMENT VS. TIME(2' EFF. PARA.)*



*- refers to Secondary Efficacy Parameter

FIGURE 13: ANKYLOSING SPONDYLITIS STUDY

INTERMALLEOR STRADDLE DISTANCE VS. TIME

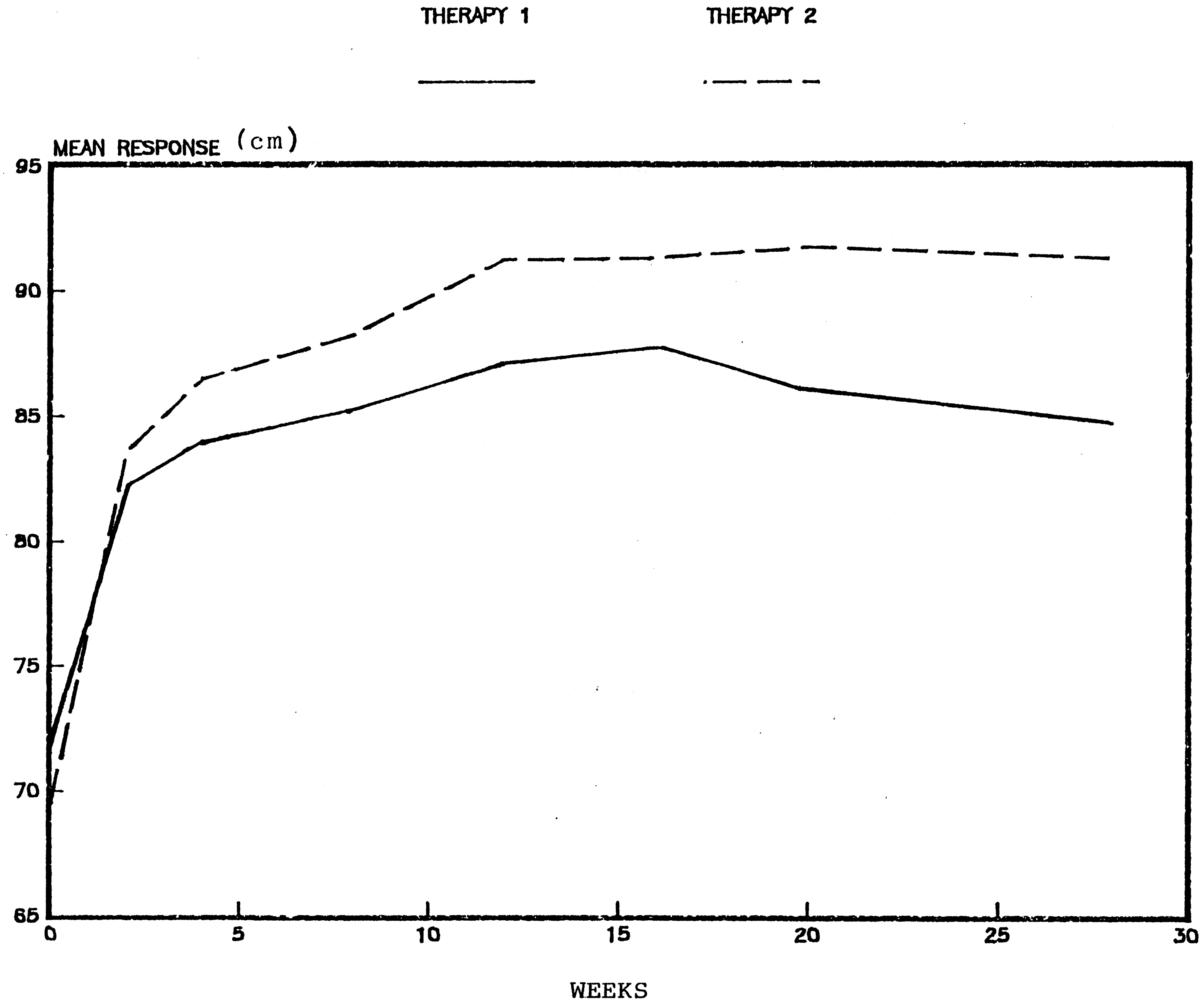
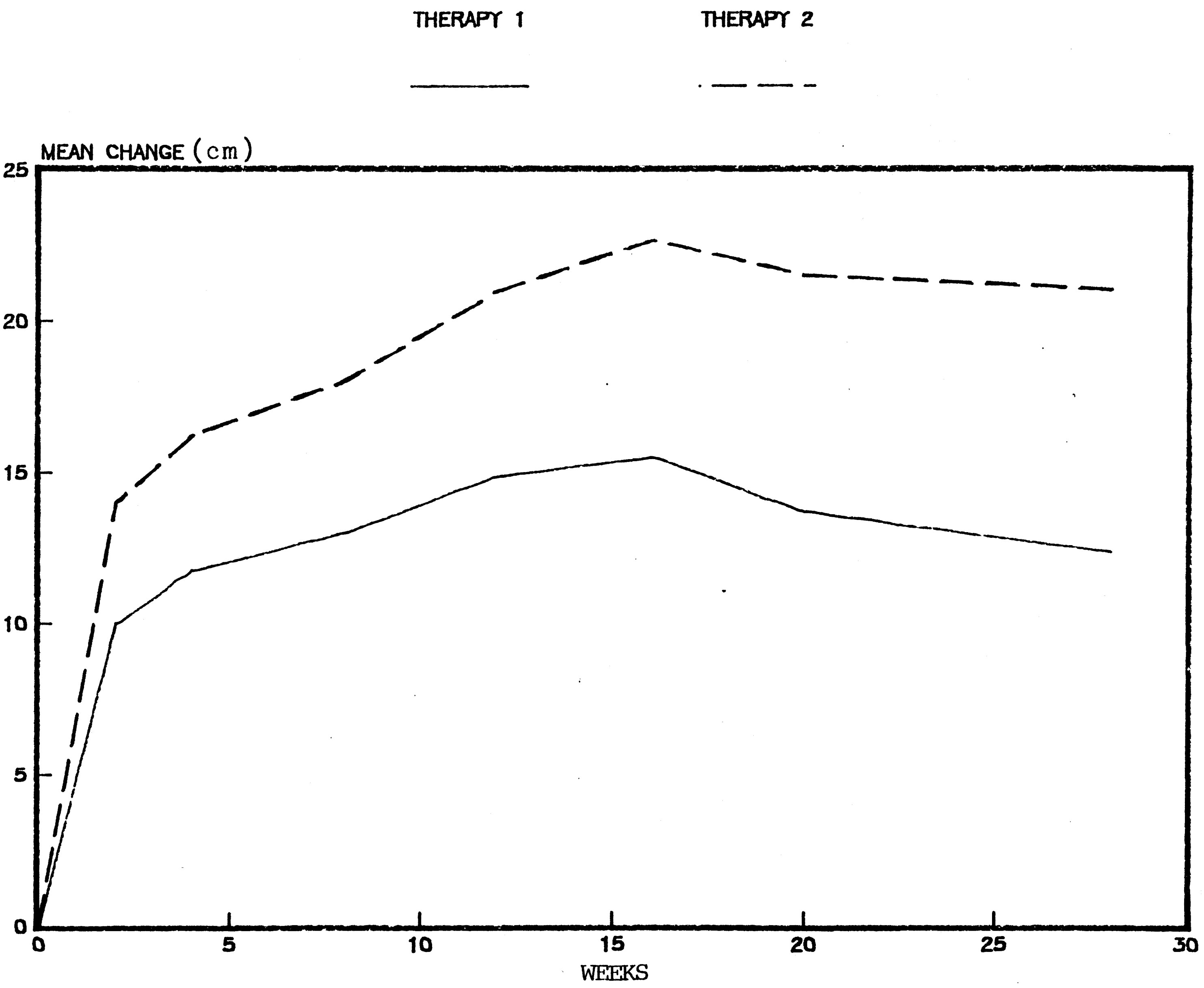


