MATERIALS AND METHODS

Study Design The study was designed as a two arm, randomized, double blind, placebo controlled clinical trial of UDCA plus methotrexate (MTX) versus UDCA plus placebo. The primary measure of treatment outcome was transplant free survival as measured by the distribution of time to transplant or death from all causes, whichever comes first. Secondary endpoints included comparison of treatment arms with respect to overall survival, time to clinical decompensation (development of ascites, hepatic encephalopathy, variceal bleeding, transplant, or death), development of varices, changes in biochemical tests, liver histology, and symptomatolgy and sense of well being. The study design was reviewed and approved by the Institutional Review Boards at each of the clinical centers.

Patient Selection Clinical investigators at 12 geographically diverse clinical centers in the U.S.A. (see appendix) screened 535 patients with PBC for possible entry into our treatment trial. During this screening period, patients' clinical records were reviewed and various clinical, laboratory, radiology, and pathology tests were performed to assure (1) that patients would satisfy our inclusion criteria for PBC; (2) that they had not already demonstrated exclusion criteria with would keep them from qualifying for the methotrexate/placebo phase, and (3) if eligible for the trial, that they would proceed to the next steps.

The intention was to study the effect of methotrexate on the progression of PBC in 20 to 69 year old patients of either sex and any race and with only moderately advanced disease at study entry. For documentation of sufficiently advanced PBC, patients were to have had a diagnosis of chronic cholestatic liver disease of at least 6 months duration, documented history of a positive antimitochondrial antibody test and alkaline phosphatase levels at least 1.5 times the upper limit of normal at their clinical center, and a liver biopsy within the 6 months prior to randomization (and on UDCA at least 6 months) with histologic findings compatible with the diagnosis of PBC. To be judged adequate for staging of disease, the liver biopsy must have been at least 2 cm long if cirrhosis was not detected. Asymptomatic patients must have had a histologic stage greater than

Stage I using the Ludwig classification. Patients could not have markedly advanced PBC, and thus patients ever having a history of serum bilirubin of 3.0 mg% or greater, a serum albumin less than 3.0 mg%, or a history of ascites, hepatic encephalopathy, or variceal bleeding were not eligible for randomization. At screening 393 of the 535 patients were judged to meet the defined inclusion criteria.

Patients were excluded from the study if they had clinical, serologic, or histologic evidence of liver disease of other etiology, had a history of alcohol abuse within the two years prior to study enrollment, were treated with immunosuppressive agents, rifampin, or dilantin in the months preceding randomization, had a history of malignant disease, were HIV positive, had other major illnesses that could limit life span, or were pregnant or unwilling to use adequate forms of birth control to avoid pregnancy. Of the 385 patients meeting the screening inclusion and exclusion criteria, 300 patients progressed to a pre-entry evaluation phase during which they were treated with UDCA alone at a dose of approximately 15 mg/kg/day. At the end of this UDCA phase, the patients were again screened for meeting the inclusion and exclusion criteria given above, as well as for having an acceptable hematologic profile, adequate renal and pulmonary function, no radiologic or ultrasound evidence of biliary obstruction, and a liver biopsy within the last 6 months consistent with a diagnosis of primary biliary cirrhosis.

Randomization Between January, 1994 and March, 1998, 265 subjects who signed informed consent documents were randomized with equal probability in a double blind fashion to receive UDCA plus MTX (132 patients) or UDCA plus placebo (133 patients). Ten patients who failed to meet only one of the eligibility criteria were reviewed by the study principal investigator and were judged suitable for randomization despite no liver biopsy within the last 6 months (1 patient on the MTX arm whose biopsy was 9.2 months prior to randomization and 1 patient on the placebo arm whose biopsy was 9.7 months prior to randomization), a percent predicted DLCO of 45% (1 patient on the MTX arm), or creatinine clearance between 50 and 60 ml / min / 1.73 meter sq (5 patients on the MTX arm and 2 patients on the placebo arm). In addition, in later, post-randomization review of medical records, two patients on the MTX arm were found to have had previous bilirubin measurements of 5.3 and 7.9 mg/dl.

Randomization was stratified according to histologic stage of liver disease according to the classification of Ludwig, et al. and as read by pathologists at the individual clinical centers. Of 126 patients initially reported to be stage 1 or 2, 62 were randomized to receive MTX and 64 to receive placebo, and of 139 patients initially reported to be stage 3 or 4, 70 were randomized to receive MTX and 69 to receive placebo. Two patients judged as stage 3 by the pathology reports at their respective clinical centers were erroneously randomized with the stage 1-2 group. In keeping with the principles of analysis by intention-to-treat, these patients were kept in the stage 1-2 group for all statistical analyses.

