Survival and Progression-Free Survival in the PBC and Methotrexate Data Set

Group 5

Presented to Dr. Scott Emerson
Biostatistics 517
Fall 2006

December 13, 2006

Summary

Debate over the use of methotrexate (MTX) as a treatment for primary biliary cirrhosis (PBC) continues, and this study aims to provide evidence to support or refute its efficacy. Specifically, the questions sought to be answered in this analysis are: Does treatment with methotrexate affect survival, and does treatment with methotrexate affect progression-free survival? This study was a randomized, double-blind, placebo controlled clinical trial conducted to evaluate UDCA plus graduated doses of MTX (n = 132) versus UDCA plus placebo (n = 133) in the treatment of PBC. For the purpose of addressing the questions in this report, the primary endpoints were the times to death or progression of disease. The study timeline included an initial data collection visit, monthly visits for the first 6 months, bi-monthly visits for the next 6 months, and a visit every three months after the first year until study withdrawal, study completion, or death, whichever came first. Data on demographics, disease history, hepatocellular damage, liver function, treatment group and duration, and treatment outcome were collected. Comparing outcomes of the MTX group compared to the placebo group, our analysis finds a hazard ratio of 1.23 (95% CI = (0.46, 3.30), log-rank p = 0.6759). The hazard ratio of progression-free survival between the two groups was 1.26 (95% CI = (0.75, 2.12), log-rank p = 0.3802). The study results do not support the use of MTX as a treatment for PBC.

Comment: multicenter

Comment: perhaps a little more detail than usual in an abstract, but OK

Comment: hazard ratio for what? death? Also: a good idea to tell how many deaths and how many progressions in each group, along with some descriptive statistics regarding length of follow-up (these would be obtained from KM statistics for the censoring distn)

Comment: Any comment on the width of the CIs? Seem awfully wide to me.

Background

PBC is a chronic disease of the liver that most often presents clinically as fatigue and pruritus, or an itching sensation that illicits a scratching response. It is a progressive cholestatic disease with an estimated prevalence of 19 to 151 cases per million people (3.9 to 15 cases per million people per year). The majority of PBC cases occur in women, and patients are usually between the ages of 20 and 80 years of age at onset of the disease. PBC begins with cholangitis that progresses to fatal cirrhosis of the liver. At this point, etiology of the disease is unknown. While hyperbilirubinemia is not common among patients presenting with the disease, bilirubin levels in the serum often rise to notably high levels during disease progression.

Comment: elicits

Many treatments for PBC have been proposed. These include liver transplant, drugs to stop disease progression, and symptom management. Many of the drugs used to prevent disease progression are immunosuppresants and anti-inflammatories because evidence suggests that PBC is a disease of the immune system[i]. MTX has been shown to have immunomodulatory or anti-inflammatory effects in patients with certain types of arthritis[ii], lupus, and certain cancers[iii].

An initial pilot study of MTX use for the treatment of PBC did show normalization of serum bilirubin and improved liver histology (as indicated by Ludwig stage) with use of MTX[iv]. Later studies, however, have not shown similar results. A randomized double-blind trial of MTX conducted in 1999 tested the use of low dose (7.5 mg/week) MTX in patients who did not have advanced liver disease[v]. The results showed a tendency to increase serum bilirubin in MTX subjects compared to placebo, and a 2.9-fold increase in rate of death or liver transplantation caused by liver disease. Therefore, the use of MTX as a treatment for PBC is debatable and is still being studied in clinical trials. However, treatment does not seem to have a positive effect in patients whose cirrhosis is advanced or whose livers have already decompensated. The question of MTX efficacy is complicated since onset of beneficial response tends to be slow and improvement may continue for up to 4 years[i].

At this juncture, results do not generally support the use of low dose MTX to treat PBC patients; however, there is still some belief, given that PBC is a disease of the immune system, that MTX's anti-inflammatory and immunomodulatory effects may indeed be beneficial.

Question of Interest

The current study, a randomized double-blind placebo controlled clinical trial, was designed to test the efficacy of MTX in increasing survival time and/or survival time without further disease progression in patients with PBC. Our analysis will answer the refined questions: Does treatment with MTX affect survival when compared to placebo treatment; similarly, does treatment with MTX affect disease progression?