FIGURE 14: ANKYLOSING SPONDYLITIS STUDY

INTERMALLEOR STRADDLE DISTANCE VS. TIME



SIG.DIFF @2,4,8,12 WKS(#1):4,8,12,20,28WKS(2)

FIGURE 15: ANKYLOSING SPONDYLITIS STUDY

SYSTOLIC BLOOD PRESSURE VS. TIME

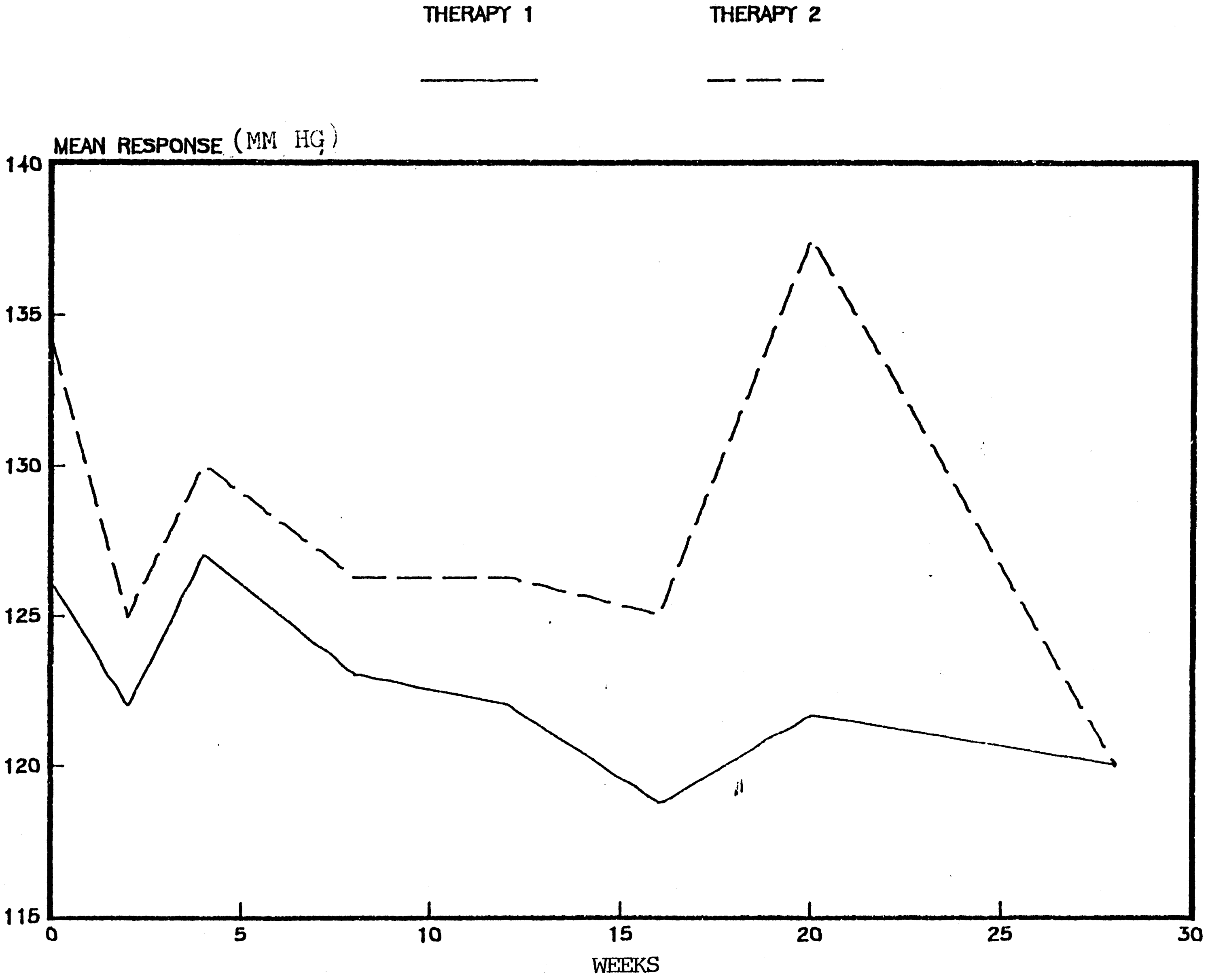


FIGURE 16: ANKYLOSING SPONDYLITIS STUDY

DIASTOLIC BLOOD PRESSURE VS. TIME

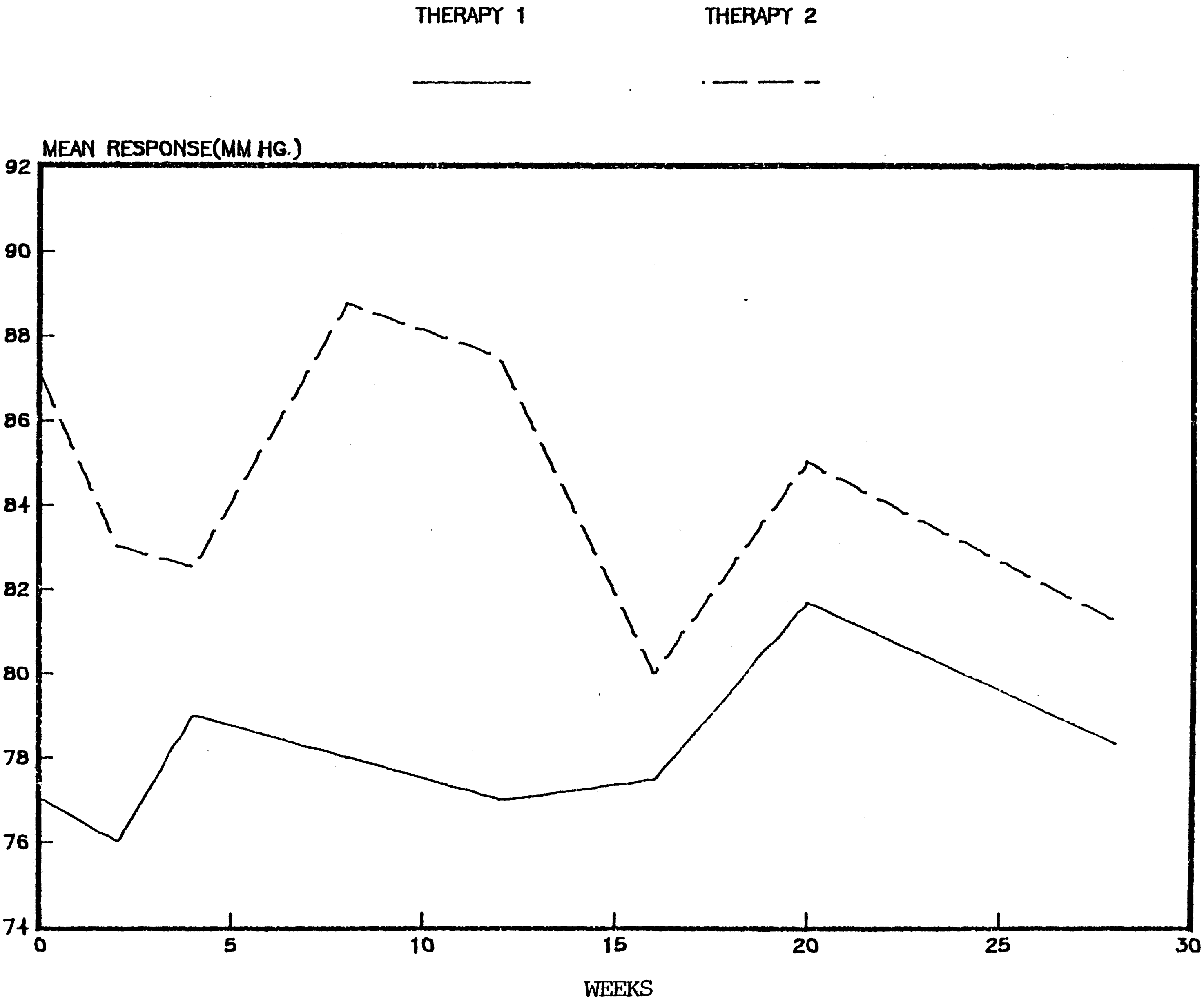
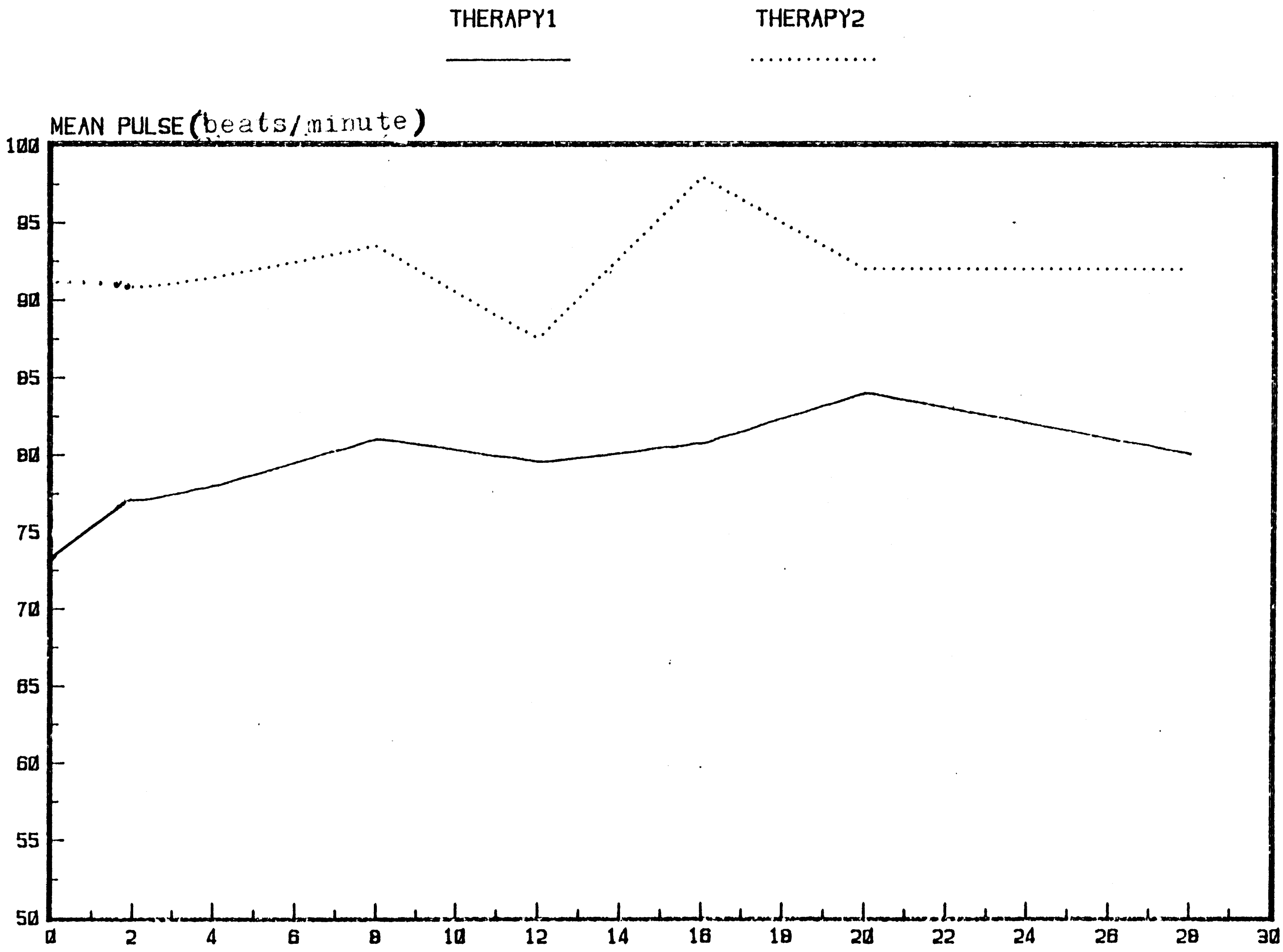


