Biostat 517 Project: **Lung Toxicity Related to Methotrexate Treatment for Primary Biliary Cirrhosis** *Group 4*

Summary:
Primary biliary cirrhosis (PBC) is a chronic disease characterized by portal inflammation and the destruction of small bile ducts, primarily affecting middle-aged women. PBC can lead to liver failure, but none of the current treatments available have any effect on improving symptoms. To address the effectiveness of methotrexate treatment for PBC, we conducted a randomized, double-blind, placebo-controlled trial using survival and progression-free survival as primary endpoints. Patients, stratified by Ludwig stage, were randomized to placebo (133 patients) or methotrexate (131 patients). To determine whether treatment with methotrexate leads to impaired lung function, one of the most common adverse side-effects of methotrexate, we used the diffusing capacity of the lung for carbon monoxide (DLCO) measurements available for our participants. DLCO was measured at baseline and at least annually, but only the baseline, final and minimum values were available for this analysis. We tested for differences between the methotrexate and placebo groups using the mean decrease in DLCO from baseline to the minimum value as well as the mean decrease to the final DLCO value. In the methotrexate group, the average decrease to a minimum DLCO was 3.35 ml/min/mm Hg (95% CI = 2.65 – 4.03 ml/min/mm Hg), while in the placebo group, the average decrease was 3.17 ml/min/mm Hg (95% CI = 2.58 – 3.77 ml/min/mm Hg). Based on our analysis, we could not reject the null hypothesis that the groups do not differ with respect to the maximal changes in DLCO (two-sided p = 0.71). With 95% confidence, we conclude that true value for the mean difference between groups for change in DLCO falls between –0.734 to 1.082 ml/min/mm Hg based on the t-test. In conclusion, methotrexate is unlikely to cause a significant decrease in pulmonary gas exchange at the doses used to treat PBC in this study.

Comment: multicenter

Comment: Why only 131?

Comment: how long did the trial last?

Comment: while on study drug post randomization

Comment: This is a rather Bayesian sentence and not justified by your analysis. Also, I would reserve the word “significant” to mean statistical significance. Use “substantial”, “meaningful”, etc. instead.

***Background*:**
Primary biliary cirrhosis (PBC) is a chronic and progressive cholestatic disease of unknown etiology, characterized by portal inflammation and the immune-mediated destruction of small bile ducts within the portal area [1, 2]. The disease leads to decreased bile secretion and the retention of toxic substances within the liver, contributing to the development of cirrhosis and eventual liver failure. PBC is considered to be an autoimmune disease and is associated with the presence of well-characterized anti-mitochondrial antibodies. It predominantly affects middle-aged women from 35-60 years, with the peak of incidence occurring after 50 years of age.

Without a clear etiology for PBC, effective treatment strategies have remained elusive. The only drug currently FDA-approved for use is the bile acid ursodeoxycholic acid. This drug is associated with improvement in the immunological disturbances caused by PBC, but it has little effect on symptoms [3]. Methotrexate is an analogue of folic acid that inhibits cellular proliferation and also has anti-inflammatory and immunomodulating properties [4] and has been proposed as a potential therapeutic option. Historically, methotrexate has been used for the treatment of cancers and rheumatic diseases, such as rheumatoid arthritis, which are also believed to have an autoimmune component [5]. However, methotrexate has not been studied in large scale trials of PBC to date. To address the effectiveness of methotrexate treatment for PBC, we conducted a

randomized, double-blind, placebo-controlled trial using survival and progression-free survival as primary endpoints.

Despite its potential benefits, methotrexate can cause a variety of side-effects, including pulmonary toxicity. Methotrexate toxicity may manifest as hypersensitivity pneumonitis, bronchiolitis obliterans, or interstitial lung disease that decreases pulmonary gas exchange across the alveolar-capillary interface [6, 7]. The decrease in gas exchange is generally measured using the diffusing capacity of the lung for carbon monoxide (DLCO) [8]. Interestingly, pulmonary toxicity does not appear to be a dose dependent phenomenon, but an idiotypic reaction. To determine whether pulmonary toxicity occurred in this trial, we performed serial DLCO measurements in the participants. The primary aim of this analysis is to determine whether methotrexate treatment affected lung function as measured by DLCO.

