CENTER FOR DRUG EVALUATION AND RESEARCH

APPLICATION NUMBER: NDA 20905

STATISTICAL REVIEW(S)

Statistical Review and Evaluation

NDA 20-905
Name of Drug: ARAVA (leflunomide 20mg)
Applicant: Hoechst Marion Roussel
Documents Reviewed: Statistical Section (Vol.222-Vol.331, Sponsor's additional submissions on 4/20/98, 5/4/98, 5/22/98, 6/16/98, 6/17/98, 6/19/98, 6/23/98, 6/24/98 and 7/1/98)
Statistical Reviewer: Laura Lu, Ph.D.
Date of Review: 7/10/98

I. Introduction

ARAVA (leflunomide) is intended for use in the treatment of active rheumatoid arthritis. NDA20-905 has been submitted for approval of leflunomide with the following claims:
- reduction in pain, articular swelling and tenderness and amelioration of the signs and symptoms of RA;
- retardation of x-ray evidence of progression of disease and prevention of new bone erosions; and
- improvement in functional ability and health-related quality of life.

Four phase III studies were conducted. Study US301 (n=482) was conducted in the US and Canada to compare the efficacy and safety of leflunomide to that of placebo and methotrexate over 52 weeks of therapy. Study MN301 (n=358) compared leflunomide to placebo and sulfasalazine over 24 weeks. Study MN302 (n=999) compared leflunomide to methotrexate over 52 weeks. Study MN303 (n=197) was an extension of MN301 for an additional 24 weeks, allowing placebo-treated subjects from MN301 to receive sulfasalazine. Studies MN301-MN303 were conducted as multinational studies in various European countries, South Africa, Australia and New Zealand. This review will focus on Study US301, Study MN301 and Study MN302.

II. Study US301

1. Protocol

Study Type: This was a U.S. multi-center, phase III, randomized, double-blind, double dummy, placebo- and methotrexate-controlled study in subjects with active RA who had never previously received methotrexate.

Primary Objective: Comparing the efficacy and safety of leflunomide (LEF) with placebo in subjects with active RA who had never previously received methotrexate (MTX).

Secondary Objectives: Comparing the efficacy and safety of LEF with MTX in subjects with active RA.

Primary Efficacy Variables: ACR success rate, defined as the percentage of subjects who completes 52 weeks of treatment and meets the ACR responder criteria at that time. An ACR responder is defined as a subject with 20% improvement in both swollen and tender joint counts, and 20% improvement in 3 of the following 5 measures: patient global assessment, physician global assessment, modified Health Assessment Questionnaire (MHAQ) or standard Health Assessment Questionnaire (HAQ), pain intensity, and either erythrocyte sedimentation rate (ESR) or C-reactive protein (CRP).

Secondary Efficacy Variables: ACR responder rate, tender joint count, swollen joint count, patient global assessment, physician global assessment, MHAQ, pain intensity assessment, ESR, CRP, morning stiffness, area under the curve (AUC) of ACR response, x-rays of hands and feet, questionnaires on quality of life and functional ability (Medical Outcomes Study (MOS) SF-36, MOS current health perceptions scale, Work Limitations Questionnaire, standard HAQ, and Problem Elicitation Technique (PET)).

Safety Variables: Adverse events; hematology, blood chemistry, urinalysis; physical examination, supine blood pressure, heart rate, oral body temperature, body weight, 12-lead ECG, chest x-ray.

Study Duration: The study was consisted of the following parts:
- a screening visit
- an initial therapy phase of 52 weeks
- an additional initial therapy phase of 52 weeks
- an alternate therapy phase.

Study Visits: Subjects were evaluated for signs and symptoms of RA biweekly from week 4 to week 12 and monthly thereafter. X-rays of the hands and feet were taken at baseline and week 52. If the subject discontinued treatment early, x-rays would also be taken at that time. Function, disability and health-related quality of life measures were assessed at baseline, week 24, week 52 or at time of early discontinuation.

Sample Size: Assuming the ACR response rate in the placebo group was 30.4%, the recruitment of 125 placebo subjects, 187 LEF subjects and 187 methotrexate subjects was to provide at least 90% power to detect a difference of 20% between treatment and placebo in ACR success rate at α=0.05, and 80% power to conclude the equivalence of LEF and MTX (assuming the difference between the methotrexate and leflunomide response rate was no more than 5% and the difference between methotrexate and placebo was approximately 25%) when the 95% confidence interval for the difference in ACR success rates overlaps zero and lay fully to the right of –15%.

Randomization: Subjects were assigned to receive LEF, MTX, or placebo in a 3:3:2 ratio by adaptive randomization (stratified according to time since last treatment with a DMARD: 8 weeks or > 8 weeks) for 52 to 104 weeks, using a double-dummy design, (i.e. all subjects received matching placebo tablets). Subjects who remained in the study for a minimum of 16 weeks but were not responders by ACR criteria or who experienced significant toxicity or persistent laboratory abnormalities were allowed to discontinue initial therapy and, after appropriate washout, enter alternate therapy on a blinded basis. (Data from alternate therapy was not included in this submission as all patients had not completed 12 months of therapy.) Blinded allocation to alternate therapy was also assigned at enrollment: if not effectively treated with initial therapy, LEF subjects would be switched to MTX therapy; MTX and placebo subjects would be switched to LEF therapy.

Statistical Methods: The statistical analyses were based on the data for the first 52 weeks of the initial therapy phase. In addition to descriptive statistics, the following tests were performed:

Baseline Comparison

Categorical variables: chi-square analysis

Continuous variables: analysis of variance (ANOVA)

Efficacy Comparison

Primary: Success rates were compared in the 2 groups using logistic regression analysis, including factors: treatment, region, duration of RA (≤2 years, > 2 years), time since last DMARD dose (< 8 weeks, ≥8 weeks), treatment × region, treatment × duration of RA, and treatment × time since last DMARD dose.

Secondary: ACR responder rates were compared by logistic regression using last observation carried forward (LOCF). All other secondary efficacy variables were analyzed by ANCOVA using LOCF.

2. RESULTS

This report describes and presents only the results for the first 52 weeks of the initial therapy phase.

Patient Disposition and Demographics

A total of 485 subjects were randomly assigned to one of the three treatment groups: LEF, placebo, or MTX; three (3) subjects did not receive study medication. Of the remaining 482 randomized subjects, 2 subjects did not have baseline scores, 480 patients (182 LEF, 118 placebo, and 180 MTX) were included in the intent-to-treat analysis. The following table summarizes the drop-out status.

Summary of Reasons for Early Withdrawal from the Treatment Phase

Reason for Withdrawal	Leflunomide (N=182)		Placebo (N=118)		Methotrexate (N=182)		Total (N=482)
	N	%	N	%	N	%	
Lack of Efficacy	31	17	62	53	44	24	137
Adverse Event	40	22	10	9	18	10	68
Lost to follow-up	1	1	0	-	2	1	3
Protocol violation	0	-	1	1	1	1	2
Noncompliance	1	1	0	-	1	1	2
Death	0	-	0	-	1	1	1
Other	13	7	8	7	10	5	31
Total	86	47	81	69	77	42	244

The distribution of demographic characteristics, RA disease history and concomitant RA medication use in the three treatment groups are presented in Table 1-3 in Appendix A.

Efficacy

The following efficacy results are from the intend-to-treat analyses.

- ACR success and responder

The table below shows that LEF is superior to placebo and equivalent to MTX in terms of ACR success rate.

Treatment Group	No. of Successful Subjects /Total N	ACR Success Rate	p-value
Leflunomide	74/182	40.7%	LEF - PBO: $p \leq 0.0001$
Placebo	22/118	18.6%	
Methotrexate	63/180	35.0%	LEF - MTX: 95% CI: -4.3 to 15.6

The results over time for ACR responder rate are plotted in Figure 1 in Appendix B. The following table presents the ACR responder rates for the treatment groups. The results are consistent with the trend of the ACR success rate.

Treatment Group	No. of ACR Responders/ Total N	ACR Responder Rate at Endpoint	p-value
Leflunomide	93/178	52.2%	LEF vs PBO: $p \leq 0.0001$
Placebo	31/118	26.3%	
Methotrexate	82/180	45.6%	LEF vs MTX: 95% CI: -3.6 to 17.0

The results from secondary ACR responder status (area under the curve (AUC) for ACR response, time to and duration of initial response and time to and duration of sustained response) support the finding of ACR success rate and responder rate. The results are include in Table 4-6 in Appendix A.

The results for individual ACR response components (tender joint count, swollen joint count, patient global assessment, physician global assessment, MHAQ, pain intensity

assessment, ESR and CRP) also demonstrated the superiority of LEF over placebo (all p<.0001). The results are listed in the following table.

Treatment Group	N	Mean Baseline	Mean Change at Endpoint	N	Mean Baseline	Mean Change at Endpoint
TJC				SJC		
Leflunomide	179	15.5	-7.7*	179	13.7	-5.7*
Placebo	118	16.5	-3.0*	118	14.8	-2.9*
Methotrexate	180	15.8	-6.6	180	13.0	-5.4
Patient Assessment				Physician Assessment		
Leflunomide	178	5.6	-2.1*	179	6.1	-2.8*
Placebo	118	5.8	0.1*	118	6.2	-1.0*
Methotrexate	179	5.4	-1.5	179	5.9	-2.4
ESR				CRP		
Leflunomide	162	38.4	-6.26*	168	2.1	-0.62*
Placebo	111	37.2	2.56*	115	2.4	0.47*
Methotrexate	163	33.9	-6.48	172	1.9	-0.50
Pain Assessment				MHAQ		
Leflunomide	178	5.9	-2.2*	178	0.78	-0.29*,**
Placebo	118	6.4	-0.5*	118	0.87	0.07*
Methotrexate	179	5.8	-1.7	179	0.79	-0.15**

*Indicates statistically significant differences between leflunomide and placebo ($p \leq 0.0001$)
**Indicates leflunomide and the active comparator were statistically different ($P \leq 0.05$)

- Subset analyses:

When comparing the ACR success rate of LEF with that of placebo, treatment × disease interaction and treatment × DMARD use interaction were significant (p-values were .03 and .06 respectively). The results for each subgroup are reported in the following table, which show that the interactions are actually quantitative, i.e., the treatment differences were in the same direction in all subgroups.

Study and Treatment Group	Leflunomide		Placebo		Active	
	n/N	%	n/N	%	n/N	%
ACR Success by Duration of RA						
≤2 yrs	31/71	43.7	3/39	7.7	29/73	39.7
>2 yrs	43/111	38.7	19/78	24.4	34/107	31.8
ACR Success by Prior DMARD Use						
No prior DMARD use	33/81	40.7	5/47	10.6	25/79	31.6
Prior DMARD use	41/101	40.6	17/71	23.9	38/101	37.6

X-Rays of Hands and Feet

A total of 352 subjects had both baseline and follow-up x-ray results available for this evaluation, representing 73% of the 480 subjects in the intent-to-treat population. The mean change from baseline in the total Sharp score was designated as the main efficacy analysis in the retardation of progression of disease, assessed by x-ray. Baseline mean scores and changes from baseline to endpoint mean scores in the intent-to-treat

population are shown below by treatment group for the total Sharp score, as well as for the erosion subscore and joint space narrowing subscore.