FIGURE17 ANKYLOSING SPONDYLITIS STUDY

PULSE VS. TIME



SIGNIFICANT (P<Ø.Ø5) DIFFERENCE BETWEEN GROUPS AT 2, WITHIN GROUPS AT 4, 16 WKS

FIGURE18. ANKYLOSING SPONDYLITIS STUDY

PULSE VS. TIME

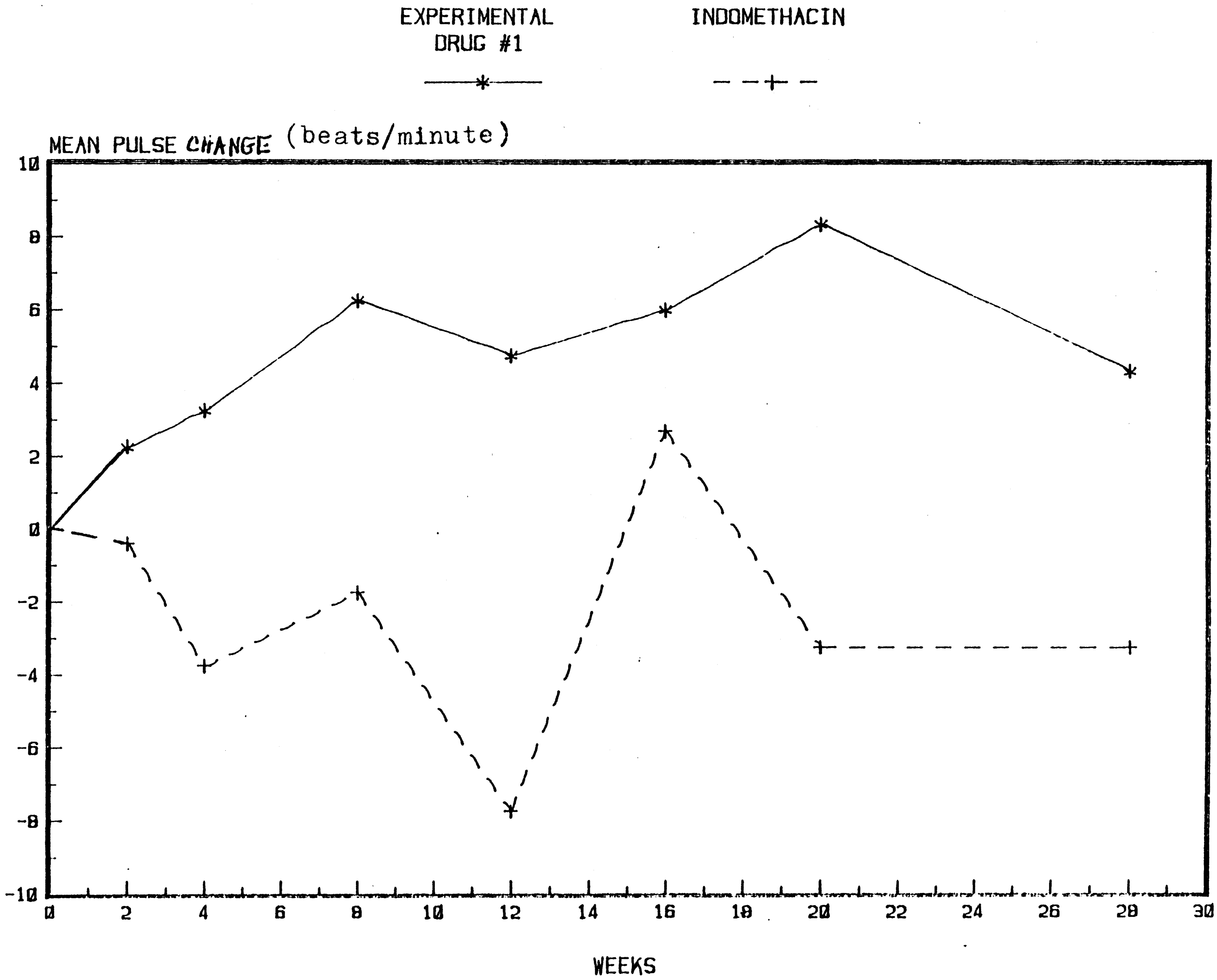


FIGURE 19: BOOTSTRAP ILLUSTRATION

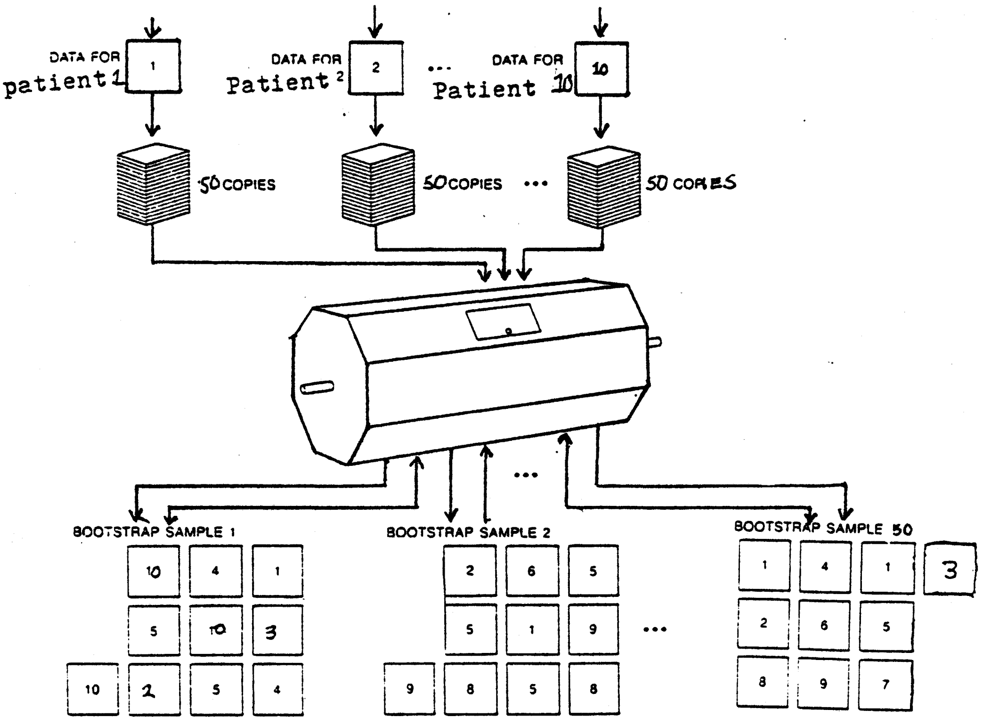
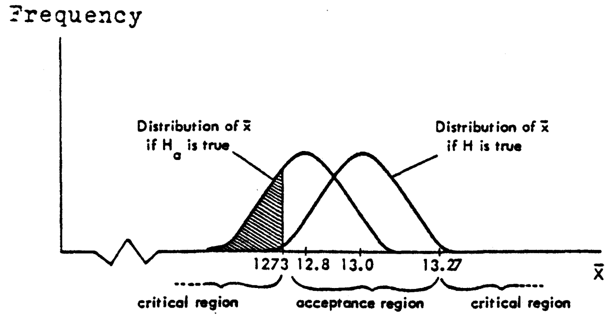
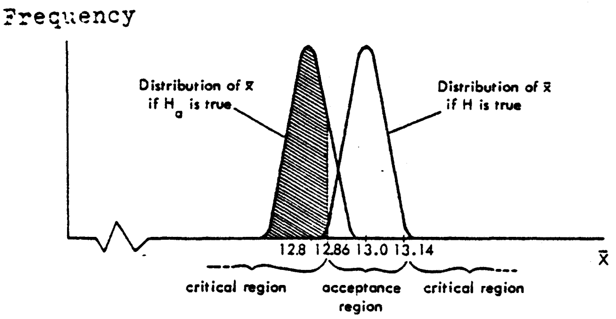


FIGURE 20: POWER OF A TEST WITH NORMAL DISTRIBUTION



Power of the test of $H:\mu = 13.0$ against $H_a:\mu = 12.8$ when $\sigma = 0.7$, $\alpha = 0.05$, $n = 25$.



Power of the test of $H:\mu = 13.0$ against $H_a:\mu = 12.8$ when $\sigma = 0.7$, $\alpha = 0.05$, $n = 100$.

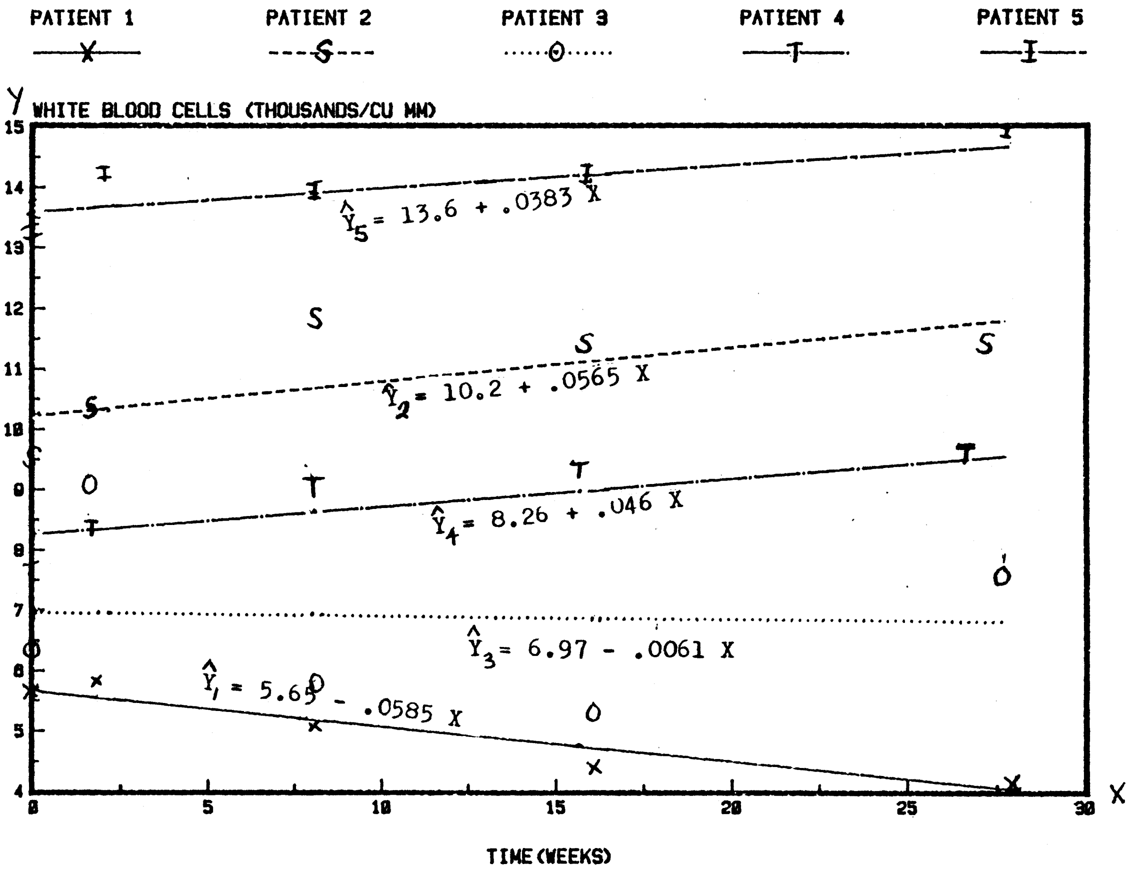
(taken from Remington and Schork, 1970)