Questions of Interest:
Our question of interest is whether treatment with methotrexate leads to impaired lung function as measured by DLCO. Our null hypothesis for the analysis, is that DLCO measurements in the placebo and methotrexate groups are similar. We used mean changes in DLCO as our primary outcome measure and determined whether the mean change in DLCO differed between the placebo and methotrexate groups. We also evaluated whether the mean DLCO measurements after randomization differed between the two groups. These evaluations provided a satisfactory answer to the question of interest.

***Source of the Data*:**
The DLCO data derives from a multicenter randomized trial of methotrexate treatment for PBC. Only patients with DLCO and FEV > 50% expected for age were included in the study. Patients, stratified by Ludwig stage, were randomized to placebo (133 patients) or methotrexate (132 patients). DLCO was measured at baseline and at least annually, but only the baseline, final and minimum values were available for this analysis. The final DLCO values were measured when the patient stopped taking the study treatment (placebo or methotrexate). At baseline, a variety of patient characteristics were measured to determine whether the groups were comparable (Table 1). Patients were followed for 46 to 3625 days. Out of the 133 patients in the placebo group, 8 had baseline DLCO measurements only and were missing minimal and final DLCO measurements. Out of the 132 patients in methotrexate group, 1 patient had no DLCO measurements and 12 patients had only baseline measurements only. Using a Chi squared analysis, the proportion of missing measurements did not differ between the treatment groups (p=0.33). Using Kaplan-Meier analysis, we could not reject the null hypothesis of no difference between the groups at 8 years(p=0.09). Similarly, disease progression did not differ between the two groups (p=0.21).

Comment: what period of time

Comment: Should be in Results

Comment: This is results. The treatment may have contributed to the missingness Also, lack of statistical significance does not mean lack of a problem, here.

Comment: with respect to what?

Comment: You did not have to comment on this, though it is actually good to. BUT…it takes an infinite sample size to establish equality. Hence, CI are very, very, very important.

***Statistical Methods*:**

Descriptive Statistics: When informative, univariate descriptive statistics, including mean, median, 95% confidence intervals, and data ranges, were calculated for all of the data included in the analysis. More detailed descriptive statistics were generated for the variables relevant to the problem as outlined in Tables 1 and 2. To assess for qualitative relationships between pairs of variables (bivariate analyses), we used visual assessment of correlation and scatter plots with LOWESS (locally weighted scatterplot smoother) curves. The LOWESS curve reveals the general trend of the relationship between the two variables. Quantitative correlation coefficients were calculated to characterize the strength of the relationship. These methods were used to reveal any linear and/or non-linear relationships between variables, and to identify potential confounders and effect modifiers – any variables that may distort the relationship between the predictors and the outcomes in question.

Comment: It is not clear where you used all this. I would not have listed these methods.

Missing data accounted for less than 10% of all DLCO measurements. Comparisons between subjects with missing and non-missing DLCO measurements did not reveal significant differences in death rate, time to progression or time to death. Based on these analyses, we determined that it was appropriate to exclude subjects with missing data from our analysis.

Comment: There is nothing in your data that can tell you whether your missing data is ignorable. That having been said, your approach here is reasonable. But I would soften the wording to "decided is was reasonable" or even better "saw no grave contraindication to"

Extreme outliers were not evident in this dataset, and the DLCO measurement distributions were slightly skewed to the right. For the purposes of subsequent analyses, we decided that that normality of the distributions was a fair assumption. To ensure that non-normal DLCO distributions were not affecting our results; however, we checked our results using the bootstrapping methods described below to generate tests of significance based on the actual distribution of DLCO data.

Comment: This did not do what you thought it did. You tested means with a ttest and medians with bootstrapping.