Treatment Group	N	Baseline Score	Change at Endpoint
Sharp Total Score			
Leflunomide	131	23.11	0.53*,**
Placebo	83	25.37	2.16*
Methotrexate	138	22.76	0.88**
Sharp Erosion Score			
Leflunomide	131	8.95	0.23*
Placebo	83	9.28	0.89*
Methotrexate	138	8.05	0.47
Sharp Narrowing Score			
Leflunomide	131	14.15	0.31*
Placebo	83	16.10	1.27*
Methotrexate	138	14.71	0.41

*Indicates statistically significant differences between leflunomide and placebo ($p \leq 0.05$)
**Indicates leflunomide and the active comparator were statistically different ($p \leq 0.05$)

The table above shows that the mean x-ray disease progression of LEF group is statistically less than that of the placebo group.

Other analyses of interest include a per-subject, intent-to-treat analysis of the proportion of subjects who progressed by more than 3 points in the total, and the proportion of subjects who developed newly involved joints (i.e., joints with baseline scores of 0 and an erosion score of 1 or more at follow-up). The results are presented in the following table.

Subjects	Leflunomide (N=131)		Placebo (N=83)		Methotrexate (N=138)	
	N	%	N	%	N	%
With progression > 3 points in total Sharp score	14	11*	21	25*	14	10
With newly involved joints	17	13	18	22	20	14

*Indicates statistical differences between LEF and placebo ($p \leq .05$)

Quality of Life

The number of evaluable subjects with completed SF-36 quality of life questionnaires was 456 (95% of the intent-to-treat population) at baseline, 402 (84%) at week 24, and 265 (55%) at week 52. The corresponding numbers for the HAQ were 464 (97%), 412 (86%), and 281 (59%).

The LEF treatment group showed significant improvement from baseline to endpoint in all HAQ scale scores, PET score and Physical Component of SF-36 score. The results are presented below.

Mean Change in HAQ, PET and SF-36 Score			
	Leflunomide	Placebo	Methotrexate
HAQ Disability Index			
N	166	101	169
Baseline Mean	1.3	1.3	1.3
Mean Change	-0.45*,**	-0.03*	-0.26**
PET Weighted Top 5			
N	166	101	170
Baseline Mean	21.2	22.4	20.4
Mean Change	-6.91*,**	-0.66*	-3.41**
PET Weighted Top 10			
N	165	101	168
Baseline Mean	17.8	18.7	16.9
Mean Change	-6.19*,**	-0.29*	-2.93**
PET Unweighted Total			
N	166	101	170
Baseline Mean	3.7	3.8	3.7
Mean Change	-0.42*,**	0.06*	-0.19**
PET Overall Health			
N	163	99	168
Baseline Mean	5.7	5.6	5.6
Mean Change	0.90*	-0.22*	0.90
SF-36			
Physical Component			
N	157	101	162
Baseline Mean	30.0	28.9	29.7
Mean Change	7.6*	1.0*	4.6
Mental Component			
N	157	101	162
Baseline Mean	46.8	48.3	48.5
Mean Change	1.5	0.8	0.9

*Indicates statistically significant differences between leflunomide and placebo ($p \leq 0.001$)
**Indicates statistically significant differences between leflunomide and methotrexate ($p \leq 0.01$)

Safety:

The overall summary of adverse events for all phase III studies are presented in Table 7 in appendix A. In Study US301, 40 patients (22%) in LEF group withdrew due to adverse events. The adverse events in digestive system caused higher withdrawal rate than did other body systems. Among the 40 patients, 23 patients withdrew due to adverse events in digestive system. The adverse events in digestive system with withdrawal rate larger than 1% are listed in the following table.

Adverse Events in Digestive System with Withdrawal Rate ≥ 1%

	Total		LEF (n=182)		Placebo (n=118)		MTX (n=180)	
	N	%	N	%	N	%	N	%
LFT abnormal*	21	4.4	13	7.1	2	1.7	6	3.3
Diarrhea**	7	1.5	5	2.7	2	1.7	0	0
Nausea	4	0.8	3	1.6	0	0	1	0.5
GI bleeding	2	0.4	2	1.1	0	0	0	0

*: the rate in LEF group is statistically higher than that in the placebo group
**: the rate in LEF group is statistically higher than that in the MTX group ($p<0.05$)

III. Study MN301/MN303

1. Protocal

Study Type: This was a European multi-center, phase III, randomized, double-blind, double-dummy, placebo- and sulfasalazine-controlled study in subjects with active RA.

Primary Objective: To demonstrate that the efficacy of LEF is superior to that of placebo in patients with active RA.

Secondary Objectives: To compare the efficacy and safety of LEF with sulfasalazine (SSZ) in subjects with active RA.

Primary Efficacy Variables: Tender joint count (28 joints), swollen joint count (28 joints), investigator's global assessment of disease activity (5-point scale), patient's global assessment of disease activity (5-point scale).

Secondary Efficacy Variables: Responder rates (Paulus and American College of Rheumatology [ACR]), treatment success rate, joint tenderness score, swollen joint score, duration of morning stiffness, pain intensity assessment, Health Assessment Questionnaire (HAQ), x-ray of both hands and forefeet, erythrocyte sedimentation rate (ESR), C-reactive protein (CRP), rheumatoid factor (RF).

Safety Variables: Observed and spontaneously reported adverse events; hematology, blood chemistry, urinalysis; physical examination, supine blood pressure, pulse, oral body temperature, body weight, 12-lead electrocardiogram, chest x-ray.

Study Duration: The study was consisted of the following parts:
- a screening visit of 1 week
- a treatment phase of 24 weeks
- a treatment-free observation phase of 8 weeks

Study Visits: Visits were scheduled at fortnightly intervals during the first 8 weeks of the treatment phase and then at 4-weekly intervals during the rest of the treatment phase and the observation phase. Vital signs, resource utilization, and adverse events were recorded at all visits. The basic laboratory program and all efficacy variables (apart from the hand and forefeet x-rays) were performed at all visits during the treatment phase and at the final visit in the observation phase.

Subjects who withdrew from the study prematurely had to undergo a complete examination as scheduled for week 24 and then enter the observation phase. If the subject was withdrawn within the first 16 weeks of treatment, x-rays of hands and forefeet were only taken at visit 0. If the subject was withdrawn after more than 16 weeks of treatment,

the hands and forefeet also had to be x-rayed 24 weeks after the start of treatment with the study drug.

Sample Size: The recruitment of 108 LEF patients, 108 SSZ patients, and 72 was to provide 84% power to detect a difference between LEF and placebo in three of the four variables with a significance level of $\alpha = 0.0375$ for each variable or $\alpha=.05$ for all four variables. The following mean differences were to be detected: 5.0 for painful joint count, 3.5 for swollen joint count, .5 for investigator's global assessment, and .5 for patient's global assessment.

Randomization: Subjects were randomized to LEF, SSZ, or placebo treatment in a ratio of 3:3:2 and were stratified according to the duration of their RA prior to study entry (≤2 years or >2 years). After 24 weeks of treatment period, patients who had received LEF or SSZ in Study 301 continued on the respective medication; subjects who had received placebo in Study 301 were switched to SSZ in a blinded manner at the start of Study 303.

Statistical Methods:
In addition to descriptive statistics, the following tests were performed:
Baseline comparability
Categorical variables: chi-square test (if cell frequency below 5, Fisher's exact test)
Continuous variables: analysis of variance (ANOVA).
Efficacy variables
Joint counts: The joint counts were compared by analysis of covariance (ANCOVA), including treatment, investigator pool, duration of RA (≤2 years, >2 years), treatment × investigator pool interaction, and treatment × duration of RA interaction as fixed effects and the baseline value as a covariate.
Global assessments: The global assessments were compared using the extended Mantel-Haenszel test (with modified ridit scores), with stratification by investigator pool, baseline value, and duration of RA (≤2 years, >2 years).
Responder and treatment success rates: logistic regression.
All other secondary variables: ANCOVA.

2. RESULTS

Patient Disposition and Demographics

A total of 358 screened subjects were randomized and treated in the study: 133 with LEF, 92 with placebo, and 133 with SSZ. Five (5) subjects had no post-baseline data and were excluded from all efficacy analyses. A total of 353 patients were included in the intent-to-treat analyses.

During the treatment phase, 128 subjects were withdrawn from study medication, 37 from the LEF group, 41 from the placebo group, and 50 from the SSZ group. Reasons for withdrawal in the individual treatment groups are summarized as follows:

Reason for Withdrawal From Study Medication	Number (%) of subjects					
	Leflunomide (N=133)		Placebo (N=92)		Sulfasalazine (N=133)	
Lack of Efficacy	10	(8)	29	(32)	14	(11)
Adverse Event (inc. 1 death)	19	(14)	6	(7)	25	(19)
Refusal/Noncompliance	5	(4)	5	(5)	7	(5)
Other	3	(2)	1	(1)	4	(3)
Total withdrawal	37	(28)	41	(45)	50	(38)

The distribution of demographic characteristics, RA disease history and concomitant RA medication use in the three treatment groups are presented in Table 1-3 in Appendix A.

Efficacy

The following efficacy results are from the intend-to-treat analyses.

a. MN301

• Primary Efficacy Variables

The most important findings of the primary analysis comparing LEF to placebo are summarized in the following table, which indicates that LEF is statistically superior to placebo (p<=.0001):

Intent-to-treat analysis of primary efficacy variables: Baseline - endpoint comparison of LEF vs placebo

Variable/Statistic[a]	Mean (SD)				Treatment Comparison p-value (95% CI)[b]
	Leflunomide (N=130)		Placebo (N=91)		
Tender Joint Count					
Baseline	18.8	(6.6)	16.3	(6.32)	
Adjusted change at endpoint	-9.1	(7.48)	-5.1	(7.52)	0.0001 (-6.12; -2.05)
Swollen Joint Count					
Baseline	16.2	(5.99)	15.8	(5.57)	
Adjusted change at endpoint	-8.1	(6.10)	-4.3	(6.10)	0.0001 (-5.42; -2.21)
Investigator's global assessment					
Baseline	3.6	(0.62)	3.5	(0.58)	
Adjusted change at endpoint	-1.0	(0.92)	-0.4	(0.92)	<0.001 (-0.90; -0.40)
Patient's global assessment					
Baseline	3.7	(0.68)	3.6	(0.67)	
Adjusted change at endpoint	-1.1	(0.99)	-0.4	(0.99)	<0.001 (-0.95; -0.41)

[a] Changes were adjusted for factors in the ANCOVA model and estimated using proportional weights (investigator pool, disease duration).

[b] p-values from ANCOVA for joint counts and from extended Mantel-Haenszel statistics (with modified ridit scores) for global assessments; confidence intervals from ANCOVA

The findings for the secondary treatment comparison (LEF vs SSZ) are presented in the following table, which indicates the equivalency of LEF and SSZ.