GRAPHS

APPENDIX B

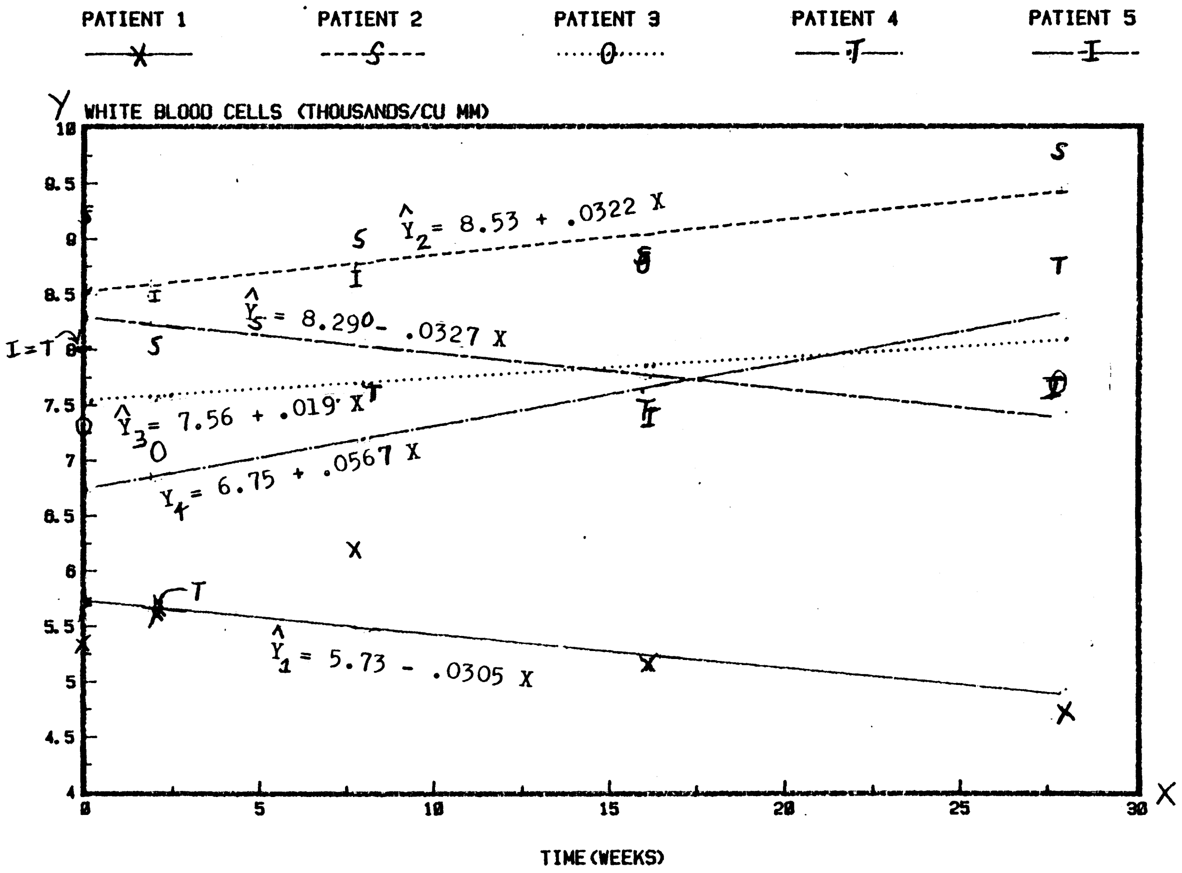
GRAPH 1

WHITE BLOOD CELLS VS. WEEKS - INDOCID GROUP



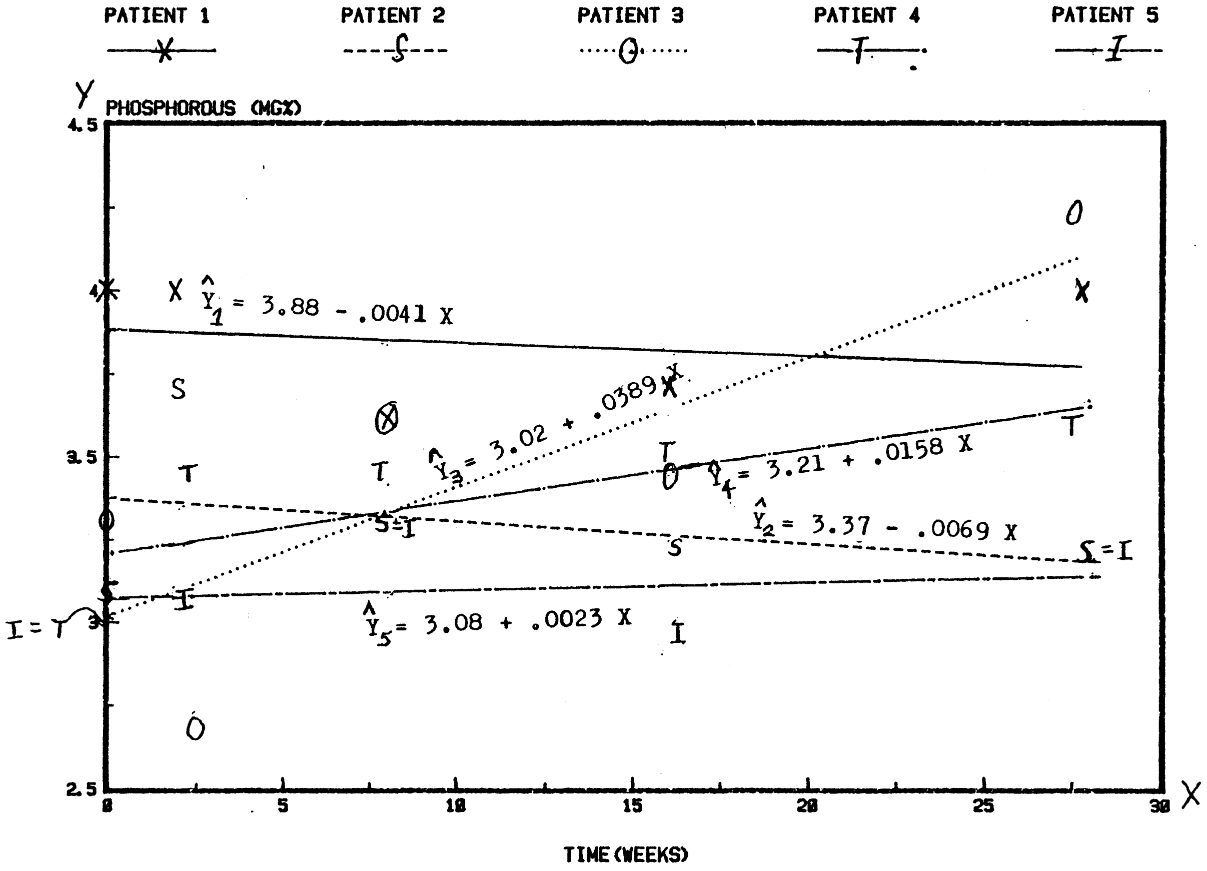
GRAPH 2

WHITE BLOOD CELLS VS. WEEKS - EXPT. DRUG #1