Hypothesis Testing: We performed the two-sided robust (not assuming equal variances) t-test to evaluate for significant differences in the mean DLCO measurements between the placebo and methotrexate groups. The null hypothesis throughout our analysis was that the treatment groups did not differ with respect to DLCO measurements. Since the baseline DLCO measurements were correlated with the minimum and final DLCO measurements, we also tested for differences between drug and placebo groups using the mean change between the minimum and baseline and between the final and baseline DLCO, thereby gaining precision. Tests were considered significant at the 5% level.

Comment: be careful with this statistical jargon. "robust" can mean different things. (but your "not assuming equal variances" made it clear)

Comment: the issue would be HIGH correlation (higher than 0.5—which there was)

Since the distributions were slightly skewed to the right, we also estimated the median difference in DLCO between treatment groups using bootstrapping. Bootstrapping is a method of estimating any statistic and its associated standard errors based on the actual distribution of the data. The main assumption is that the data is representative of the population. Given the samples sizes > 130 subjects, the study data should be reasonably representative of the population. The results of this analysis were expected to confirm the hypothesis tests using the t-test based on the means, indirectly verifying our assumption of normality.

Comment: I think this a bad choice for an analysis of toxicity. We generally expect few people to have toxicity. But if someone does, there is often a big outlier. We would want to detect that. Hence, you should use a summary measure that is sensitive to outliers.

Comment: we do not need an assumption of normally distributed data. We only need the sample means to be normally distributed. (I have such a hard time getting students to unlearn this malarkey about a need to verify the assumption of normality.)

In this manner, we estimated the median and the standard errors associated with the median of the minimum DLCO measurements (and the final DLCO measurements). Similarly, we used bootstrapping to estimate the median and the standard errors associated with the differences between minimum and baseline DLCO, as well as the final and baseline DLCO. We then used a hypothesis test based on the standard normal to determine possible significant differences. Again, the significance level was 5%.

All of the analyses above were performed using Intercooled Stata 9 statistical software (StatCorp, College Station, TX).

Results:

At randomization, the placebo and methotrexate groups were similar with respect to age, sex, weight, height, duration of disease, Ludwig pathologic stage, baseline laboratory values and baseline DLCO measurement (Tables 1 and 2). The distributions of the DLCO measurements (baseline, minimum and final) were roughly normally distributed without extreme outliers as discussed in the Methods section and outlined in Table 2. This allowed us to use the arithmetic means for our primary comparisons rather than using medians, geometric means or log transformations. Hypothesis tests were intentionally not performed on the between-group baseline variables [8].

Based on the t-test, the mean final DLCO was significantly lower than the mean baseline DLCO in both groups ($p<0.001$; Table 2 and Figure 1). Similarly, the mean minimum DLCO was lower than the mean baseline and final DLCO values ($p<0.001$). The three DLCO measurements (baseline, minimum, and final) were highly correlated, with a pairwise correlation coefficient of $r > 0.65$ for each variable. To gain precision, the change between DLCO at baseline and during the trial were calculated and used for subsequent comparisons. To maximize the sensitivity for detecting differences in DLCO between the groups, we calculated the maximal DLCO change during the study period by subtracting the baseline DLCO from minimum DLCO. The absence of a difference between the placebo and methotrexate groups for this change in DLCO would be the strongest evidence to support our null hypothesis of no effect of methotrexate on DLCO.

Table 3 describes differences between the placebo and methotrexate groups with respect to minimum, final and changes in mean DLCO. In the methotrexate group, the average decrease to the minimum DLCO was 3.35 ml/min/mmHg (95% CI = 2.65 - 4.03 ml/min/mmHg), while in the placebo group the average decrease was 3.17 ml/min/mmHg (95% CI = 2.58 – 3.77 ml/min/mmHg). Based on this analysis, we cannot reject the null hypothesis that the groups do not differ with respect to the maximal changes in DLCO (two-sided t test $p = 0.71$). With 95% confidence, we conclude that the true value for the mean difference between groups for change in DLCO falls between -0.734 to 1.082 ml/min/mmHg based on the t test. We reached the same conclusion when using bootstrap methods on the median differences to minimize the effects of non-normal distributions ($p = 0.17$).