Materials and Methods

Sample

The data for this analysis were collected from a randomized, double-blind, placebo controlled clinical trial conducted to evaluate UDCA plus methotrexate (MTX) versus UDCA plus placebo in the treatment of primary biliary cirrhosis (PBC). For the purpose of addressing the questions in this report, the primary endpoints were the time to progression of disease or death. The study design was reviewed and approved by the Institutional Review Boards at each of the 12 participating clinical centers in the United States.

This study screened 535 patients with PBC for entry into the trial. Of these patients, 385 met inclusion and exclusion criteria; requirements included age within the range of 20 to 69 years and liver disease that was only moderately advanced. For the purposes of this study, moderately advanced disease was defined by multiple factors including at least 6 months duration, elevated alkaline phosphatase levels, and history of positive antimitochondrial antibodies. Patients with advanced PBC, such as those with serum bilirubin 3.0mg/dL or greater, or a serum albumin less than 3.0g/dL, were excluded. Furthermore, all patients with evidence of other forms of liver disease, with a history of alcohol abuse, with other potentially life threatening illnesses, who were pregnant, or who were treated with immunosuppressive agents were excluded from the trial. Of these initially screened individuals, 265 patients entered the trial after a final screening process to ensure adequate renal and pulmonary function, no evidence of biliary obstruction, a positive liver biopsy within the previous 6 months reflective of PBC, and a satisfactory hematologic profile. 12 patients had minor deviations from the eligibility criteria but were judged acceptable by the principal investigator.

Treatment Randomization

The patients who passed screening signed informed consent documents and were enrolled in the trial between January 1994 and March 1998. The patients were randomized in a double-blind fashion into the UDCA plus MTX (132 patients) or UDCA plus placebo (133 patients) treatment arms. Subjects were stratified by the Ludwig staging of liver disease[vi]; the 162 patients with stage 1 or 2 disease were equally divided to receive MTX (62 patients) and the placebo (64 patients); the 139 patients initially diagnosed with stage 3 or 4 were divided to receive MTX (70) and placebo (69). Post-randomization assessment of pathology reports revealed that two patients originally randomized as stage 1-2 should have been randomized with the stage 3 group (these patients remain in the stage 1-2 group for this analysis).

Throughout the course of enrollment, all patients received a UDCA dose of 13-15 mg/kg/day in 300 mg capsules (Ciba-Geigy/Novartis). Likewise, the study drug (MTX or placebo) was administered in 2.5 mg tablets (Lederle Labs/Wyeth-Ayerst Labs) with a graduated dosage. The dosage was increased in 2.5 mg increments each week from 7.5mg to 15mg per 1.73 m^2 body surface area, with a maximum dose of 20mg per week. All patients continued their treatment program until the end of the study, unless prohibited by liver transplantation, excessive drug toxicity, the

development of cancer, or voluntary withdrawal. Patients with adverse reactions to MTX had their dosage schedule adjusted based on the categories of mild, moderate, or severe toxicity. In essence, according to the category, the dosage was stopped entirely or reduced to well below toxic levels and increased 2.5mg per week back to a non-toxic level.

Data Collection

At the time of randomization, multiple variables were measured for each patient to determine the current status of liver function. During the course of the study, patients visited the clinic on a monthly, then bi-monthly (2nd half of first year) schedule, and then at a 3 month interval (after 1st year to end of study). During each visit, a history and blood test were obtained to determine if the had disease progressed as defined by development of variceal bleeding, hepatic encephlopathy, ascites, disabling pruritus, development of new varices (endoscopy once every 2 years), or change of disease stage from liver histology (biopsy once every 2 years). In addition, MTX toxicity data were collected for safety monitoring purposes. Data available for use in our analysis include the following measured at randomization: demographics (age, sex, weight, height), disease history (duration of disease from diagnosis to study randomization), hepatocellular damage (alkaline phosphatase levels in serum, ALT serum levels), and liver function (blood bilirubin level, clotting time, albumin level, serum cholesterol level). Also available were variables measuring treatment group assignment (treatment group, number of days on treatment) and treatment outcome (time until first of disease progression or last follow-up, time until first of death or last follow-up) were recorded.

Comment: actually ought to use the term "prothrombin time", as there are other measures that are "clotting time"

Statistical Methods

All methods used for analyzing this study are standard statistical analysis techniques. They were chosen since they are appropriate for examining censored data (such as the survival status and progression-free survival status that are of interest in this study). Comparisons between treatment groups were made under an intent-to-treat assumption, using treatment assignment, rather than time on study drug, as a basis for comparison. This is justified since the times to when patients stopped taking drug were not found to be significantly different between treatment groups. All analyses were conducted using standard statistical software (STATA 9, College Station, Texas).