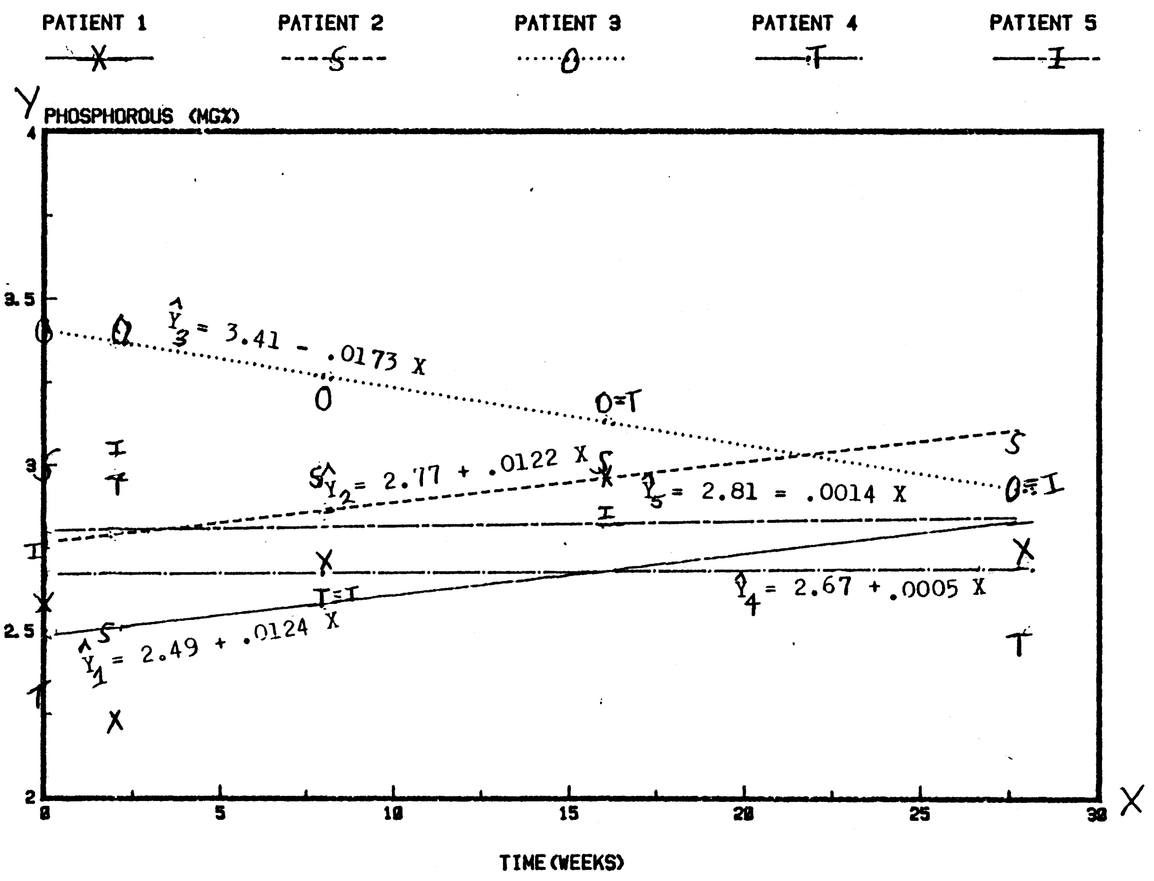
GRAPH 3

PHOSPHOROUS VERSUS WEEKS-INDOMETHACIN GROUP



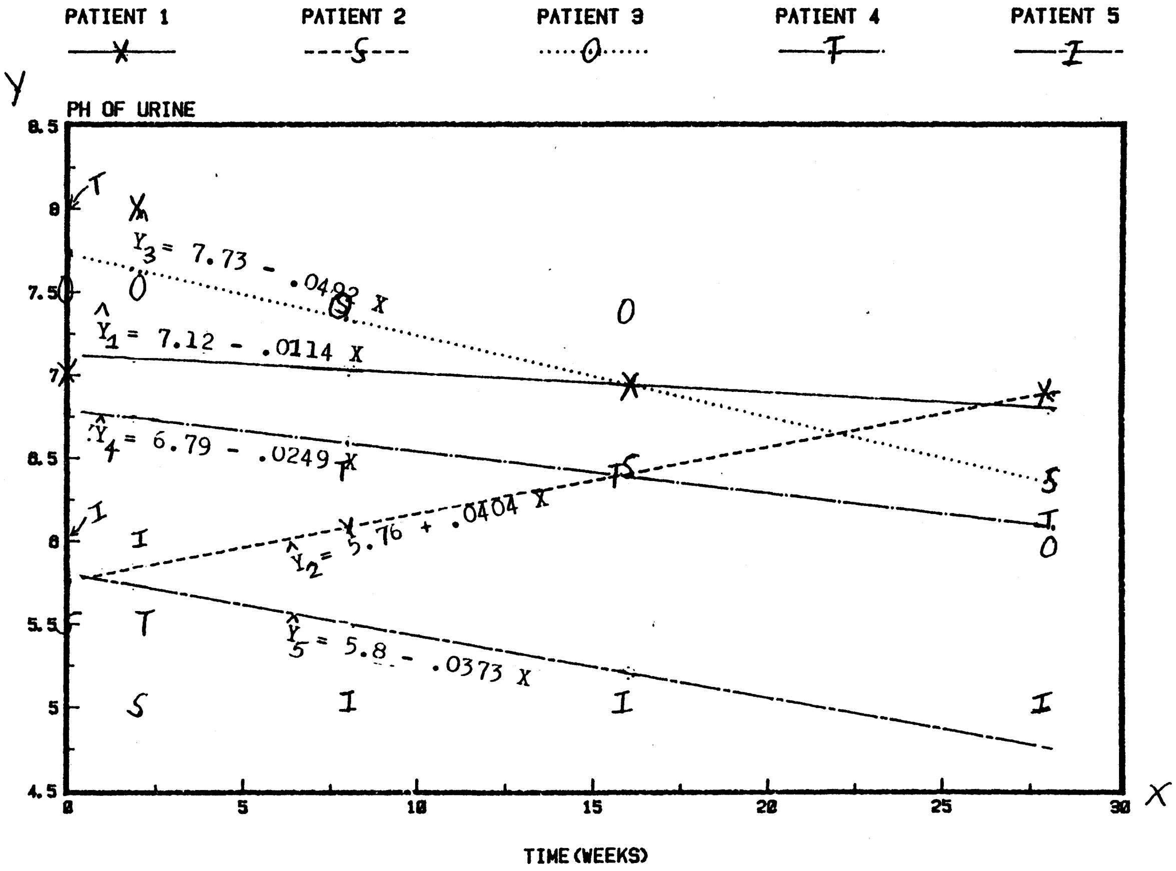
GRAPH 4

PHOSPHOROUS VERSUS WEEKS-EXPT. DRUG #1 GROUP