Comment: who cares?

Comment: as these were results, the discussion should have been here

Comment: NO!! You should choose summary measures based on the scientific question.

Comment: No need to put this sentence here. You were correct in not doing it

Comment: You need to tell us about what happened to the patients in the trial. How long were they treated. How many dropped out? Differential missingness between treatment arms is potentially a result related to toxicity.

Comment: Okay, but this is why we have a control group. We know the patients got older over the course of the study.

Comment: Okay, but I would have just anticipated this and decided to look at the change. Well, really I would have adjusted for baseline in regression, but that is Biost 518 material.

Comment: It is also worthy of comment to look at patients experiencing a large drop. There may not be statistical significance, but the anecdotal evidence might be enough to make us lose interest in further study of MTX in PBC.

Comment: Are these differences clinically important? Did we have enough precision?

Comment: And, again, normal distn or not, if we are interested in the mean, we are interested in the mean. And the ttest is the way to assess it. Please unlearn your emphasis on normality. It is not important.

Further exploratory analyses revealed no significant differences between the groups for the average DLCO values (baseline, minimum and final) or the decrease in DLCO between baseline and final values. The DLCO measurements and differences were not correlated with age, duration of disease or time on study drug, so confounding by these variables despite randomization is not likely to have obscured a difference between the groups. Height was correlated with the DLCO measurements, but height did not differ between the treatment groups, so height is also unlikely to have confounded any association between DLCO and treatment. Finally, we did not find significant differences between the groups with respect to medians or geometric means for the DLCO values or differences.

Comment: time on study drug is a post randomization variable. Randomization does not protect you here.

You need to be careful how you assessed this "correlation". Did you do it within treatment arms? Also, a small correlation can still cause substantial confounding.

For the most part your discussion here is OK, though. I would have downplayed it all. Considered the possibility, but not written it up.

***Discussion*:**

From analysis of this randomized, controlled pharmacologic trial of methotreaxate in patients with PBC, there appears to be no significant difference in DLCO with the use of methotrexate as compared to placebo. It should be noted that this is only relevant for the dose given in this study and for patients with PBC and not other hepatic diseases. Further, whether DLCO alone is the best measure of methotrexate toxicity is not clear although pathophysiologically additional data, such as the degree of hypoxemia and alveolar-arterial oxygen difference, may provide better lung functional measures. The data analysis indicates that DLCO decreased in the setting of PBC over the study period, but the decrease was not different between treatment and placebo. Further data, including quality of life assessment, functional status, arterial blood gas, formal pulmonary functional studies, total lung volumes, and pulmonary imaging, may be informative.

Comment: So was this lack of statistical precision, or lack of a clinically important difference.

Limitations of the current analysis include 1) the analysis was performed on a truncated dataset from the larger database, and 2) no formal multivariate analysis was completed. The current dataset gives only 3 data points for DLCO (baseline, minimum value anytime during follow up, and measurement at discontinuation of treatment or the study). In order to better assess DLCO alterations over the study duration, sequential DLCO, time to the minimum DLCO and whether methotrexate therapy was altered due to the DLCO value (as noted in research design) would be beneficial additions to the drug safety analysis. Given that methotrexate treatment may have changed based on DLCO, we were unable to determine whether DLCO was a contributor to death; this would be an important clinical endpoint.

Comment: Why is this a limitation. A well designed study can be analyzed simply

Comment: And this would be of interest. Though the major toxicity causing change in MTX would be bone marrow suppression

Although a formal multivariate analysis was not performed, we are reasonably assured that there are no significant single variable confounders in the dataset based on our bivariate analyses. However, multiple variables added into a statistical model may discover certain subsets of patients at risk. Further, addition of certain cutoff values for hepatic disease, such as poor hepatic synthetic function as measured by an arbitrary value of albumin and/or PT, may identify additional at risk populations. We chose not to complete these analyses as these have not been prospectively validated. The model of end stage liver disease (MELD) or Child-Pugh scores cannot be calculated with the data provided. These may be considered for future analyses.