Intent-to-treat analysis of primary efficacy variables:
Baseline - endpoint comparison of LEF vs SSZ

Variable/Statistic[a]	Mean (SD)				Treatment
	Leflunomide (N=130)		Sulfasalazine (N=132)		Comparison p-value (95% CI)[b]
Tender Joint Count					
Baseline	18.8	(6.6)	16.7	(6.33)	
Adjusted change at endpoint	-9.2	(7.19)	-8.5	(7.20)	0.4661 (-2.42; 1.11)
Swollen Joint Count					
Baseline	16.2	(5.99)	15.3	(5.61)	
Adjusted change at endpoint	-7.1	(5.68)	-6.4	(5.67)	0.3184 (-2.09; 0.68)
Investigator's global assessment					
Baseline	3.6	(0.62)	3.5	(0.57)	
Adjusted change at endpoint	-1.0	(0.88)	-1.0	(0.88)	0.834 (-0.22; 0.21)
Patient's global assessment					
Baseline	3.7	(0.68)	3.6	(0.65)	
Adjusted change at endpoint	-1.1	(0.99)	-1.1	(0.96)	0.401 (-0.28; 0.19)

[a] Changes were adjusted for factors in the ANCOVA model and estimated using proportional weights (investigator pool, disease duration).

[b] p-values from ANCOVA for joint counts and from extended Mantel-Haenszel statistics (with modified ridit scores) for global assessments; confidence intervals from ANCOVA

• Secondary Efficacy Variables

ACR success and responder: The numbers (%) of subjects classified as responders and as treatment successes are presented in the following table, which indicate that LEF is statistically superior to placebo and not statistically different from SSZ.

Study and Treatment Group	No. Subjects /Total N	Rate	p-value
ACR Success			
Leflunomide	63/130	48.5%	LEF vs PBO: $p \leq 0.0026$
Placebo	26/91	28.6%	
Sulfasalazine	59/132	44.7%	LEF vs SSZ: 95% CI: -8.3 to 15.8
ACR Responder			
Leflunomide	71/130	54.6%	LEF vs PBO: $p \leq 0.0001$
Placebo	26/91	28.6%	
Sulfasalazine	75/132	56.8%	LEF vs SSZ: 95% CI: -14.2 to 9.8

The responder rates along time for the three treatment groups are presented in Figure 2 in Appendix B. The findings for AUC, time to and duration of first response, and time to and duration of sustained response support the ACR responder's results (see Table 4 - 6 in Appendix A).

• Subset Analyses

When comparing the ACR success rate of LEF with that of placebo, treatment × age interaction was found significant (p-values=.06). The results for each subgroup are reported in the table below, which show that the interaction is quantitative.

ACR Success by Age						
Study and Treatment Group	Leflunomide		Placebo		Active	
	n/N	%	n/N	%	n/N	%
<65	41/88	46.6	15/65	23.1	35/81	43.2
>=65	22/42	52.4	11/26	42.3	24/51	47.1

X-Rays of Hands and Feet

The x-ray results are included in the following table, which indicate that LEF is statistically superior to placebo and not statistically different from SSZ.

Treatment Group	N	Baseline Score	Change at Endpoint
Sharp Total Score			
Leflunomide	89	46.26	-0.06*
Placebo	62	46.18	5.60*
Sulfasalazine	85	41.86	1.44†
Sharp Erosion Score			
Leflunomide	89	16.25	0.17*
Placebo	62	16.68	1.97*
Sulfasalazine	85	15.16	0.78†
Sharp Narrowing Score			
Leflunomide	89	30.01	-0.22*
Placebo	62	29.50	3.63*
Sulfasalazine	85	26.69	0.66†

*Indicates statistically significant differences between leflunomide and placebo ($p \leq 0.05$)
†Indicates leflunomide and the active comparator were not statistically different (95% CI overlapping 0)

Safety:

The overall summary of adverse events for all phase III studies are presented in Table 7 in Appendix A. In Study MN301, 19 patients (14.3%) in LEF group withdrew due to adverse events. The adverse events in digestive system caused higher withdrawal rate than other body systems. Among the 19 patients, 9 patients withdrew due to adverse events in digestive system. The adverse events in digestive system with withdrawal rate larger than 1% are listed in the following table.

Adverse Events in Digestive System with Withdrawal Rate ≥ 1%

	Total		LEF		Placebo		SSZ	
	N	%	N	%	N	%	N	%
Diarrhea*, **	11	3.1	8	6.0	1	1.1	2	1.5
Gastrointestinal pain	4	1.1	1	0.8	0	0	3	2.3
Liver function test abnormal	4	1.1	1	0.8	1	1.1	2	1.5
Nausea	9	2.5	3	2.3	0	0	6	4.5
Vomiting**	6	1.7	5	3.8	1	1.1	0	0

*: the rate in LEF group is statistically higher than that in the placebo group
**: the rate in LEF group is statistically higher than that in the SSZ group

b.MN303

Efficacy

A total of 197 (86%) of the 230 subjects who completed 24 weeks of treatment in Study MN301 entered Study MN303 (80 LEF, 76 SSZ, 41 placebo/SSZ). A total of 168 subjects completed the 24-week treatment phase of Study MN303 (71 LEF, 68 SSZ, 29 placebo/SSZ). The adjusted mean changes in joint counts and global assessments between baseline (visit 0, Study MN301) and endpoint (Study MN303) in the intention-to-treat population were:

	Mean (SD)								p- value
	Leflunomide (n=78)				Sulfasalazine (n=74)				LEF vs
Variable	Baseline		Adj. Change		Baseline		Adj. change		SSZ
Tender joint count	18.7	(6.49)	-10.1	(6.64)	15.8	(5.67)	-10.8	(6.64)	0.5215
Swollen joint count	16.3	(5.91)	-9.6	(5.32)	15.0	(4.90)	-10.6	(5.53)	0.2196
Inv. Global assessment	3.7	(0.66)	-1.3	(0.90)	3.5	(0.55)	-1.4	(0.90)	0.635
Pat. Global assessment	3.7	(0.69)	-1.3	(0.89)	3.6	(0.62)	-1.4	(0.89)	0.787

Safety

Only 2 patients withdrew due to adverse events: one was due to RA, the other one was due to Bronchiectasis.

IV. Study MN302

1. Protocol

This was a multinational, randomized, double-blind, parallel-group, MTX-controlled study. The primary objective of this study was to demonstrate that the efficacy of LEF was equivalent to that of MTX in patients with active RA. The study design of MN302 was similar to that of Study MN301 except the following:

Study Duration: the study was consisted of the following parts:

- a screening visit of 1 week
- a treatment phase of 52 weeks
- a treatment-free observation phase of 8 weeks

Sample Size: A total of 670 patients (335 in LEF group, 335 in MTX group) were recruited to show equivalence for all four primary variables using 95% confidence intervals with power 80%. The clinically meaningful differences (Δ) and standard deviations used for the sample size calculation are listed in the following table.

Variable	Δ	SD
Tender joint count	2.5	8.5
Swollen joint count	2.5	6.0
Investigator's global assessment	0.25	1.0
Patient's global assessment	0.25	1.0

Randomization: Subjects were randomized to LEF or MTX treatment in a ratio of 1:1. All subjects who completed the full 52-week treatment phase were offered the opportunity to participate in a 1-year, double-blind extension study (HWA 486/6/MN/304/RA). Subjects who completed the study but did not continue into the extension study were followed up for 8 weeks in an observation phase. Patients who withdrew from the study medication also entered the observation phase.

2. Results

Patient Disposition and Demographics

A total of 1244 patients were enrolled in the study, and 244 were withdrawn in the screening phase. One (1) subject was randomized to MTX but was not treated. A total of 999 subjects were randomized and treated in the study (501 LEF, 498 MTX). Among the 999 patients, 6 subjects were excluded from efficacy analyses because assessments could not be validated. Additionally, 3 LEF subjects and 6 MTX subjects had no baseline or post-baseline efficacy assessments. The intent-to-treat population included 495 patients in LEF group and 498 patients in MTX group.

During the treatment phase, 263 subjects were withdrawn from study medication, 152 from the LEF group (listed in Table 24) and 111 from the MTX group. Reasons for withdrawal are summarized below:

	Number (%) of subjects withdrawn from study medication			
Reason for withdrawal	Leflunomide (N=501)		Methotrexate (N=499)	
Lack of efficacy	37	(7)	15	(3)
Adverse events	94	(19)	74	(15)
Death	4	(1)	5	(1)
Refusal/Noncompliance	11	(2)	14	(3)
Other	6	(1)	3	(<1)
Total	152	(30)	111	(22)

The distribution of demographic characteristics, RA disease history and concomitant RA medication use in the three treatment groups are presented in Table 1-3 in Appendix A.

Efficacy

The following efficacy results are from the intend-to-treat analyses.

• Primary Efficacy Variables

All of the covariates included in the ANCOVA model (investigator pool, duration of RA, and baseline status) showed statistically significant effects ($p < 0.05$) on the reductions in joint counts and improvements in global assessments. No interactions were detected between treatment and investigator pool or between treatment and disease duration. The most important findings of the primary analysis comparing LEF to MTX are summarized as follows, which indicate that MTX is superior to LEF.

Study and Treatment Group	N	Mean Baseline TJC	Mean Change at Endpoint	Confidence Interval *	P-value
Tender Joint Count					
Leflunomide	495	17.2	-8.3	(0.37, 2.16)	0.006
Methotrexate	489	17.7	-9.7		
Swollen Joint Count					
Leflunomide	495	15.8	-6.8	(1.03, 2.63)	0.0001
Methotrexate	489	16.5	-9.0		
Patient's Global					
Leflunomide	495	3.6	-0.9	(0.12,0.35)	0.001
Methotrexate	489	3.6	-1.2		
Investigator's Global					
Leflunomide	495	3.5	-1.0	(0.16, 0.38)	0.001
Methotrexate	489	3.6	-1.2		

* C.I. of the difference of LEF and MTX in terms of mean change

• Secondary Efficacy Variables

ACR success and responder: The numbers (%) of subjects classified as responders and as treatment successes are presented in the following table, which indicate that LEF is superior to MTX.

Study and Treatment Group	No. Subjects /Total N	Rate	p-value
ACR Success			-
Leflunomide	215/495	43.4%	
Methotrexate	277/489	56.6%	LEF vs MTX: 95% CI: -19.4 to –7.0
ACR Responder			
Leflunomide	253/495	51.1%	
Methotrexate	319/489	65.2%	LEF vs MTX: 95% CI: -20.2 to –8.0

The responder rates at different time poinrs for the three treatment groups are presented in Figure 3 in Appendix B. The findings for the analysis on response as AUC, time to and duration of first response, and time to and duration of sustained response support the ACR responder's result and are listed in Table 4 - 6 in Appendix A.

Other secondary efficacy variables: MTX is statistically superior to LEF in joint tenderness score (p=.0188), swollen joint score (p=.0001), duration of morning stiffness

(p=.0048), pain intensity assessment, HAQ (p=.0126), ESR (p<.0001), and CRP (p=.0002).

• Subset analyses
When comparing the ACR success rate of LEF with that of MTX, no interaction was found significant between the treatment groups and other relevant factors.

X-Rays of Hands and Feet

The numbers of patients with both baseline and change from baseline Sharp x-ray result are included in the following table, which indicate that MTX and LEF were not statistically different in Sharp x-ray scores.