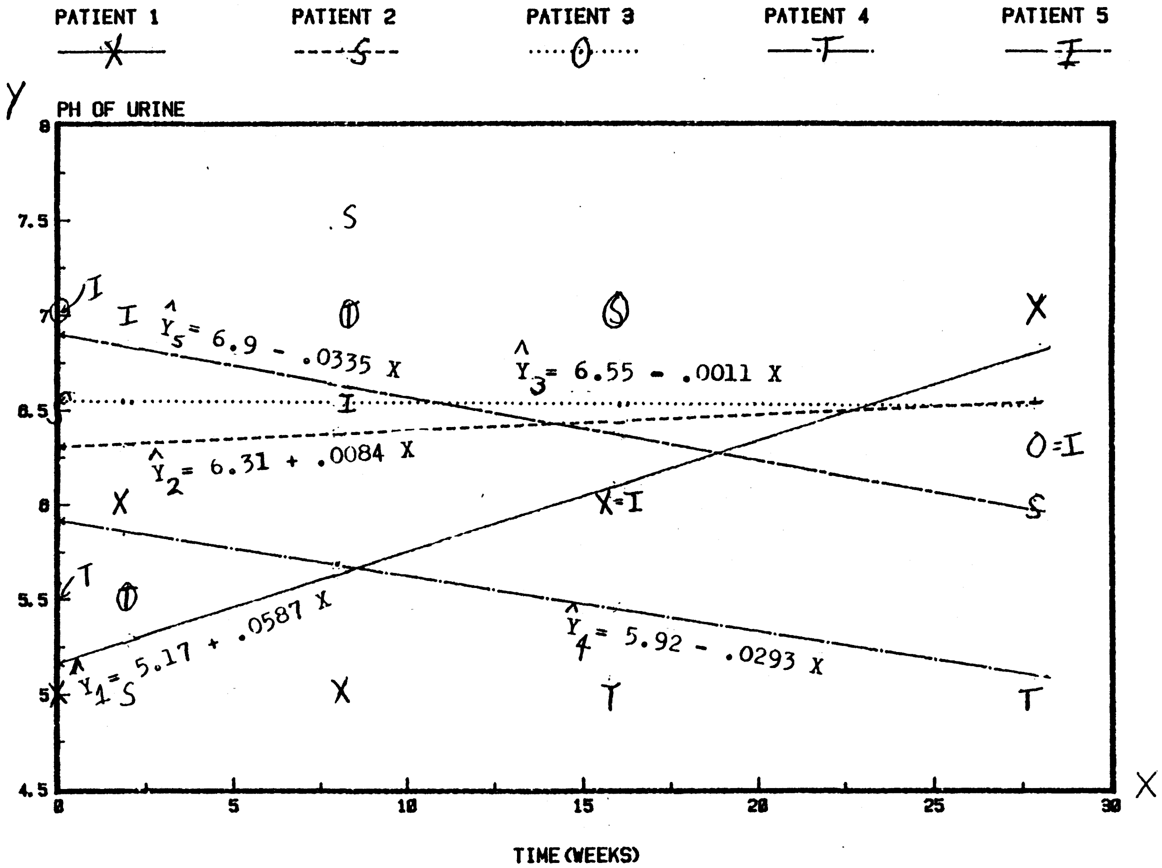
GRAPH 5

PH OF URINE VERSUS WEEKS - INDOCID GROUP

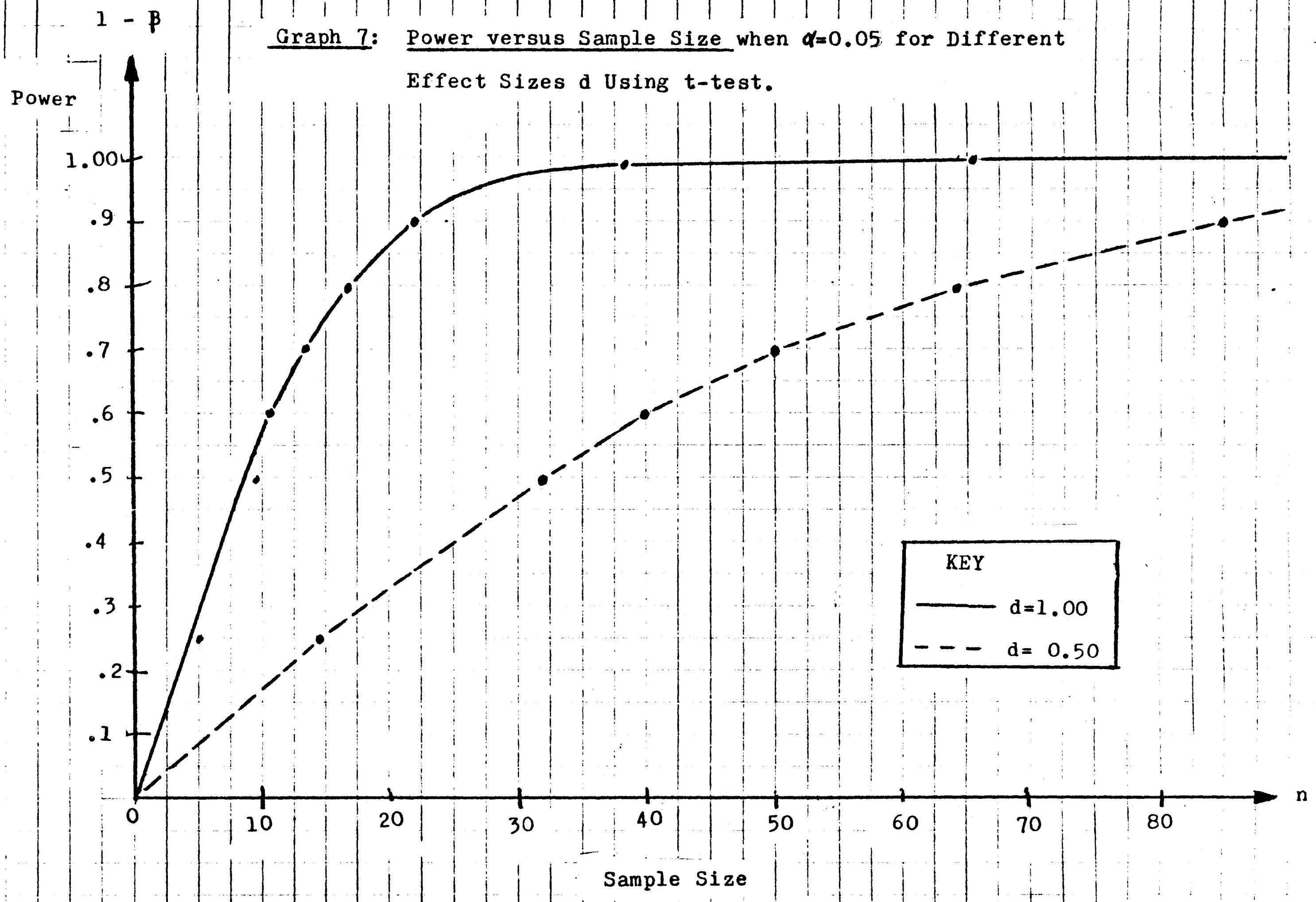


GRAPH 6

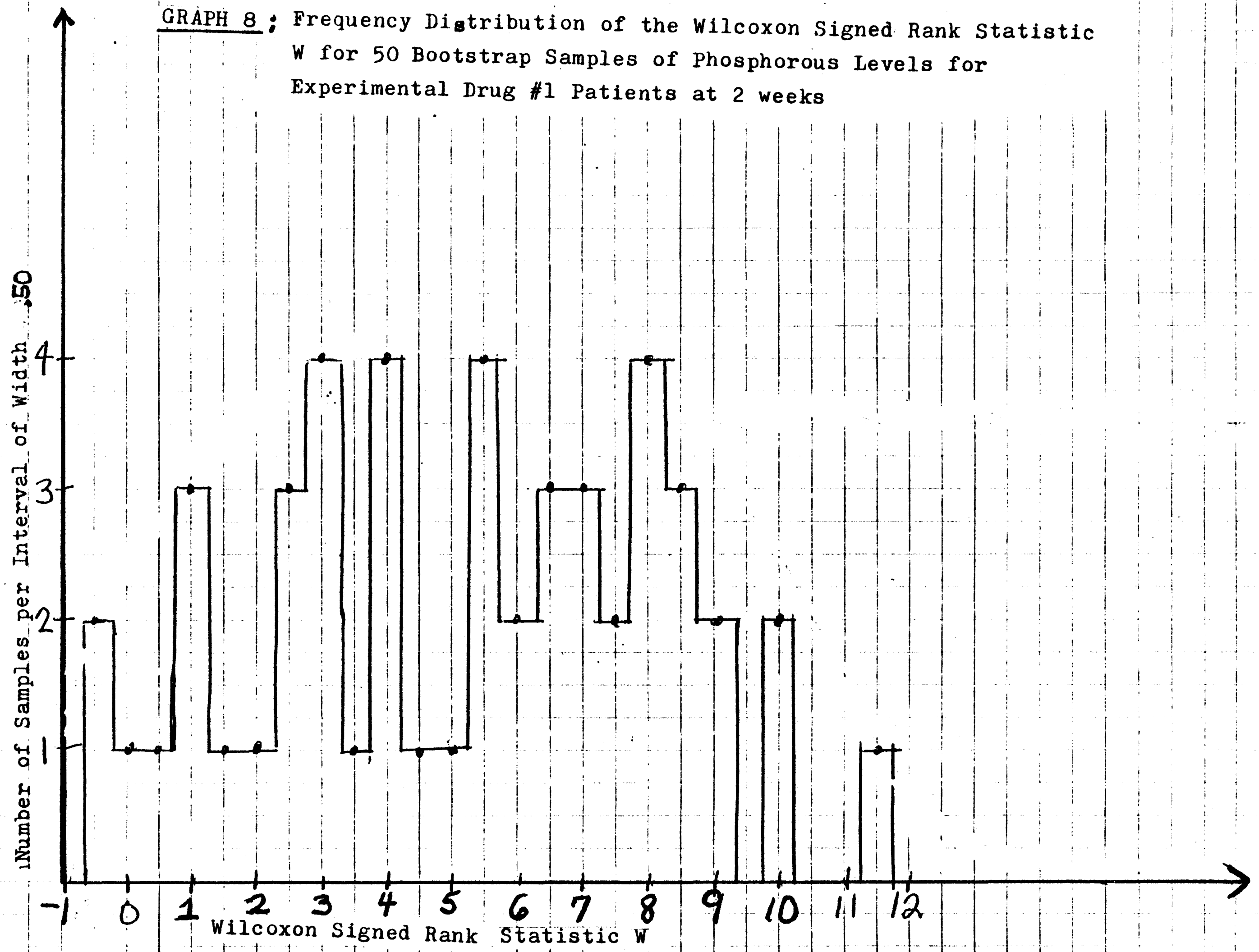
PH OF URINE VS. WEEKS - EXPT. DRUG #1 GROUP