Comment: You probably mean a multiple regression analysis rather than multivariate. But this is statistical jargon.

Table 1. Characteristics of the placebo and methotrexate treatment groups

Characteristic	Placebo N	Placebo Mean or %	Placebo SD	Treatment N	Treatment Mean or %	Treatment SD
Age (years)	133	52.2	8.63	132	50.4	8.67
% Male	133	7.5%	-	132	7.6%	-
Weight (kg)	132	73.3	16.2	132	70.1	14.4
Height (cm)	132	163.7	8.1	132	163.7	7.9
Disease duration (days)	133	1339	1195	132	1280	1262
Stage						
I	24	18%	-	13	10%	-
II	40	30%	-	47	36%	-
III	47	35%	-	56	42%	-
IV	22	17%	-	16	12%	-
% splenomegaly	133	10.5%	-	131	8.4%	-
Bilirubin (mg/dl)	133	0.72	0.39	132	0.66	0.44
Albumin (g/dl)	133	4.0	0.34	132	4.0	0.35
Alkaline phosphatase (U/l)	133	245	187	132	243	146
ALT (U/l)	132	50	42	131	54	41
Prothrombin time (sec)	133	11.4	1.1	128	11.2	1.1
Cholesterol (mg/dl)	130	236	59	130	239	58
Platelets (1000 cells/mm^3)	133	235	83	132	244	89

SD = standard deviation

Comment: What about time on study? This would be very much of interest.

Comment: months or years would have been better

Table 2. Descriptive statistics and 95% confidence intervals for DLCO measurements at baseline, minimum and end of study

Treatment	N	Mean	S.D.	Median	Interquartile Range	Range	95% CI
Baseline							
Placebo	133	19.7	5.02	18.4	5.9	11.5 - 41	18.8, 20.6
Mtx	131	20.3	5.05	20.0	7.1	8.9 - 37.3	19.4, 21.1
Final							
Placebo	125	18.4	4.96	18.2	5.1	8.5 - 39.4	17.5, 19.2
Mtx	120	18.7	4.81	18.8	6.0	8.6 - 35.8	17.8, 19.5
Minimum							
Placebo	125	16.7	4.17	16.2	4.3	8.6 - 32.7	16.0, 17.4
Mtx	120	17.0	4.37	17.2	4.7	7.0 - 34.6	16.1, 17.7

Comment: It was good to include the min and max here.

Table 3. Hypothesis testing for differences in mean DLCO between placebo and methotrexate groups using the Student's t-test.

		DLCO measurement	Mean difference between groups	95% CI	p value
A.	i.	Minimum	-0.25	-1.33, 0.83	0.65
	ii.	Final	-0.33	-1.56, 0.90	0.60
B.	i.	Minimum - Baseline	0.17	-0.73, 1.08	0.71
	ii.	Final - Baseline	0.08	-0.95, 1.11	0.88

A. Differences between the groups for (i) mean minimum DLCO and (ii) DLCO at the end of the study were compared to the null hypothesis value of 0.
B. Differences between the groups for (i) mean change between baseline and minimum DLCO and (ii) mean change between baseline and final DLCO were compared to the null hypothesis value of 0.
CI = confidence interval

Comment: Since you didn't give it elsewhere, it probably would have been a good idea to give a min and maz here, as well. We do look for anecdotal toxicity.

Comment: These figures add next to nothing that is not in the table. I doubt an editor would let you publish them