Treatment Group	N	Baseline Score	Change at Endpoint
Sharp Total Score			
Leflunomide	304	24.94	2.19*
Methotrexate	331	24.60	1.04
Sharp Erosion Score			
Leflunomide	304	9.48	0.87*
Methotrexate	331	9.56	0.25
Sharp Narrowing Score			
Leflunomide	304	15.46	1.32*
Methotrexate	331	15.04	0.79

*Indicates LEF and MTX were not statistically different (95% CI overlapping 0)

Safety:

The overall summary of adverse events for all phase III studies are presented in Table 7 in Appendix A. In Study MN302, 94 patients (18.8%) in LEF group withdrew due to adverse events. The adverse events in digestive system caused higher withdrawal rate than did other body systems. Among the 94 patients, 40 patients withdrew due to adverse events in digestive system. The adverse events in digestive system with withdrawal rate larger than 1% are listed in the following table.

Adverse Events in Digestive System with Withdrawal Rate ≥ 1%

	Total		LEF		MTX	
	N	%	N	%	N	%
Anorexia	9	0.9	4	0.8	5	1.0
Diarrhea	17	1.7	11	2.2	6	1.2
Gastrointestinal pain	9	0.9	4	0.8	5	1.0
Liver function test abnormal	26	2.6	10	2.0	16	3.2
Nausea	22	2.2	9	1.8	13	2.6
Vomiting	13	1.3	6	1.2	7	1.4

No statistically significant difference was found among the above adverse event rates between LEF and MTX.

V. Reviewer Comments

1. The Influence of missing Sharp X-ray Score

The x-rays for US301 were read only in pairs by Dr. Sharp in a blinded fashion (i.e., Dr. Sharp read the baseline and endpoint x-ray at the same time and was blinded to the timing of the x-rays). Subsets for whom only baseline film was available were thus not read. Sets of single baseline x-rays for MN301 and MN302 were read, but some of the x-ray results were not readable.

In terms of the patient's disposition in x-ray score, the patient population can be classified into the following three categories:
NB: patients with no baseline x-ray score;
YBNF: patients with baseline x-ray score but without follow-up x-ray score; and
YBYF: patients with both baseline x-ray and follow-up x-ray score.

The distribution of patients in the above three categories are listed in the following table for all three Phase III trials.

	US301	MN301	MN302
NB	120 (25%)	43 (12%)	130 (13%)
YBNF	6 (1%)	74 (21%)	219 (22%)
YBYF	354 (74%)	235(67%)	635 (65%)

The x-ray category × clinical disposition tables and the x-ray category × clinical disposition table stratified by treatment groups for each study are attached (Tables 8 - 19 in Appendix A). To assess how the x-ray missing values could affect the inference, the reviewer explored the following questions:

Q1. Are there any difference between the patients with/without baseline Sharp x-ray score in terms of the clinical measurements?

In Study US301, 78% of the patients in group NB also withdrew from the clinical study early (Table 8 in Appendix A). So clinically, the NB group behaved similarly to the early terminated patient group. The ACR success rate in YB (YBNF+YBYF) group was significantly higher than that in the NB group (36% vs. 15%, p=.001; χ^2 test). Statistical significance is also found for the baseline measurements and the changes from baseline clinical measurements. The results from ANOVA tests are listed below.

	PNSCAL0		MHAQ0		WLSCAL0	
	NB	YB	NB	YB	NB	YB
lsmean	64.19	59.03	.90	.78	5.94	5.45
stderr	1.99	1.16	.05	.03	.21	.12
Pvalue	.026		.04		.037	

PNSCAL0: baseline pain intensity evaluation
MHAQ0: baseline quality of life evaluation
WLSCAL0: baseline patient evaluation

	DIFFSWL		DIFFMHAQ		DIFFMD		DIFFESR	
	NB	YB	NB	YB	NB	YB	NB	YB
lsmean	-3.57	-5.01	-.04	-.15	-1.54	-2.23	.86	-4.74
stderr	.56	.33	.05	.03	.25	.14	2.04	1.18
Pvalue	.025		.05		.02		.02	

DIFFSWL: change in swollen joint count
DIFFMHAQ: change in quality of life
DIFFMD: change in physician evaluation
DIFFESR: change in ESR

In Study MN301, 35% of the patients in group NB also withdrew from the clinical study early, which was similar to the rate in the overall population (35%). The ACR success rate in YB (YBNF+YBYF) group was significantly lower than that in the NB group (40% vs. 56%, p=.044; χ^2 test). No statistical significance is found for the baseline clinical measurements and the changes from baseline clinical measurements.

In Study MN302, 32% of the patients in group NB also withdrew from the clinical study early, which was higher than the rate in the overall population (26%). The ACR success rate in YB (YBNF+YBYF) group was not significantly different from that in the NB group (50% vs. 45%, p=.29; χ^2 test). Baseline physician's evaluation score is the only measurement found significantly different between the NB and YB group (NB group has a higher mean with p=.01; ANOVA).

Overall, there are no consistent finding for all studies in terms of baseline clinical measurements and the change from baseline clinical measurements.

Q2. How does the x-ray missing structure interact with treatment group in clinical measurements?

No x-ray missing structure (NB, YB) by treatment interactions are found among all baseline and the change from baseline clinical scores.

Q3. Are the clinical measurements different in the YBNF group and the YBYF group in terms of baseline x-ray score and clinical measurements?

Since Study US301 has only 6 patients in the YBNF group, the following discussion focuses on Study MN301 and Study MN302.

The YBNF group and YBYF group were not statistically different in baseline Sharp x-ray score. In Study MN301, 77% of the patients in group YBNF terminated the clinical study early. In Study MN302, 71% of the patients in group YBNF withdrew from the clinical study early. So clinically, the YBNF group behaved similar to the patients group which withdrew from the clinical study early. In Study MN301, the ACR success rate is significantly higher in the YBYF group than that in the YBNF group (62% vs. 16%, p=.001; χ^2). The YBNF group has significant higher value (smaller improvement) than the YBYF group in change of morning stiffness (p=.016; ANOVA), change of pain intensity (p=.0018; ANOVA), change of physician's evaluation (p=.0003; ANOVA), and change in patient's evaluation (p=.005; ANOVA). In Study MN302, the ACR success rate is significantly higher in the YBYF group than that in the YBNF group (48% vs. 14%, p=.001; χ^2). The YBNF group has significant higher value (smaller improvement) than the YBYF group in all changes of clinical measurements (p≤..004; ANOVA). No x-ray missing structure (YBNF, YBYF) × treatment interactions are found.

Overall, due to the high clinical withdrawal rate, the YBNF group has less clinical improvement than the YBYF group.

Q4. How Robust is the X-ray Result ?

Due to the high rate of missing value in x-ray score, the following analyses (Analyses A) was requested by the statistical reviewer to assess the robustness of the results.

Step 1. A random subset of non-missing baseline, change-from-baseline x-ray scores, and covariates in the x-ray ANOVA model was created for each of the two treatment groups (LEF and Placebo) and each of the studies (MN301, US301).

Step 2. Each time a missing change score was encountered, a new change score, along with its corresponding baseline value and covariates, was selected as a substitute from the randomly ordered subset of the opposite treatment group. Thus, selection was made without replacement.

Step 3. Treatment group means and p-values from the resulting ANCOVA analysis were recorded.

Step 4. Step 1, 2, and 3 were repeated 100 times.

Step 5. Resulting Statistics (means, standard deviation, medians, upper and lower quartiles, ect.) were computed and tabulated for the 100 resulting mean values from each treatment group.

Step 6. The number of significant p-values out of the 100 runs were calculated and presented.

The results of the above analysis showed that, in Study US301, 55 of the 100 samples resulted in p-value less than 0.05, and in Study MN301, 35 of the 100 samples resulted in p-value ≤ 0.05.

The medical officer also requested sensitivity analysis (Analysis B, see Appendix C). The results showed that, to make the p-value of .0007 for the non-missing value group comparison to disappear s, the boundary mean value for the LEF missing value group need to be worse than that of the placebo group.

One thing needs to be concerned for the above two analyses (medical officer's and statistical reviewer's) is that, when doing the ANOVA, the covariates were not the true covariates of the missing value group. Since none of the covariates were reported significant in the ANOVA analysis for the observed group and the there were no baseline imbalances for the covariates among treatment groups, another approach would be to calculate the p-value theoretically for analyses A and B without any covariate adjustment. The calculation was done as the following for Analysis A.

Step 1. Assume the LEF observed x-ray scores follow a normal distribution $N(\mu_1, \sigma_1^2)$, and the placebo observed x-ray scores follow a normal distribution $N(\mu_2, \sigma_2^2)$. Further, assume the LEF missing x-ray scores follow the normal distribution $N(\mu_2, \sigma_1^2)$, and the placebo missing x-ray scores follow the normal distribution $N(\mu_1, \sigma_2^2)$.

Step 2. Combining the missing and observed groups, the mean x-ray scores in LEF group approximately follow a normal distribution $N(m_1, v_1^2)$, and the mean x-ray scores in placebo group approximately follow a normal distribution $N(m_2, v_2^2)$, where $m_1=(n_1*\mu_1+(N_1-n_1)*\mu_2)/N_1$, $m_2=(n_2*\mu_2+(N_2-n_2)*\mu_1)/N_2$, $v_1^2=(n_1*\sigma_1^2+(N_1-n_1)*\sigma_2^2)/N_1^2$, $v_2^2=(n_2*\sigma_2^2+(N_2-n_2)*\sigma_1^2)/N_2^2$, n_1, n_2 denote the sample sizes of the observed groups, and N_1, N_2 denote the total sample sizes of LEF and placebo group, respectively. The parameters m_1, m_2, v_1 and v_1 can be estimated by replacing μ_1, μ_2, σ_1, and σ_2 by the sample means and standard deviations in the observed x-ray groups. We denote these estimators as em_1, em_2, ev_1, and ev_2.

Step 3. The test statistic for testing hypothesis H_0: $m_1=m_2$ is $\dfrac{em_2 - em_1}{\sqrt{ev_1^2 + ev_2^2}}$.

Step 4. P-value can be obtained based on the value of the Z-statistic.

The calculation results are presented in the following table.

Treatment	Observed Mean	Observed SD	n	N	em	ev	T-statistic	P-value
US301								
LEF	0.534	4.532	131	182	0.989637	0.335934	1.39066	0.164
PLB	2.16	3.946	83	118	1.677712	0.363259		
MN301								
LEF	-0.06	12.33	89	130	1.725077	1.081413	1.38656	0.166
PLB	5.6	9.83	62	91	3.796264	1.030464		

In Analysis B, the boundary value (Δ) for the increment/decrement such that p=0.05 (see Appendix C) can be calculated by the following formula:

$$\Delta=(\bar{x}_2-\bar{x}_1-Z_{.975}\times\sqrt{ev_1^2+ev_2^2})/(\frac{N_1-n_1}{N_1}+\frac{N_{21}-n_2}{N_2}),$$

where $\bar{x}_1$ and $\bar{x}_2$ are the observed means of the LEF group and the placebo group, respectively.