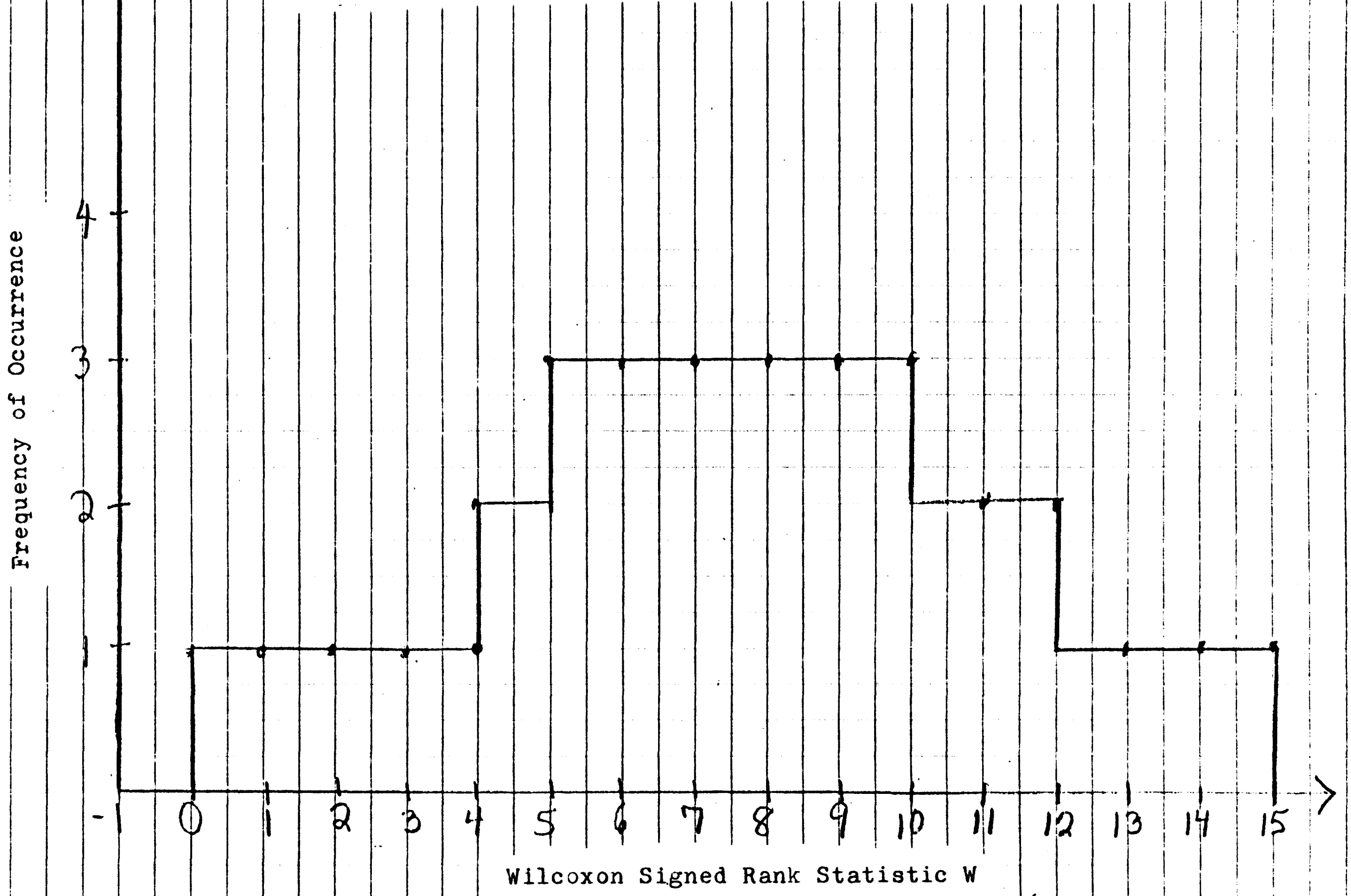
Graph 7: Power versus Sample Size when α=0.05 for Different Effect Sizes d Using t-test.



GRAPH 8: Frequency Distribution of the Wilcoxon Signed Rank Statistic W for 50 Bootstrap Samples of Phosphorous Levels for Experimental Drug #1 Patients at 2 weeks



GRAPH 9: Expected Frequency Distribution of the Wilcoxon Signed Rank Statistic W based on a Sample Size of 5



APPENDIX C

PAIRED T-TEST, σ UNKNOWN

The paired t-test tests the null hypothesis:

H_o: $\mu_d = 0$ versus the alternate H_a: $\mu_d \neq 0$

where μ_d is the mean of the population of differences; that is $\mu_d = \mu_1 - \mu_2$ and μ_1 is the mean of population 1 and μ_2 is the mean of population 2. The assumptions for this test are:

(1) The population of differences are normally distributed, and

(2) The number of paired differences 'n' are a random sample.

The resulting test statistic is:

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

where $d = x_1 - x_2$, the x's being the individual values of the variable in question, s_d is the sample standard deviation of the observable differences and n is the number of pairs. The distribution under H_o of the test statistic is a Student's t distribution with n-1 degrees of freedom. The two-sided critical region at level of significance α is:

$t \leq t_{\alpha/2}$ or $t \geq t_{1-(\alpha/2)}$.

ONE-WAY ANALYSIS OF VARIANCE

The One-Way analysis of variance (ANOVA) tests for any differences among the population means; specifically, the null hypothesis is:

H_o: $\mu_1 = \mu_2$ versus H_a: $\mu_1 \neq \mu_2$

where μ_1 is the first population mean and μ_2 is the second population mean. The assumptions for this test are:

(1) Each sample is a random sample from the corresponding population, and observations from different populations are independent.

(2) The measurement variable 'x' is normally distributed in each of the populations.

(3) The populations all have the same variance (homoscedasticity).

The above is an upper two-sided test and the resulting test statistic follows the F distribution with k-1 degrees of freedom in the numerator and n-k denominator degrees of freedom, where k = number of populations sampled and n = the total number of observations. The upper two-sided critical region at level of significance α is: $F \geq F_{(1-\alpha)}$

$$\text{where } F = \frac{\sum_{i=1}^{k} n_i(\bar{x}_{i.} - \bar{x})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i.})^2/(n-k)} .$$