Comment: skip these sentences and lead off with the statistical techniques used: KM estimatates, PH regression with logrank statistic. The mention of intent-to-treat can come first as you have it, or after the techniques. But it is very good to mention that.

Then, the more descriptive nature of the statistical methods can come for the purposes of communicating with collaborators, though they probably would not make it into the manuscript itself.

Our statistical tests provide point estimates, 95% confidence intervals, and p-values. Confidence intervals put a result from our limited sample in the context the true population. For instance, looking at differences in 4 year survival probability between drug- and placebo-randomized groups, the confidence interval says that the difference calculated from our sample would be typical if the true difference in the population of patients as a whole lay within the confidence interval. In other words, we would say that we are 95% confident that, given our study results, the true difference between the groups lies within the range defined by our confidence interval. P-values, on the other hand, provide a decision criterion for hypothesis testing. The p-value tells us the probability that any difference we observed was due to random chance. Thus the lower the p-value, the more certain we are that we should reject the null hypothesis. In the example above, our null hypothesis would be that there is no difference between survival probabilities for patients on drug versus patients on placebo.

Comment: when there was no true treatment effect

We use the Kaplan-Meier estimator as a descriptive statistic for survival (or progression-free survival) at two key time points for each study group; for this descriptive analysis, we chose years 4 and 8 as time points representative of mid-term and late outcomes for this 9 year study. The Kaplan-Meier statistic estimates a probability of survival at each given time point during the study by taking into account the surviving patients lost to follow-up as well as the patients who died at each time point.

We use the log-rank test for our primary analysis of treatment effect. The log-rank test looks for overall effect of a treatment, rather than just at a given time point. It was chosen because not only

is it appropriate for censored data, but it provides greater statistical power for detecting an overall effect than does the Kaplan-Meier (which looks at particular years), providing more confidence should we obtain a positive result. Log-rank assumes that events (death or death/deterioration of condition) are equally expected between the two treatment groups, and then compares how many events were actually observed for each group. From this we provide a hazard ratio, which is simply the relative risk of death (or death/deterioration) between the treatment groups, and test significance using a p-value. From a clinical viewpoint that higher bilirubin or later stage disease are possibly predictive of worse outcome, a secondary analysis of this same type was suggested; for this, we compared treatment groups within subgroups defined by high bilirubin (>=1.2 mg/dL), low bilirubin (<1.2 mg/dL), early stage (I or II), and late stage (III or IV) liver disease.

The literature reports that there may be a drug effect up to 4 years[1], so we considered using the Wilcoxon rank-sum test, which is more sensitive to early effects than log-rank. However, previous, shorter, studies failed to find a significant treatment effect, so it was decided that log-rank would be a more appropriate analysis for our longer-term (9 year) study.

Missing data did not present a problem for this analysis, since we are looking at major outcomes of the study, and most baseline data were required to be present for inclusion into the study.

Results

Table 1 shows baseline descriptive statistics by treatment group and for all subjects together. All, except weight and height, are taken directly from data supplied by the principle investigator. Weight and height used to calculate body mass index (BMI); this combined factor was used to determine study drug dosage. No variables used in our analysis required data transformations as distributions tended to be close to normal. In addition, there were no outliers that caused sufficient concern to exclude them. The two treatment groups are very similar across these baseline characteristics.

Comment: Was your interest looking at effect modification or was it only looking for added precision. It is not clear from this what your motivation was.

Comment: This would probably go better in the Discussion. Here you should talk about what you did do. Later you could talk about how it fit in with other studies.

It was good to consider whether the Wilcoxon form of logrank might have been better (this is not exactly the same as the Wilcoxon rank-sum, though it is looking at the same summary measure). Of course, you have to consider whether a treatment that conferred benefit only over 4 years was clinically important in this chronic disease. That, to me, would be more important a criterion than whether I was merely duplicating prior findings. (But placing your results in the context of prior findings is also important.)

Comment: You could have remarked here that the validity of your analyses assumes that any subjects lost to follow-up is not informative censoring.