In the LEF missing value group, the boundary mean will be $\bar{x}_1+\Delta$, and in the placebo missing value group, the boundary mean will be $\bar{x}_2-\Delta$. The Δ value and the boundary value of the means (m) in the missing-value groups are presented in the following table.

Treatment	Observed Mean	Observed SD	n	N	Δ	m
US301						
LEF	0.534	4.532	131	182	1.120491	1.65
PLB	2.16	3.946	83	118	1.120491	1.04
MN301						
LEF	-0.06	12.33	89	130	4.261957	4.20
PLB	5.6	9.83	62	91	4.261957	1.34

The above results confirm the results of analyses A and B. That is, to make the p-value reach .05, the boundary mean value for the LEF missing value group (1.65 in US301, 4.20 in MN301) needs to be worse than that of the placebo group (1.04 in US301, 1.34 in MN301), but the boundary mean value for the LEF missing value group is not quite as large as the observed placebo progression (2.16 in US310, 5.6 in MN301) and the boundary mean value of the placebo missing value group was not quite as good as the observed LEF result (.534 in US301, -.06 in MN301).

2. Correlation Between Sharp x-ray Score and Clinical Measurements

In Study US301, the mean Sharp x-ray score was marginally higher (deteriorated more) in the ACR non-responder group than that in the ACR responder group (p=.07; T-test). The Sharp x-ray score was significantly correlated with the change in swollen joint count, and AUC of responder status, but the correlation coefficients were low. The p-values and the correlation coefficients are listed in the table below.

	R**2	R	P-value
DIFFSWL	.016	.126	.03
AUC	.023	.152	.01

DIFFSWL: change in swollen joint count
AUC: area under the curve of response time

In Study MN301, no correlation was found between the Sharp x-ray score and clinical measurements.

In Study MN302, the mean Sharp x-ray score was marginally higher (deteriorate more) in the ACR non-responder group than that in the ACR responder group (p=.003; T-test). The Sharp x-ray score was significantly correlated with the following clinical measurements, but the correlation coefficients were extremely low. The p-values and the correlation coefficients are listed in the table below.

	R**2	R	P-value
DIFFSWL	.007	.084	.024
DIFFPAIN	.005	.071	.056
DIFFESR	.008	.089	.019
DIFFPN	.007	.084	.037
DIFFMD	.012	.11	.007

DIFFSWL: change in swollen joint count
DIFFPAIN: change in tender joint count
DIFFESR: change in ESR
DIFFPN: change in pain intensity
DIFFMD: change in physician evaluation

Overall, Study US301 and MN302 showed very little correlation between the Sharp x-ray score and some clinical measurements, and Study MN301 showed none.

3. Comparison of LEF and MTX

The results of the comparison of LEF and MTX in Study US301 and MN302 were not consistent. In Study MN302, LEF was statistically superior to MTX in all clinical endpoints (ACR success rate, ACR responder rate, the individual components of ACR response, morning stiffness and HAQ). In Study US301, although LEF was clinically equivalent to MTX by the prespecified equivalence criteria (the 95% confidence interval of ACR success rate overlaps 0 and the lower bound is above -15%), LEF showed higher ACR responder rate than MTX since month 1, and LEF showed more improvement in all individual components of ACR response than MTX since month 5.

Besides some possible clinical reasons, the difference in the study design might have contributed to the discrepancies discussed above. In Study US301, subjects who remained in the study for a minimum of 16 weeks but were not responders by ACR criteria or who

experienced significant toxicity or persistent laboratory abnormalities were allowed to discontinue initial therapy and, after appropriate washout, entered alternate therapy on a blinded basis. While in Study MN302, no alternate therapy was offered to the early terminated patients. In Study US301, there was a sharp increase in the number of discontinuations at the 4 month visit. The overall number of dropout and the number of dropouts at the 4 month visit are listed in the following table:

	LEF (N=182)		MTX (N=180)		Placebo (N=118)	
	n	%	n	%	n	%
Overall Dropouts	31	17.0%	44	24.2%	62	52.5%
Dropouts at Month 4	13	7.1%	28	15.6%	38	32.2%
Dropouts at Month 4 Entered Alternate Therapy	13	7.1%	20	11.1%	36	30.5%

This reviewer suspects that the alternate therapy might have attracted some patients with slow improvement in the first 4 month, but could become ACR responders later, and these patients were more likely to be from the MTX group than from the LEF group due to the following reasons:

1. At month 4 visit, there was a higher drop-out rate in MTX patients (15.6%) than in the LEF patients (7.1%).
2. Figure 1 shows that, during the first 4 months of treatment, MTX had slower improvement in almost all components of ACR response (except CRP) than that of LEF.

So the alternate therapy option might have artificially enlarged the difference between the ACR success rate of LEF and MTX.

VI. Overall Summary and Conclusion

1. LEF vs. Placebo

• Study US301 and MN301 demonstrated the superiority of LEF over placebo in terms of ACR success rate and the individual components of the ACR success rate.

• Although, in Study US301 and MN301, the LEF group showed statistically significantly less disease progression than the placebo group in terms of Sharp x-ray score (see table below), the high rate of missing x-ray score (26% in US301, 33% in MN301) should be taken into consideration. Although some sensitivity analyses on the influence of x-ray missing values were done, and the results support the Sponsor's conclusion, the full extent of the meaning of the high rate of missing values remains worrisome.

• Study US301 showed that the LEF treatment group was statistically superior to the placebo group in terms of HAQ scale scores, PET score and Physical Component of SF-36 score.

2. LEF vs. MTX

• The efficacy results of LEF vs. MTX were inconsistent between Study US301 and MN302. Study US301 showed that LEF and MTX were statistically equivalent. On the other hand, although not statistically significant, the ACR success rate and the individual components of the ACR success response were consistently better for LEF. Study MN302 showed that MTX was statistically superior to LEF in terms of ACR success rates and the individual components of the ACR success rate. The ACR success rate of MTX in Study MN302 was much higher than that in US301 (56.5% vs. 35.0%).

• The results in Study US301 and MN302 were inconsistent in terms of x-ray score. Study US301 showed LEF was statistically superior to MTX in terms of Sharp x-ray score, while Study MN302 did not show any statistical significance. Again, the high rate of missing x-ray values should be taken into consideration.

• Study US301 showed that the LEF treatment group was statistically superior to MTX group in terms of HAQ scale scores, but there was no statistically significant difference in PET score and Physical Component score of SF-36.

3. LEF vs. SSZ

• The efficacy of LEF and SSZ were only compared in Study MN301. It showed that LEF and SSZ were statistically equivalent in terms of ACR success rate and the individual components of the ACR success response.

• Study MN301 showed that LEF and SSZ were not statistically different in terms of change in Sharp x-ray score.

Laura Lu, Ph.D.
Mathematical Statistician

Concur:

8/18/98

Stan Lin, Ph.D.
Team Leader

CC:

HFD-550/MO/Johnson
HFD-550/PM/Cook
HFD-550/MO/Hyde
HFD-550/Div. File
HFD-725/Lu
HFD-725/Lin ST.
HFD-725/Huque
HFD-725/Div. File

Appendix A. Tables

Table 1 Demographic Characteristics

	Age			Gender		Race	
Study and Treatment Group	Mean (years)	<65 yrs N (%)	≥65 yrs N (%)	Male N (%)	Female N (%)	Caucasian N (%)	Other Race N (%)
US301 (12 months)							
Leflunomide	54.2	141 (78%)	41 (23%)	50 (27%)	132 (73%)	161 (89%)	21 (12%)
Placebo	54.6	97 (82%)	21 (18%)	35 (30%)	83 (70%)	103 (88%)	14 (12%)
Methotrexate	53.3	147 (81%)	35 (19%)	45 (25%)	137 (75%)	162 (89%)	20 (11%)
MN301 (6 months)							
Leflunomide	58.3	90 (68%)	43 (32%)	32 (24%)	101 (76%)	115 (87%)	18 (14%)
Placebo	58.8	65 (71%)	27 (29%)	23 (25%)	69 (75%)	82 (89%)	10 (11%)
Sulfasalazine	58.9	82 (62%)	51 (38%)	41 (31%)	92 (69%)	124 (93%)	9 (7%)
MN301/303 (12 mos.)							
Leflunomide	57.8	51 (64%)	29 (36%)	20 (25%)	60 (75%)	69 (86%)	11 (14%)
Sulfasalazine	58.7	46 (61%)	30 (40%)	26 (34%)	50 (66%)	69 (91%)	7 (9%)
MN302 (12 mos.)							
Leflunomide	58.3	347 (70%)	154 (31%)	147 (29%)	354 (71%)	495 (99%)	6 (1%)
Methotrexate	57.8	348 (70%)	150 (30%)	143 (29%)	355 (71%)	491 (99%)	7 (1%)

Table 2 RA Disease History

	Duration of RA				DMARD Use	
Study and Treatment Group	Mean (yrs)	≤2 yrs N (%)	>2-10 yrs N (%)	>10 yrs N (%)	Prior DMARD N (%)	Mean No. DMARDs Used
US301 (12 months)						
Leflunomide	7.0	71 (39%)	66 (36%)	45 (25%)	101 (56%)	0.8
Placebo	6.9	39 (33%)	53 (45%)	25 (21%)	71 (60%)	0.9
Methotrexate	6.5	73 (40%)	72 (40%)	37 (20%)	102 (56%)	0.9
MN301 (6 months)						
Leflunomide	7.6	50 (38%)	43 (32%)	40 (30%)	80 (60%)	1.2
Placebo	5.7	41 (45%)	29 (32%)	22 (24%)	43 (47%)	0.9
Sulfasalazine	7.4	56 (42%)	40 (30%)	37 (28%)	65 (49%)	1.0
MN301/303 (12 mos.)						
Leflunomide	6.4	33 (41%)	27 (34%)	20 (25%)	49 (61%)	1.2
Sulfasalazine	6.5	32 (42%)	27 (36%)	17 (22%)	37 (49%)	0.9
MN302 (12 mos.)						
Leflunomide	3.7	219 (44%)	270 (54%)	12 (2%)	332 (66%)	1.1
Methotrexate	3.8	215 (43%)	265 (53%)	18 (4%)	333 (67%)	1.1

Table 3 Concomitant RA Medication Use			
Study and Treatment Group	NSAID Use N (%)	Corticosteroid Use N (%)	Both N (%)
US301 (12 months)			
Leflunomide	137 (75%)	98 (54%)	74 (41%)
Placebo	77 (65%)	65 (55%)	46 (39%)
Methotrexate	127 (70%)	96 (53%)	66 (36%)
MN301 (6 months)			
Leflunomide	114 (86%)	60 (45%)	49 (37%)
Placebo	80 (87%)	41 (45%)	35 (38%)
Sulfasalazine	104 (78%)	61 (46%)	47 (35%)
MN301/303 (12 mos.)			
Leflunomide	68 (85%)	29 (36%)	24 (30%)
Sulfasalazine	56 (74%)	32 (42%)	23 (30%)
MN302 (12 mos.)			
Leflunomide	402 (80%)	358 (72%)	285 (57%)
Methotrexate	431 (87%)	323 (65%)	279 (56%)