Drug Treatment All patients received UDCA in 300 mg capsules provided by Ciba-Geigy and, subsequently, Novartis, in a single dose of 13-15 mg/kg/day taken orally at bedtime. In addition, methotrexate or its placebo, provided as 2.5 mg tablets by Lederle Laboratories initially, then Wyeth-Ayerst Laboratories, was administered orally once a week in a single dose at bedtime. The initial dose was one-half of the maximum dose and was increased each month by 2.5 mg per week to the maximum dose of 15 mg per 1.73 m^2 body surface area, with a maximum dose of 20 mg per

week, provided toxicity was absent or mild. Patients taking cholestyramine or colestipol were asked to take the medication at least 2 hours before or after intake of UDCA and methotrexate or its placebo. Patients were to be continued on UDCA along with methotrexate or its placebo until the closure of the study despite progression of disease unless liver transplantation or death without transplantation ensued, drug toxicity necessitated withdrawal, the patient developed a cancer, or voluntary withdrawal ensued.

Modification of Methotrexate Dose Because there is no current evidence that UDCA affects blood elements or induces side effects other than diarrhea in a small number of patients, the development of a cytopenia, mucositis, significant nausea or anorexia were initially considered to be related to methotrexate, and methotrexate dose was altered in accord with the following rating for the common side effects and bone marrow toxicity of methotrexate (Table 2). Toxicity was rated as either mild (acceptable), moderate (requiring alteration of dose), or severe (requiring discontinuation of therapy).

For moderate toxicity, weekly dosage was reduced by a quarter or a third, and the toxicity was monitored weekly until resolved. The dosage of methotrexate was then increased by 2.5 mg per week until a dose of 2.5 mg less than the original toxic dose was reached, provided toxicity did not recur. Return to the original dose at which toxicity occurred was attempted carefully.

For severe toxicity, methotrexate was stopped completely while the toxic reaction was being managed. Gastrointestinal and hematologic findings usually improve fairly rapidly. Once better, methotrexate was to be restarted at half the toxic dose, and then increased 2.5 mg per week at monthly intervals provided toxicity did not recur, until a weekly dose 2.5 mg less than the original toxic dose was reached. If recurrent toxicity was not observed, cautious increase to full dose was attempted.

If severe toxicity did not improve within a week or two, or if it was judged to be life threatening, leucovorin factor was to be administered at a dose of 5 mg po or IV every 12 hours for at least 48 hours in order to facilitate recovery. Controversy exists about duration and dosage of

leucovorin factor in this type of toxicity.

Other reasons listed in the protocol for decreasing the dose or stopping methotrexate included the appearance of allergic reactions, severe skin rash, pulmonary symptoms or chest x-ray findings suggestive of pulmonary fibrosis, severe exacerbation of liver disease (as judged by liver biopsy histology or by prothrombin time, serum bilirubin and/or albumin levels), and worsening of renal function. Methotrexate was to be withdrawn if evidence of alcohol abuse arose or if the patient became pregnant or would no longer practice birth control. Study medication was stopped in patients developing a cancer.

Dose modifications could be carried out without the local investigator breaking the medication code, since in all instances dosage would be temporarily reduced or stopped. Nevertheless, when deemed necessary by our external safety monitors, the treatment code could be broken for their use in assisting with the management of our patients.

Schedule of Patient Visits and Investigations According to the study protocol, patients were to be seen and have blood drawn at weeks 2 and 4, then monthly for the first 6 months, bimonthly for the next 6 months, then at 3 month intervals for the duration of the study. Blood was to be drawn one week after the preceding dose of methotrexate and on the day of, but preceding the next dose of methotrexate. Symptoms of liver disease and of potential toxicity were to be assessed at each visit by history and with the aid of a diary. At each visit, blood was to be obtained for a CBC, differential and platelet count; at the monthly and each later visit for bilirubin, alkaline phosphatase, AST and ALT; at 3 monthly intervals for total protein and albumin, and at 6 monthly intervals for prothrombin time (INR). Complete histories, physical examinations, chest x-rays and pulmonary function studies, including measurements of diffusing capacity (DLCO) were to be obtained at least annually. Patients were to have a liver biopsy and upper endoscopy after 24 months on methotrexate or its placebo, and subsequently at additional intervals of 2 years.

Evaluation of Compliance Patients were given known quantities of medicine at appropriate intervals and instructed in how to keep a log of medicine intake. The log was checked, and unused medicine

counted at appropriate return visits, and before a new supply of medicine was given to the patient. The log and pill counts were kept in the permanent record for each patient.