Comment: principal

Comment: This comment could have been omitted, as BMI is pretty standard. If you were using something a little unusual, you would have described the transformations under Methods.

Comment: Normality is immaterial here! A uniform distribution would also never suggest any transformations. Heavily skewed data might lead us to geometric means instead of means, but that is how I would talk about it: Geometric means of the original data, rather than means of transformed data.

Comment: I NEVER, EVER, EVER would exclude “outliers” in the report of a clinical trial. I would never consider it, So I certainly would not comment on the fact in a manuscript. I would presume that all my readers were sufficiently knowledgeable to know this would not be done in a quality study.

Table 1: Baseline Characteristics, by treatment group

	treatment = placebo (133 subjects)				treatment = MTX (132 subjects)			
variable	**N**	**mean (SD)**	**median**	**range**	**N**	**mean (SD)**	**median**	**range**
age	133	52.19 (8.64)	53	25, 67	132	50.38 (8.67)	50	31, 70
sex	133	0.0752 *(10/133 subjects are male)*			132	0.0758 *(10/132 subjects are male)*		
BMI*	132	27.3 (5.4)	26.8	17.1, 53.4	132	26.1 (4.7)	25.3	17.6, 43.2
disease stage	133	2.5 (1.0)	3	1, 4	132	2.6 (0.8)	3	1, 4
durdis **	133	3.7 (3.3)	2.4	0.5, 14.4	132	3.5 (3.5)	2.7	0.6, 17.9
splenomeg	133	0.1053 *(14/133 subjects had splenomegaly at study start)*			131	0.0840 *(11 /131 subjects had splenomegaly at study start)*		
bilirubin (mg/dl)	133	0.7 (0.4)	0.7	0.1, 2.3	132	0.7 (0.4)	0.6	0.1, 2.8
albumin (g/dl)	133	4.0 (0.3)	4.0	3.1, 4.9	132	4.0 (0.4)	4.0	3, 5.9
alk. phosph. (U/l)	133	244.9 (187.4)	200	63, 1127	132	242.8 (145.7)	208.5	50, 930
ALT (U/l)	132	50.2 (41.6)	38	9, 314	131	54.3 (41.4)	42	9, 199
pt time ***	133	11.4 (1.1)	11.5	8.6, 13.9	128	11.2 (1.1)	11.35	8.5, 14.6
Cholesterol (mg/dl)	130	235.8 (58.8)	224.5	128, 560	130	239.2 (58.2)	233	137, 472
Platelets (1000 cells/cu mm)	133	234.7 (83.2)	235	77, 561	132	243.5 (88.6)	231.5	86, 619

* BMI = weight(kg) / height(m^2)
**disease duration (years): time between diagnosis with PBC and randomization into study
*** prothrombin time (seconds)

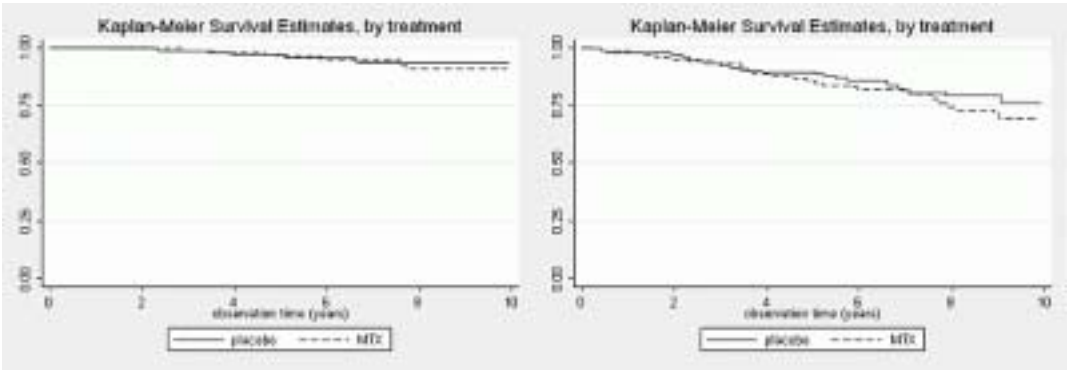
Comment: This formatting of the table makes it difficult to read. There are better ways it could be presented.

Comment: Use English rather than cryptic abbreviations. The footnote could be used for exact definitions of the measurements, but a better label should be used for each row.