Table 4 Summary of Mean AUC		
Study and Treatment Group	Mean AUC(weeks)	p-value
US301 (12 months)		
Leflunomide	23.7	LEF vs PBO: $p \leq 0.0001$
Placebo	12.6	
Methotrexate	22.7	LEF vs MTX: 95% CI: -3.8 to 6.2
MN301 (6 months)		
Leflunomide	11.8	LEF vs PBO: $p \leq 0.0001$
Placebo	5.5	
Sulfasalazine	10.5	LEF vs SSZ: 95% CI: -0.8 to 3.6
MN301/303 (12 mos.)		
Leflunomide	21.9	
Sulfasalazine	20.1	LEF vs SSZ: 95% CI: -2.6. to 6.5
MN302		
Leflunomide	23.0	
Methotrexate	25.4	LEF vs MTX: 95% CI: -4.8 to 0.1

Table 5 Summary of Time to and Duration of Initial Response by ACR Criteria				
Study and Treatment Group	% of Subjects Responding at Any Time N	%	Mean Time to Initial Response (weeks)	Mean Duration of Initial Response (weeks)
US301 (12 months)				
Leflunomide	143	78.6	8.5	19.3
Placebo	63	53.4	10.4	13.8
Methotrexate	139	77.2	9.5	16.2
MN301 (6 months)				
Leflunomide	103	79.2	7.3	12.5
Placebo	42	46.2	10.0	8.6
Sulfasalazine	97	73.5	8.3	10.9
MN301/303 (12 mos.)				
Leflunomide	104	80.0	7.6	20.9
Sulfasalazine	99	75.0	8.9	18.7
MN302 (12 mos.)				
Leflunomide	379	76.6	10.6	22.5
Methotrexate	404	82.6	14.4	24.3

Table 6 Summary of Time to and Duration of Sustained Response by ACR Criteria				
Study and Treatment Group	**Subjects Achieving a Sustained Response**		**Mean Time to Sustained Response (weeks)**	**Mean Duration of Sustained Response (weeks)**
	N	**%**		
PIVOTAL STUDIES				
US301 (12 months)				
Leflunomide	96	52.7	10.7	33.4
Placebo	40	33.9	14.7	26.4
Methotrexate	103	57.2	14.0	29.6
MN301 (6 months)				
Leflunomide	68	52.3	7.3	18.8
Placebo	18	19.8	9.7	17.3
Sulfasalazine	63	47.7	8.3	17.6
SUPPORTIVE STUDIES				
Active-Controlled Studies				
MN301/303 (12 mos.)				
Leflunomide	71	54.6	8.1	33.7
Sulfasalazine	67	50.8	9.3	30.6
MN302				
Leflunomide	206	41.6	9.3	40.2
Methotrexate	243	49.7	11.2	39.2

Table 7 Adverse Events with Frequency of 1% or Greater in all LEF Group and Frequency in LEF Group in Controlled Studies Equal to or Greater than Placebo	All LEF TX'D RA		Placebo-Controlled Studies MN301 and US301								Active Control Study MN302			
			LEF		PL		MN301/3 SSZ		US301 MTX		MN302 MTX		LEF	
	N 1339	% 100	N 315	% 100	N 210	% 100	N 133	% 100.0	N 182	% 100.0	N 498	% 100	N 501	% 100
Body as a Whole														
Abdominal Pain	74	5.5	16	5.1	8	3.8	5	3.8	14	7.7	19	3.8	29	5.8
Back Pain	66	4.9	20	6.3	7	3.3	5	3.8	16	8.8	34	6.8	38	7.6
Accidental Injury	65	4.9	22	7.0	11	5.2	4	3.0	20	11.0	34	6.8	32	6.4
Asthenia	43	3.2	20	6.3	8	3.8	7	5.3	10	5.5	16	3.2	13	2.6
Flu Syndrome	32	2.4	11	3.5	5	2.4	0	0.0	13	7.1	0	0.0	0	0.0
Pain	26	1.9	12	3.8	5	2.4	2	1.5	9	4.9	2	0.4	4	0.8
Infection	23	1.7	4	1.3	2	1.0	3	2.3	7	3.8	8	1.6	10	2.0
Allergic Reaction	20	1.5	15	4.8	5	2.4	0	0.0	10	5.5	9	1.8	5	1.0
Cardiovascular System														
Hypertension	138	10.3	28	8.9	9	4.3	5	3.8	5	2.7	20	4.0	49	9.8
Chest Pain	22	1.6	12	3.8	5	2.4	3	2.3	8	4.4	9	1.8	7	1.4
Tachycardia	20	1.5	3	1.0	2	1.0	0	0.0	2	1.1	1	0.2	10	2.0
Gastrointestinal System														
Diarrhea	227	17.0	84	26.7	25	11.9	13	9.8	35	19.2	50	10.0	111	22.2
Nausea	124	9.3	41	13.0	23	11.0	25	18.8	33	18.1	90	18.1	64	12.8
Dyspepsia	66	4.9	32	10.2	21	10.0	12	9.0	24	13.2	35	7.0	29	5.8
LFT Abnormal	65	4.9	32	10.2	5	2.4	5	3.8	19	10.4	84	16.9	29	5.8
GI Pain/Abd.Pain	61	4.6	18	5.7	9	4.3	9	6.8	15	8.2	38	7.6	40	8.0
Vomiting	38	2.8	16	5.1	9	4.3	5	3.8	5	2.7	17	3.4	16	3.2
Anorexia	34	2.5	9	2.9	5	2.4	7	5.3	3	1.6	15	3.0	13	2.6
Flatulence	16	1.2	9	2.9	2	1.0	3	2.3	6	3.3	6	1.2	5	1.0
Cholelithiasis	15	1.1	3	1.0	0	0.0	1	0.8	0	0.0	3	0.6	3	0.6
Oral														
Mouth Ulcer	33	2.5	15	4.8	8	3.8	4	3.0	18	9.9	28	5.6	17	3.4
Tooth Disorder	24	1.8	9	2.9	4	1.9	1	0.8	5	2.7	20	4.0	7	1.4
Salivary Gland	19	1.4	3	1.0	0	0.0	1	0.8	0	0.0	4	0.8	5	1.0
Stomatitis	15	1.1	4	1.3	0	0.0	0	0.0	1	0.5	9	1.8	9	1.8
Hemic & Lymphatic System														
Leukopenia*	38	2.8	7	2.2	0	0.0	3	2.3	2	1.1	13	2.6	22	4.4
Metabolic & Nutritional Disorders														
Weight Loss	47	3.5	7	2.2	1	0.5	2	1.5	0	0.0	9	1.8	9	1.8
Hypokalemia	16	1.2	10	3.2	2	1.0	1	0.8	1	0.5	1	0.2	4	0.8
Musculo-Skeletal System														
Joint Disorder	54	4.0	8	2.5	4	1.9	3	2.3	3	1.6	28	5.6	39	7.8
Tenosynovitis	36	2.7	7	2.2	0	0.0	1	0.8	4	2.2	5	1.0	24	4.8
Myalgia	18	1.3	8	2.5	4	1.9	4	3.0	7	3.8	6	1.2	9	1.8
Arthralgia	17	1.3	11	3.5	7	3.3	0	0.0	16	8.8	1	0.2	3	0.6
Leg Cramps	16	1.2	11	3.5	5	2.4	3	2.3	10	5.5	12	2.4	5	1.0

Table 7	Adverse Events with Frequency of 1% or Greater in all LEF Group and Frequency in LEF Group in Controlled Studies Equal to or Greater than Placebo													
	All LEF TX'D RA		Placebo-Controlled Studies MN301 and US301								Active Control Study MN302			
			LEF		PL		MN301/3 SSZ		US301 MTX		MN302 MTX		LEF	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%
	1339	100	315	100	210	100	133	100.0	182	100.0	498	100	501	100
Central Nervous System														
Headache	91	6.8	42	13.3	24	11.4	16	12.0	38	20.9	39	7.8	48	9.6
Dizziness	56	4.2	16	5.1	7	3.3	8	6.0	9	4.9	31	6.2	35	7.0
Paresthesia	29	2.2	9	2.9	3	1.4	1	0.8	4	2.2	14	2.8	18	3.6
Vertigo	23	1.7	4	1.3	2	1.0	5	3.8	0	0.0	9	1.8	7	1.4
Respiratory System														
Resp. Infection	202	15.1	66	21.0	43	20.5	27	20.3	58	31.9	122	24.5	133	26.5
Bronchitis	87	6.5	16	5.1	4	1.9	5	3.8	12	6.6	34	6.8	40	8.0
Pharyngitis	41	3.1	6	1.9	3	1.4	2	1.5	1	0.5	13	2.6	14	2.8
Rhinitis	32	2.4	15	4.8	5	2.4	5	3.8	5	2.7	10	2.0	12	2.4
Pneumonia	29	2.2	10	3.2	0	0.0	0	0.0	2	1.1	11	2.2	11	2.2
Sinusitis	26	1.9	15	4.8	10	4.8	0	0.0	18	9.9	7	1.4	7	1.4
Dyspnea	18	1.3	8	2.5	2	1.0	5	3.8	9	4.9	10	2.0	8	1.6
Skin														
Rash	132	9.9	39	12.4	14	6.7	14	10.5	16	8.8	48	9.6	54	10.8
Alopecia	130	9.7	28	8.9	3	1.4	8	6.0	11	6.0	49	9.8	81	16.2
Pruritis	57	4.3	15	4.8	4	1.9	4	3.0	4	2.2	10	2.0	29	5.8
Dry Skin	28	2.1	10	3.2	5	2.4	2	1.5	0	0.0	6	1.2	17	3.4
Eczema	25	1.9	4	1.3	1	0.5	1	0.8	1	0.5	9	1.8	17	3.4
Rash, Macpap	24	1.8	9	2.9	2	1.0	4	3.0	2	1.1	10	2.0	8	1.6
Skin Disorder	24	1.8	9	2.9	2	1.0	4	3.0	0	0.0	13	2.6	7	1.4
Herpes Simplex	18	1.3	4	1.3	1	0.5	3	2.3	10	5.5	12	2.4	11	2.2
Herpes Zoster	16	1.2	1	0.3	0	0.0	2	1.5	2	1.1	9	1.8	8	1.6
Acne	13	1.0	4	1.3	2	1.0	2	1.5	3	1.6	5	1.0	5	1.0
Special Senses														
Conjunctivitis	21	1.6	9	2.9	2	1.0	4	3.0	1	0.5	17	3.4	6	1.2
Urogenital System														
Cystitis	18	1.3	6	1.9	1	0.5	2	1.5	2	1.1	8	1.6	9	1.8
Dysuria	17	1.3	6	1.9	4	1.9	1	0.8	1	0.5	2	0.4	5	1.0
Increased Frequency	16	1.2	6	1.9	4	1.9	0	0.0	2	1.1	0	0.0	6	1.2

TABLE 8. Patient Disposition: X-ray by clinical measurements
(US301)