Evaluation of Adverse Experiences The adverse experiences reported by patients during their study visits were grouped within broad categories defined by organ system. In addition, adverse experiences were categorized across all organ systems infections, bleeding events, neoplasms events, and cancers. Serious adverse experiences occurring at any clinical center were reported to a central committee monitoring such events.

Evaluation of Treatment Response The primary and several secondary measures of treatment outcome were based on the distribution of time to treatment failure as defined by a hierarchy of clinical and subclinical outcomes as listed in Table 4. Times to death, transplant, activation for transplant, and clinical deterioration as defined by development of variceal bleeding, hepatic encephalopathy, ascites, or disabling pruritus were obtained from the routine follow-up of patients. Subclinical deterioration was defined as a doubling of serum bilirubin from baseline to at least 2.5 mg/dl, a decrease in serum albumin to a level less than 2.5 g/dl, or an increase in PTINR to 1.3. In order to be judged a subclinical deterioration, the corresponding threshold must have been exceeded on two consecutive clinic visits.

Liver histology was evaluated as the average stage and fibrosis scores on liver biopsies obtained every two years and scored independently by a panel of 5 pathologists in a central core. Development of varices was evaluated by endoscopies performed every two years according to the study protocol. Because patients with varices at screening were eligible for randomization, only a subset of the trial participants were evaluable for the endpoint of development of new varices.

Patients who terminated study treatment prior to liver transplant or death were encouraged to continue all regular clinic visits, and patients who agreed were thus considered evaluable for all measures of treatment response. Some patients declined to have further biopsies, endoscopies, and/or serum chemistries measured, but were willing to be followed for clinical events, and these patients are considered fully evaluable for all endpoints that could be observed without invasive

procedures. Patients who withdrew consent for all further follow-up contribute information only up to the time of their withdrawing their consent to be studied, although as described in the statistical methods, exploratory analyses imputed missing measurements for these patients.

Monitoring of the Clinical Trial The accruing data were monitored on a semiannual basis by an independent Data Safety and Monitoring Board (DSMB), who reviewed the data for safety, as well as making recommendations for early termination of the trial for reasons of demonstrated efficacy of methotrexate over placebo or for reasons of the inability to demonstrate a statistically credible, clinically important benefit. In making such a recommendation for early termination of the trial, the DSMB was guided by a formal stopping rule as described in Statistical Methods. Study procedures called for the DSMB to remain blinded to treatment assignment, unless specific safety issues arose that necessitated their becoming unblinded. Hence, at each of their meetings, the DSMB was provided with statistics broken down by study arm, but labeled only by treatment A or B. In the conduct of the study, the DSMB remained blinded until the formal interim analysis at which study termination was recommended.

In their review of the data on October 31, 2002, the DSMB recommended termination of the study for reasons of futility: The estimate of treatment effect at that interim analysis was such that the hypothesized treatment effect for which the study was powered had been ruled out with high confidence. On November 1, 2002, clinical centers began advising patients to discontinue study drug. At the time of such notification, patients were asked to continue clinical visits for monitoring of clinical progression and laboratory measurement of blood chemistries. The more invasive study procedures of liver biopsy and endoscopy were no longer performed. It was hoped that such monitoring would continue for at least one year post discontinuation of study drug to observe the effect of withdrawal of MTX therapy.

Statistical Analysis All statistical analyses of treatment effect were performed according to the principle of intent to treat. The distribution of time to treatment failure was estimated using Kaplan-Meier curves for each of the hierarchical definitions of treatment failure of 1) time to the earliest of

death or liver transplantation, 2) time to the earliest of death, liver transplantation, or clinical deterioration as defined by development of hepatic encephalopathy, variceal bleeding, or ascites, and 3) time to the earliest of death, liver transplantation, clinical deterioration, development of disabling pruritus, or an observation at two successive clinic visits of subclinical progression defined by a doubling from baseline of serum bilirubin to at least 2.5 mg/dl, a decrease of serum albumin to less than 2.5 g/dl, or an increase in PTINR to 1.3. Patients still alive without failure at the time of study termination or their withdrawing consent were censored at the time of their last clinic visit. Unadjusted comparisons of the treatment arms with respect to time to treatment failure were made using the logrank statistic (65) stratified according to the stratification variable used in randomization. The Cox proportional hazards model (66) was used to compare the treatments while adjusting for prognostic variables. Such variables included the baseline values of serum bilirubin, serum albumin, age, and prothrombin time.