During the course of this trial, 58 of the 265 enrolled patients were found to have a progression in their liver disease, and 16 died. Figure 1 shows the Kaplan-Meier estimated survival curves for survival and progression-free survival by treatment group. Note that any observable separation tends to occur at the extreme ends of the graphs, and may be considered to be artifacts of the Kaplan-Meier estimator algorithm.

Comment: First tell us what the length of follow-up was, as might be obtained from a KM estimate of the censoring distn. I might then comment on the drug compliance while on study, though you did not have full data to be able to analyze this. Then, it is good to include this info, though I would have done it by treatment group.

Comment: I would instead remark that this is where the estimates are less precise, thus the KM estimates are less reliable.



survival probability estimates progression-free survival probability estimates

Figure 1: Outcomes, by treatment group

Table 2 shows Kaplan-Meier estimates for survival probability and progression-free survival probability by treatment group. As with the accompanying figure above, the two groups appear very similar with regard to estimated survival probability and with regard to estimated progression-free survival probability; this is true for mid-study (year 4) and late-study (year 8) comparisons.

Table 2: Outcomes, by treatment group (Kaplan Meier Survival Estimates)

Kaplan-Meier Estimate for Survival Probability		
year	**survivor function (95% confidence interval)**	
	treatment = placebo	*treatment = MTX*
4	0.9663 (0.9126 , 0.9872)	0.9753 (0.9254 , 0.9920)
8	0.9345 (0.8664 , 0.9685)	0.9083 (0.8281 , 0.9521)
Kaplan-Meier Estimate for Progression-Free Survival Probability		
year	**survivor function (95% confidence interval)**	
	treatment = placebo	*treatment = MTX*
4	0.8897 (0.8209 , 0.9332)	0.8835 (0.8142 , 0.9281)
8	0.7880 (0.6975 , 0.8542)	0.7435 (0.6515 , 0.8146)

Table 3 shows hazard ratios for survival and progression-free survival. The survival hazard ratio says that those in the MTX group were 1.23-times as likely to experience death as those in the placebo group. Similarly, those in the MTX group were estimated to be 1.26-times as likely to experience death or deterioration of their condition as those in the placebo group. However, neither of these results are significant (p = 0.6759, p = 0.3802). Note that for both survival and progression-free survival, the 95% confidence intervals are rather wide, and given these intervals, it would be plausible that the true hazards for the population were < 1, indicating a reversal in which treatment may be considered superior. Again, the results are non-significant.

Comment: Very good to note briefly here, as you did. (Hopefully, you will discuss this further in the Discussion)

Comment: Omit this redundant sentence.

Table 3: Results, by treatment group

	Hazard Ratio (95% confidence interval)	**log-rank p-value (2-sided)**
Survival	1.23 (.46 , 3.30)	0.6759
Progression-Free Survival	1.26 (0.75 , 2.12)	0.3802

Hazard ratios and Kaplan-Meier estimates for survival probability and progression-free survival were also calculated for subgroups defined by early (1 or 2) and advanced (3 or 4) stage disease at randomization, and low (≤1.2 mg/dL) and high (>1.2 mg/dL) bilirubin levels. Results were insignificant within these groups, as they were for the whole sample.

Comment: The issues could be that by doing a stratified analysis, you would have more precision, or that by looking at tx effects within subgroups would show some evidence of effect modification. In the latter case (which you seem to be more interested in), you would definitely lack precision due to the lower sample size. So the mere lack of significance is not the important issue. Instead you should give estimates.

Discussion

Many studies on the efficacy of MTX in treating PBC have been performed with mixed outcomes, with more tending to refute efficacy than support it. This is the largest randomized, double-blind placebo controlled trial of MTX that has been completed at this point in time. In addition, this trial followed patients for up to 9 years, well beyond the 4 years indicated in the literature for maximum benefit[i] and one of the longest follow up times among studies of this drug.

Since PBC patients have various symptoms on liver inflammation and hepatocellular damage, many clinical measurements in this dataset were available to monitor liver function, liver inflammation and hepatocellular damage. As presented above, statistical tests point to the conclusion that MTX does not improve survival or progression free survival of PBC patients. The design of our analysis was not ideal, since our sub-group analysis opened up the problem of multiple comparisons, should we have found a significant effect.