Frequency Percent Row Pct Col Pct	com	loe	oth	saf	Total
NBNF	26	42	22	30	120
	5.42	8.75	4.58	6.25	25.00
	21.67	35.00	18.33	25.00	
	10.92	30.66	57.89	44.78	
YBNF+YBYF	212	95	16	37	360
	44.17	19.79	3.33	7.71	75.00
	58.89	26.39	4.44	10.28	
	89.08	69.34	42.11	55.22	
Total	238	137	38	67	480
	49.58	28.54	7.92	13.96	100.00

com=completer, loe=lack of efficacy, oth=other reason, saf=safety

Table 9. Patient Disposition: X-ray by clinical measurements in LEF group
(US301)

Frequency Percent Row Pct Col Pct	com	loe	oth	saf	Total
NBNF	11	9	6	18	46
	6.04	4.95	4.40	9.89	25.27
	23.91	19.57	17.39	39.13	
	11.46	29.03	53.33	45.00	
YBNF+YBYF	85	22	7	22	136
	46.70	12.09	3.85	12.09	74.73
	62.50	16.18	5.15	16.18	
	88.54	70.97	46.67	55.00	
Total	96	31	15	40	182
	52.75	17.03	8.24	21.98	100.00

com=completer, loe=lack of efficacy, oth=other reason, saf=safety

Table 10. Patient Disposition: X-ray by clinical measurements in MTX group
(US301)

Frequency Percent Row Pct Col Pct	com	loe	oth	saf	Total
NBNF	9	14	9	8	40
	5.00	7.78	5.00	4.44	22.22
	22.50	35.00	22.50	20.00	
	8.57	31.82	64.29	47.06	
YBNF+YBYF	96	30	5	9	140
	53.33	16.67	2.78	5.00	77.78
	68.57	21.43	3.57	6.43	
	91.43	68.18	35.71	52.94	
Total	105	44	14	17	180
	58.33	24.44	7.78	9.44	100.00

com=completer, loe=lack of efficacy, oth=other reason, saf=safety

Table 11. Patient Disposition: X-ray by clinical measurements in Placebo group (US301)

Frequency Percent Row Pct Col Pct	com	loe	oth	saf	Total
NBNF	6	19	5	4	34
	5.08	16.10	4.24	3.39	28.81
	17.65	55.88	14.71	11.76	
	16.22	30.65	55.56	40.00	
YBNF+YBYF	31	43	4	6	84
	26.27	36.44	3.39	5.08	71.19
	36.90	51.19	4.76	7.14	
	83.78	69.35	44.44	60.00	
Total	37	62	9	10	118
	31.36	52.54	7.63	8.47	100.00

com=completer, loe=lack of efficacy, oth=other reason, saf=safety

Table 12. Patient Disposition: X-ray by clinical measurements (MN301)

Frequency Percent Row Pct Col Pct	com	loe	oth	saf	Total
NBNF	28	3	5	7	43
	7.93	0.85	1.42	1.98	12.18
	65.12	6.98	11.63	16.28	
	12.17	5.66	21.74	14.89	
YBNF	17	14	12	31	74
	4.82	3.97	3.40	8.78	20.96
	22.97	18.92	16.22	41.89	
	7.39	26.42	52.17	65.96	
YBYF	185	36	6	9	236
	52.41	10.20	1.70	2.55	66.86
	78.39	15.25	2.54	3.81	
	80.43	67.92	26.09	19.15	
Total	230	53	23	47	353
	65.16	15.01	6.52	13.31	100.00

com=completer, loe=lack of efficacy, oth=other reason, saf=safety

Table 13. Patient Disposition: X-ray by clinical measurements in LEF group
(MN301)

Frequency Percent Row Pct Col Pct	com	loe	oth	saf	Total
NBNF	10	0	3	3	16
	7.69	0.00	2.31	2.31	12.31
	62.50	0.00	18.75	18.75	
	10.42	0.00	50.00	16.67	
YBNF	5	4	3	13	25
	3.85	3.08	2.31	10.00	19.23
	20.00	16.00	12.00	52.00	
	5.21	40.00	50.00	72.22	
YBYF	81	6	0	2	89
	62.31	4.62	0.00	1.54	68.46
	91.01	6.74	0.00	2.25	
	84.38	60.00	0.00	11.11	
Total	96	10	6	18	130
	73.85	7.69	4.62	13.85	100.00

com=completer, loe=lack of efficacy, oth=other reason, saf=safety

Table 14. Patient Disposition: X-ray by clinical measurements in placebo group
(MN301)

Frequency Percent Row Pct Col Pct	com	loe	oth	saf	Total
NBNF	9	3	1	0	13
	9.89	3.30	1.10	0.00	14.29
	69.23	23.08	7.69	0.00	
	17.65	10.34	14.29	0.00	
YBNF	5	5	4	2	16
	5.49	5.49	4.40	2.20	17.58
	31.25	31.25	25.00	12.50	
	9.80	17.24	57.14	50.00	
YBYF	37	21	2	2	62
	40.66	23.08	2.20	2.20	68.13
	59.68	33.87	3.23	3.23	
	72.55	72.41	28.57	50.00	
Total	51	29	7	4	91
	56.04	31.87	7.69	4.40	100.00

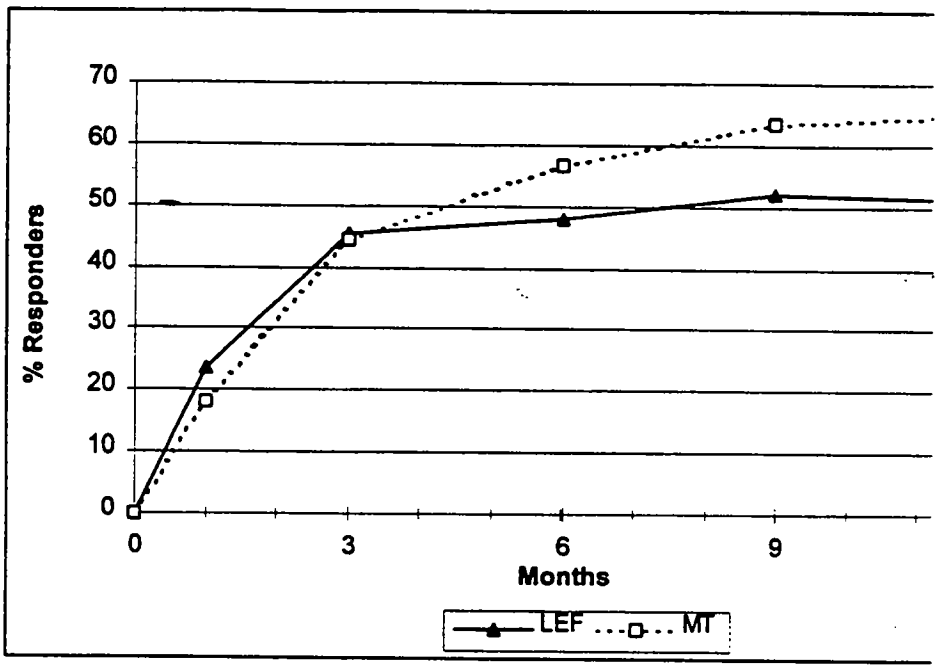
com=completer, loe=lack of efficacy, oth=other reason, saf=safety

Table 15. Patient Disposition: X-ray by clinical measurements in SSZ group
(MN301)

Frequency Percent Row Pct Col Pct	com	loe	oth	saf	Total
NBNF	9	0	1	4	14
	6.82	0.00	0.76	3.03	10.61
	64.29	0.00	7.14	28.57	
	10.84	0.00	10.00	16.00	
YBNF	7	5	5	16	33
	5.30	3.79	3.79	12.12	25.00
	21.21	15.15	15.15	48.48	
	8.43	35.71	50.00	64.00	
YBYF	67	9	4	5	85
	50.76	6.82	3.03	3.79	64.39
	78.82	10.59	4.71	5.88	
	80.72	64.29	40.00	20.00	
Total	83	14	10	25	132
	62.88	10.61	7.58	18.94	100.00

com=completer, loe=lack of efficacy, oth=other reason, saf=safety

Figure 3 MN302 ACR Responders Over Time



APPEARS THIS WAY
ON ORIGINAL

Appendix C

Analyses B (requested by medical officer)

1. A randomly ordered subset of non-missing baseline and change-from-baseline x-ray scores were created for each of the two treatment groups (Leflunomide and Placebo) and each of the two studies (MN301 and US301).

2. Each time a missing change score was encountered, a new change score and corresponding baseline value were sequentially selected as a substitute from the randomly ordered subset of the same treatment group. Thus, selection was made without replacement.

3. Treatment group means and p-values from the resulting ANCOVA analysis were recorded.

4. The minimum mean significant difference (i.e. the smallest difference which would still result in $p \leq 0.05$) between the overall Leflunomide cohort and the overall placebo cohort (i.e. including the randomly sampled values) was calculated. Then, the corresponding mean difference between the Leflunomide missing value cohort and the placebo missing value cohort was calculated. This was used to estimate the "minimum mean difference for the missing value cohort".

5. The estimated value for the minimum mean difference for the missing value cohort was used as a starting point to rerun the model. Each randomly sampled value in the Leflunomide missing value cohort was increased by ½ of the estimated mean difference and each randomly sampled value in the placebo missing value cohort was decreased by the same amount.

6. The analysis was rerun with these adjusted values for the missing value cohorts to determine the p-value. This step was iterated, increasing the increment/decrement on each run (by 0.1 on runs 1-10, by 0.15 on runs 10-15 and by .20 on runs 16-30), to determine the boundary value (Δ) for the increment/decrement such that $p=0.05$ (i.e., the iteration continued until the p-value exceeded $p=0.05$; the boundary value was taken as the increment/decrement from the previous run where $p \leq 0.05$).

7. The above steps were repeated 100 times.

STATISTICAL REVIEW AND EVALUATION
(Carcinogenicity Review)

JUN 22 1998

NDA #: 20-905

APPLICANT: Hoechst Marion Roussel, Inc.

NAME OF DRUG: Arava™ Tablets (leflunomide)

DOCUMENTS REVIEWED: Volumes 1.44-1.48 (Mouse Study) and 1.49-1.55 (Rat Study) of NDA 20-905. Data on CD-ROM supplied by the sponsor.

REVIEWING PHARMACOLOGIST: Asoke Mukherjee, Ph.D. (HFD-550).

I. BACKGROUND

In this NDA submission, two animal carcinogenicity studies (Study 2822 in mice and Study 2779 in rats) were included. These two studies were conducted to investigate whether Arava affects tumor incidence in mice and rats when administered orally by stomach tube at some selected dose levels for up to 24 months.

II. THE MOUSE STUDY (Study 2822)

IIa. Design

Groups of 50 male and 50 female CD-1 mice were treated with Arava in concentrations of 0 (control 1), 0 (control 2), 1.5 (low), and 5 (medium) mg/kg; and, groups of 70 male and 70 female CD-1 mice were treated with 15.0 (high) mg/kg in the feed for up to 24 months.

IIb. Reviewer's Analysis

This reviewer independently performed analyses on the survival and the tumor data provided by the sponsor on a CD-ROM. For survival data analysis, methods described in the papers by Cox (1972) and Gehan (1965) were used. The tumor data were analyzed using the methods described in the paper of Peto et al. (1980) and the method of exact permutation trend test developed by the Division of Biometrics, FDA. The results are included in the Appendix.