Comparisons of histologic stage were made using the t-test which allows for unequal variances, and the prevalence of varices was compared across treatment groups using the chi-squared test. Analyses of these endpoints adjusting for prognostic covariates were effected using linear regression and logistic regression, respectively. Comparisons of laboratory measures of bilirubin, albumin, or PTINR were similarly made using linear regression models which adjusted for baseline measurements and the randomization stratum. To account for the variation in time of exact time of measurement of the serum and blood chemistries, linear interpolation was used to estimate measurements at exact 3 monthly intervals. Analyses of all of the histology, endoscopy, and laboratory measurements were subject to possible informative censoring due to death, liver transplant, and study withdrawal. Exploratory analyses imputed values for such missing data using the “last one carried forward” approach, as well as methods based on extrapolating trends observed in the year prior to censoring. In this last imputation approach for laboratory measurements, any linear trend over the previous two years toward progressive disease (increasing serum bilirubin or PTINR or decreasing serum albumin) was used to impute the missing values through the end of the

interventional phase of the study, up to a maximum bilirubin of 34.4 mg/dl, a minimum albumin of 1.5 g/dl, or a maximum PTINR of 3.0. If no trend toward progressive disease was evident or if all measurements were within normal limits (serum bilirubin < 1.2 mg/dl, serum albumin > 3.0 g/dl, and PTINR < 1.2), the average measurement observed over the preceding two years was used.

Analysis of rates of adverse experiences were performed using the chi square test, with adjustment made for small sample sizes. Analysis of forced expiratory volume and DLCO were made using linear regression models which adjusted for baseline values as a covariate.

In order to examine possible “rebound” effects on stopping treatment, the bilirubin, albumin, and PTINR measurements were analyzed for the first year after stopping study drug for all patients who had taken study drug for at least four years, whether or not the date of stopping study drug coincided with the end of the interventional phase of the study. Linear interpolation between available measurements were used to estimate the laboratory values at exact 3 monthly intervals, and these values were then used to compute the change in laboratory measurement from the last value obtained while the patient was taking study drug. The mean changes in laboratory measurements after stopping study drug were compared across treatment arms using a linear regression model which adjusted for randomization stratum and the last measurement obtained on study drug.

Stopping Rule and Sample Size The clinical trial was conducted using a group sequential stopping rule. In November, 1997, prior to any formal interim analyses and when only two transplants and one death had been observed, the DSMB and investigators adopted a stopping rule based on a level .025 one-sided group sequential test with O’Brien-Fleming boundary relationships for the efficacy endpoint and a futility stopping boundary corresponding to a choice of boundary shape parameters P=0.8, R=0, A=0 within the unified family of group sequential designs (Kittelson JM and Emerson SS, A unifying family of group sequential test designs. *Biometrics* 55:874-882, 1999). Such a stopping rule preserves the nominal one-sided type I error at .025 (corresponding to a two-sided level 0.05 test) while allowing early stopping either for strong evidence of efficacy of

methotrexate or for the futility of further continuation of the trial in the absence of evidence for sufficient benefit due to methotrexate. It was planned to continue the trial until 50 primary events had been observed, unless stopping was recommended by our Data Safety Monitoring Board (DSMB) as guided by the specified stopping rule. Using the sampling plan suggested by this stopping rule, the planned maximal sample size of 50 transplant or death events would provide 80%, 90%, 95%, and 97.5% statistical power to detect a benefit of methotrexate therapy corresponding to hazard ratios of 0.45, 0.39, 0.35, and 0.32, respectively. Hence, a failure to reject the null hypothesis can be interpreted as having 95% confidence that the true treatment effect was less than would correspond to a 68% decrease in risk of transplant or death. At the time of study design, a hazard ratio of 0.32 was estimated to correspond roughly to five year transplantation-free survival rates of 83% on placebo and 94% on methotrexate.

Formal interim analyses were to be performed according to the funding cycle of this NIH sponsored study. At the first such competing renewal, the number of events was so small as to make early termination virtually impossible. In October, 2002, the second renewal period was drawing to a close with a total of 30 transplant or death events observed. Implementation of the stopping rule was effected using the constrained boundary approach (Burington BE and Emerson SS, Flexible implementation of group sequential stopping rules using constrained boundaries. *Biometrics* 59: 770-777, 2003). Computation of estimates of treatment effect, confidence intervals, and P values adjusted for the bias introduced by a stopping rule using the bias adjusted mean and the sample mean ordering (Emerson SS and Fleming TR, Parameter estimation following group sequential hypothesis testing. *Biometrika* 77:875-892, 1990). All reported P values are two-sided.