The accuracy of our results may also be compromised by the lack of information on dosage changes for each patient due to toxicity. For example, if a large fraction of the MTX group experienced toxicity and were administered minimal dosage for most of the study, then our results do not reflect the effects of MTX (at the dosage indicated by study protocol) versus placebo. Furthermore, we do not know the root cause of disease progression for each patient; our results may be limited if the primary marker of liver disease progression was only evaluated once every 2 years providing an overly discrete time sampling (i.e. if most patients were diagnosed with disease progression only after biopsy). Likewise, if the cause of disease progression differed between the treatment arms and since the causes were sampled at different time points (monthly/yearly versus every other year), the reported statistical results may lack accuracy and precision.

In this data set, the variables supplied were limited. Most notably, toxicity data were not included. Therefore, the number of variables we were able to adjust for were limited. Additionally, since the number of male subjects was quite small compared to the number of females (due to differences in prevalence between sexes), treatment versus placebo could not accurately be compared in the sexes separately. While there are shortcomings to our data and analysis, we believe that the overall conclusions still firmly stand.

Conclusion

This trial fails to support an MTX-treatment related difference in survival or progression-free survival for patients with primary biliary cirrhosis. MTX efficacy as a treatment for PBC has been studied numerous times. We believe that this study confirms that further studies on this use of the drug are unnecessary. Even if a treatment effect is found, it is likely to be within a subgroup that is so small that the cost of future studies to uncover this effect would be too high to justify them. Resources would best be spent investigating other treatments for PBC.

Comment: Good to note. It would be much stronger if you told us the KM estimates of median follow-up.

It would probably be a good idea to comment on whether the patients were taking their meds for all this time, as well. You can't change whether they did or not, but it might also indicate problems with toxicity that made them stop taking the drug.

Comment: I would not apologize for exploratory analyses. You put proper emphasis on the "bottom line" analysis. It is true that your failure to give estimates for the subgroups do not allow me to judge whether I should have any further interest in the treatment.

Comment: Very good to note, and this mitigates somewhat my comment above. You do have data on time they were taking drug. Ostensibly they were taking the highest dose they could tolerate, if they had a low dose prescribed. So, compliance is hard to measure: We need to consider 1) the dose prescribed, 2) how often they "forgot" to take the prescribed dose, and 3) whether they quit taking the drug entirely. And then we need to wonder why they stopped taking the drug—perhaps it was toxicity that was not explicitly reported.

Comment: Excellent point! If toxicity makes a patient go to the doctor more often, that might introduce bias into the primary endpoint of progression (though hopefully not death).

Comment: How would you have adjusted for the post-randomization variable of progression. Hopefully, you would not. And of course, there is no huge need to adjust for baseline variables, though you can gain a bit precision if you do. I do note that you have the most important variables from this population (we did not have very many with really severe disease at baseline, so other strongly predictive variables would have had little variation in our sample).

Comment: Any comment on the extremely wide CI? Would a HR of 0.5 be clinically important? If so, is this study of any use?

Comment: I agree.

References

[i] Nishio, A, Keeffe, EB, Ishibashi, H, and EM Gershwin. 2000. Diagnosis and treatment of primary biliary cirrhosis. Medical Science Monitor 6(1): 181-193.

[ii] Ramanan AV, Whitworth P, and EM Baildam. 2003. Use of methotrexate in juvenile idiopathic arthritis. Archives of disease in childhood. 88: 197-200.

[iii] Mulne AF, Ducore JM, Elterman RD, Friedman HS, Krischer JP, Kun LE, Shuster JJ, Kadota RP. 2000. Oral methotrexate for recurrent brain tumors in children: a Pediatric Oncology Group study. Journal of Pediatric Hematology and Oncology. 22(1): 41-44.

[iv] Kaplan, MM and TA Knox. 1991. Treatment of primary biliary cirrhosis with low-dose weekly methotrexate. Gastroenterology. 101(5): 1440-142.

[v] Hendrickse MT, Rigney E, Giaffer MH, Soomro I, Triger DR, Underwood JCE, and D Gleeson. 1999. Low-dose methotrexate is ineffective in primary biliary cirrhosis: long-term results of a placebo-controlled trial. Gastroenterology. 117: 400-407.

[vi] Ludwig J, Dickson ER, and GSA McDonald. 1978. Staging of chronic nonsuppurative destructive cholangitis (syndrome of primary biliary cirrhosis). Virchows Archiv A-Pathological Anatomy and Histopathology. 379:103-12.