Survival Analysis: The purpose of the survival analysis was two-fold:
(1) To examine the differences in the survival distributions among different dose groups (referred to as the test of homogeneity), and

(2) To determine the significance of a positive linear trend in proportions of deaths with respect to dose levels (called the test of linear trend).

For the theoretical background of these analyses, please refer to Lin et al. (1994) and Thomas et al. (1976).

The following results for survival analysis are contained in the Appendix:

- Tables 1a and 1b summarize the intercurrent mortality data for the male and female mice respectively. For the male mice, in the time-interval of 92-105 weeks, there appears to be an increased mortality in the high dose group as compared to other dose groups. For the female mice, in the time-interval of 92-105 weeks, more animals died in high dose group than in the other groups.

- Figures 1a and 1b depict the Kaplan-Meier survival distributions for males and females respectively. For the male mice, after 60 weeks, there appears to be an increased mortality in the high dose group when compared to the other doses. For the female mice, the curves for different dose groups (except the control 1 group) intertwine each other suggesting that there is no significant difference between their survival patterns. The test of homogeneity yields significant results for the male mice and non-significant results for the female mice (Tables 2a and 2b in the Appendix).

- Tables 2a and 2b display the p-values of the test of homogeneity and of positive linear trends for males and females, respectively, using the Cox test and the generalized Kruskal-Wallis (Gehan) test. It is well known that the Kruskal-Wallis test gives more weight to early differences in death rates between groups than the Cox test which gives equal weight to all deaths. The test of homogeneity and the test of linear trend yield significant results for the male mice which confirm the graphical findings of Figure 1a.

Tumor Analysis: The tumor data analysis was performed to detect, for a selected tumor type in a selected organ/tissue, the significance of a positive linear trend in the proportions of discovered tumors with respect to dose levels. The tumor types were classified as fatal and non-fatal. Table 3 (Part I) displays selected organs and organ codes. Table 3 (Part II) displays tumors and tumor codes.

Following Peto et al. (1980), this reviewer applied the death-rate method and the prevalence method to fatal and non-fatal tumors respectively. For tumors that caused death for some, but not all animals, a combined analysis was performed. The exact permutation trend test was used to calculate the p-values of all trend tests, except when the tumor was found in both categories, in which case the continuity corrected normal test was used. The scores used were 0, 0, 1.5, 5, and 15 for the control 1, control 2, low, medium, and high dose groups.

respectively. This was done in order to reflect the actual dose levels of 0, 0, 1.5, 5, and 15 mg/kg of Arava. The time-intervals used were 0-52, 53-78, 79-91, 92-105, 106 and beyond for males and females.

The tumor analysis results are displayed in the Appendix. Tables 4a and 4b describe the p-values for the test of trend based on the tumor data for males and females, respectively. The rule proposed by Haseman (1983) could be used to adjust for the effect of multiple testings in pairwise comparisons. A similar rule proposed by Lin and Rahman (1998) for trend tests was used in this review. This rule for trend tests says that in order to keep the false-positive rate at the nominal level of approximately 0.1, tumor types with a spontaneous tumor rate of 1% or less (rare tumors) should be tested at a 0.025 significance level, otherwise (for common tumors) a 0.005 significance level should be used.

On the basis of the rule for trend tests described above, the following significant linear dose tumor-trends were indicated for the male mice.

The number of males with malignant lymphoma for the haemolymphoret. sys. for various dose groups is described below (Table 4a).

Male Mice			Tumor Rate					Trend Test p-value
Organ	Tumor Name	Tumor Type	Control 1 N=50	Control 2 N=50	Low N=50	Medium N=50	High N=70	
Haemolymphoret. Sys.	Lymphoma Malignant	Mixed	3	5	2	4	12	0.0017

IIc. Additional Statistical Analyses

At the request of the reviewing pharmacologist, three additional tumor analyses were performed for both sexes:

Analysis #1: all adenoma and carcinoma together for specific organs,
Analysis #2: all lymphomas together, and
Analysis #3: all sarcomas and hemangiosarcomas together.

The tumor analysis results for all three analyses are displayed in Table 5a (for males) and 5b (for females). The significant results are produced below.

The number of males with malignant lymphoma for the haemolymphoret. sys. for various dose groups is described below (Table 5a).

Male Mice			Tumor Rate					Trend Test p-value
Organ	Tumor Name	Tumor Type	Control 1 N=50	Control 2 N=50	Low N=50	Medium N=50	High N=70	
Haemolymphoret. Sys.	Lymphoma Malignant	Mixed	3	5	2	4	12	0.0017

The number of females with adenoma bronchiolo-alveolar and carcinoma branchiolo-alveolar combined for the lungs for various dose groups is described below (Table 5b).

Female Mice			Tumor Rate					Trend Test p-value
Organ	Tumor Name	Tumor Type	Control 1 N=50	Control 2 N=50	Low N=50	Medium N=50	High N=70	
Lungs	Adenoma bronchiolo-alveolar and Carcinoma bronchiolo-alveolar	Mixed	3	1	7	9	15	0.0025

IId. Summary of Mouse Study (Study 2822)

The results of the statistical tests show that, for the male mice, there is an increased mortality in the high dose group when compared to the other doses.

For male mice, a significant linear dose tumor-trend was indicated for malignant lymphoma for the haemolymphoret. sys. For female mice, a significant linear dose tumor-trend was indicated for adenoma bronchiolo-alveolar and carcinoma bronchiolo-alveolar for the lungs.

III. THE RAT STUDY (Study 2779)

IIIa. Design

Groups of 50 male and 50 female Wistar rats were treated with Arava in concentrations of 0 (control 1), and 0 (control 2) mg/kg; groups of 60 male and 60 female Wistar rats were treated with Arava in concentrations of 0.5 (low), and 1.25 (medium) mg/kg; and, groups of 80 male and 80 female Wistar rats were treated with Arava in concentrations of 3.0 (high), and 6.0 (maximum) mg/kg orally by stomach over 24 months. Mortality increased markedly in the 6 mg/kg group after 1 year of treatment, especially in male rats.

IIIb. Reviewer's Analysis

This reviewer independently performed analyses on the survival and the tumor data provided by the sponsor on a CD-ROM. For survival data analysis, methods described in the papers by Cox (1972) and Gehan (1965) were used. The tumor data were analyzed using the methods described in the paper of Peto et al. (1980) and the method of exact permutation trend test developed by the Division of Biometrics, FDA. The results are included in the Appendix.

Survival Analysis: The purpose of the survival analysis was two-fold:
(1) To examine the differences in the survival distributions among different dose groups (referred to as the test of homogeneity), and
(2) To determine the significance of a positive linear trend in proportions of deaths with respect to dose levels (called the test of linear trend).

For the theoretical background of these analyses, please refer to Lin et al. (1994) and Thomas et al. (1976).

The following results for survival analysis are contained in the Appendix:

- Tables 6a and 6b summarize the intercurrent mortality data for the male and female rats respectively. For the male rats, mortality increased markedly in the 6 mg/kg group after 1 year of treatment; all animals in the group died before week 85. For the female rats, in the time-interval of 92-105 weeks, there appears to be an increased mortality in the low dose group as compared to other dose groups.

- Figures 2a and 2b depict the Kaplan-Meier survival distributions for males and females respectively. For the male rats, after 60 weeks, there appears to be an increased mortality in the maximum dose group when compared to the other doses. For female rats, the curves are intertwined. The test of homogeneity yields significant results for the male rats and non-significant results for the female rats (Table 7a and 7b in the Appendix).

- Tables 7a and 7b display the p-values of the test of homogeneity and of positive linear trends for males and females using the Cox test and the generalized Kruskal-Wallis (Gehan) test. It is well known that the Kruskal-Wallis test gives more weight to early differences in death rates between groups than the Cox test which gives equal weight to all deaths. The test of homogeneity and the test of linear trend yield significant results for the male rats.

Tumor Analysis: The tumor data analysis was performed to detect, for a selected tumor type in a selected organ/tissue, the significance of a positive linear trend in the proportions of discovered tumors with respect to dose levels. The tumor types were classified as fatal and non-fatal. Table 8 (Part I) displays selected organs and organ codes. Table 8 (Part II) displays tumors and tumor codes.

Following Peto et al. (1980), this reviewer applied the death-rate method and the prevalence method to fatal and non-fatal tumors respectively. For tumors that caused death for some, but not all animals, a combined analysis was performed. The exact permutation trend test was used to calculate the p-values of all trend tests, except when the tumor was found in both categories, in which case the continuity corrected normal test was used. The scores used were 0, 0, 0.5, 1.25, 3.00, and 6.00 for control 1, control 2, low, medium, and high dose groups respectively. This was done in order to reflect the actual dose levels of 0, 0, 0.5, 1.25, 3.00, and 6.00 mg/kg of Arava. The time-intervals used were 0-52, 53-78, 79-91, 92-105, 106 and beyond for males and females.

The tumor analysis results are displayed in the Appendix. Tables 9a and 9b describe the p-values for the test of trend based on the tumor data for males and females, respectively. The rule proposed by Haseman (1983) could be used to adjust for the effect of multiple testings in pairwise comparisons. A similar rule proposed by Lin and Rahman (1998) for trend tests was used in this review. This rule for trend tests says that in order to keep the false-positive rate at the nominal level of approximately 0.1, tumor types with a spontaneous tumor rate of 1% or less (rare tumors) should be tested at a 0.025 significance level, otherwise (for common tumors) a 0.005 significance level should be used.

On the basis of the rule for trend tests described above, the following significant linear dose tumor-trends were indicated for the female rats.

The number of females with polyp\glandular for the uterus for various dose groups is described below (Table 9b).

Female rats			Tumor Rate						Trend Test p-value
Organ	Tumor Name	Tumor Type	Control 1 N=50	Control 2 N=50	Low N=60	Medium N=60	High N=80	Maximum N=80	
Uterus	Polyp\glandular	Incidental	0	0	0	1	0	4	0.0121

IIIc. Summary of Rat Study (STD03592)

The results of the statistical tests show that, for the male rats, there is an increased mortality in the maximum dose group when compared to the other doses.

For female rats, a significant linear dose tumor-trend was indicated for polyp\glandular for the uterus.

IV. SUMMARY

Mouse Study (Study 2822)

The results of the statistical tests show that, for the male mice, there is an increased mortality in the high dose group when compared to the other doses.

For male mice, a significant linear dose tumor-trend was indicated for malignant lymphoma for the haemolymphoret. sys. For female mice, a significant linear dose tumor-trend was indicated for adenoma bronchiolo-alveolar and carcinoma bronchiolo-alveolar for the lungs.

Rat Study (Study 2779)

The results of the statistical tests show that, for the male rats, there is an increased mortality in the maximum dose group when compared to the other doses.

For female rats, a significant linear dose tumor-trend was indicated for polyp\glandular for the uterus.

Baldeo K. Taneja, Ph.D.
Mathematical Statistician (Biomed)

Concur: Dr. Lin 6/22/98

Dr. Huque Huque 6/22/98

cc: Archival NDA 20-905
HFD-550/Mukherjee, Cook, Division File
HFD-725/Taneja, Lin, Huque, Division File, Chron.