Figure 1A: Mean Baseline and Minimum DLCO Values with 95% CI

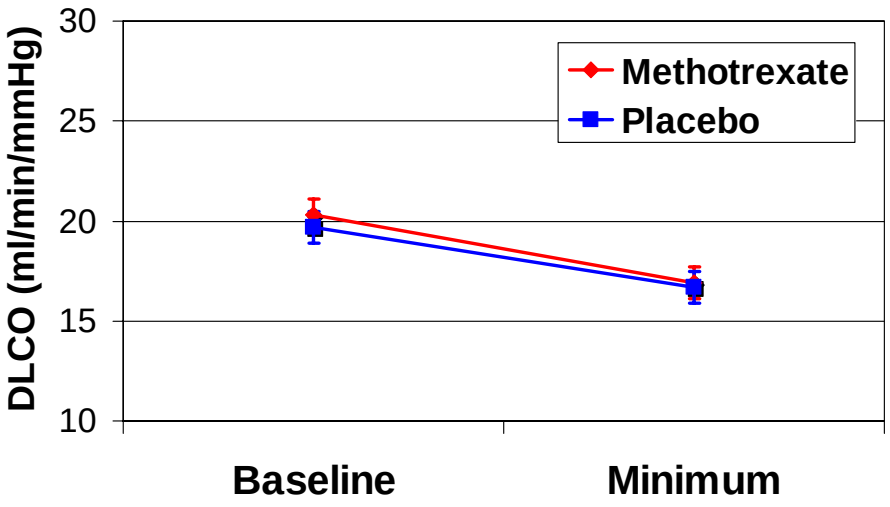
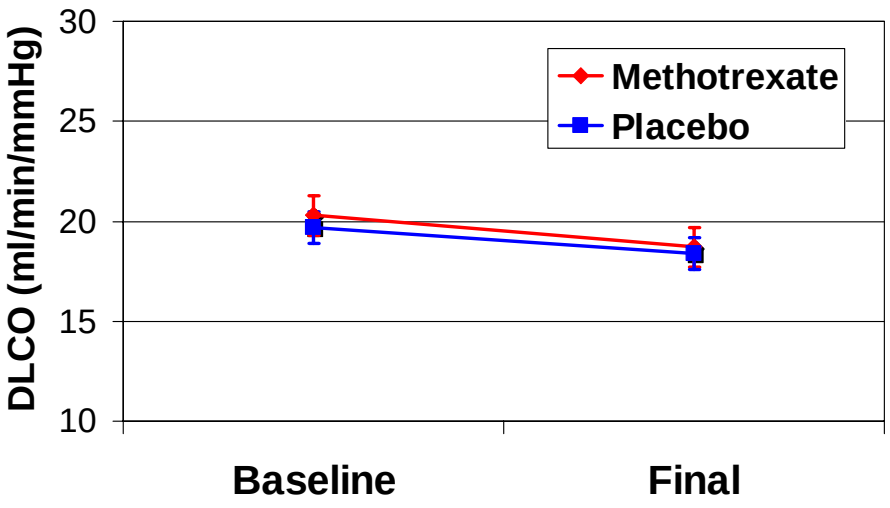


Figure 1B: Mean Baseline and Final DLCO Values with 95% CI



References:

1. Kaplan, M.M., *Primary biliary cirrhosis.* N Engl J Med, 1996. **335**(21): p. 1570-80.
2. Pares, A. and J. Rodes, *Treatment of primary biliary cirrhosis.* Minerva Gastroenterol Dietol, 2000. **46**(3): p. 165-74.
3. Goodman, T.A. and R.P. Polisson, *Methotrexate: adverse reactions and major toxicities.* Rheum Dis Clin North Am, 1994. **20**(2): p. 513-28.
4. Willkens, R.F. and M.A. Watson, *Methotrexate: a perspective of its use in the treatment of rheumatic diseases.* J Lab Clin Med, 1982. **100**(3): p. 314-21.
5. Searles, G. and R.J. McKendry, *Methotrexate pneumonitis in rheumatoid arthritis: potential risk factors. Four case reports and a review of the literature.* J Rheumatol, 1987. **14**(6): p. 1164-71.
6. Kremer, J.M., et al., *Clinical, laboratory, radiographic, and histopathologic features of methotrexate-associated lung injury in patients with rheumatoid arthritis: a multicenter study with literature review.* Arthritis Rheum, 1997. **40**(10): p. 1829-37.
7. Saydain, G., et al., *Clinical significance of elevated diffusing capacity.* Chest, 2004. **125**(2): p. 446-52.
8. Moher D, Schulz KF, Altman D for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA, 2001. **285**(15):p. 1987-91.