Overview of the hepatitis data[1]

Currently there are 5 identifiable forms of viral hepatitis namely A, B, C, D and E. All of these viruses are hepatotrophic (i.e. liver is the primary site of infection). Approximately 4 million people in U.S. and 100 million people worldwide are infected with the HepatitisC (HepC) virus. Approximately 85% of persons with acute HepC develop chronic hepatitis as determined by persistently abnormal serum enzymes and/or viremia (HepC virus (RNA)). Both the acute and the chronic illnesses are predominantly asymptomatic. For this reason and because of chronic illness runs in extremely protracted course, it has been difficult to accurately define the frequency and the rate of progression to symptomatic or end stage liver disease and death.

The response for the current treatment for the Hepatitis C virus is only about 30-40%. And has many side effects and is expensive. Liver biopsy is usually the most specific test to assess the nature and severity of the liver disease. So it is proposed in the medical literature to treat stage III and IV liver disease on liver biopsy. Liver has a rich vascular supply, therefore there are some complications associated with the liver biopsies. Approximately 1-3% of patients require hospitalizations for complications after liver biopsy. Complications include transient, localized discomfort at the biopsy site; pain requiring analgesia; and mild transient hypotension. Approximately ¼ th of the patients have pain in the right upper quadrant or right shoulder after liver biopsy. Although very rare clinically significant intraperitoneal hemorrhage, is the most serious bleeding complication of liver biopsy.

Therefore in order to avoid complications associated with liver biopsy and to predict the severity of the disease early so that the treatment (medical or surgical (i.e. Liver transplant) can be started early. In our study we aim to predict the stage of the disease (I, II, III, IV) using data mining software WEKA thus avoiding Liver Biopsy.

In medical literature different modes of the disease transmission has been documented. The most common modes of transmission are IVDA (Intravenous Drug Abuse), usage of nasal Cocaine, Blood Transfusion (Tx), Needle Stick (N) in occupation, for example accidental needle sticks in work place (Nurses, doctors, emergency medical technicians and other health care professionals), presence of Tattoo marks, Sexual transmission of the disease. It has been documented that some people do not have any of the above risk factors and fall into the category of No Risk Factors (NRF). Co-infection with the Hepatitis B virus (HBV) or with Human Immune Deficiency Virus (HIV) is also an important consideration.
Alcohol use (ETOH), Obesity, co-infection with HBV and HIV makes Liver Disease progression faster and worse. The current treatment options for the patients inflicted with the HepC virus is Interferon and Ribavarin. There are 6 Genotypes (GT) of the Hepatitis virus: 1, 2, 3, 4, 5, and 6. About 70% of the patients in US have subtype 1. Only 50% of the genotype1 respond the above treatment. In genotypes other then 1 have about 70% response rate.

Liver Function test (LFT) is used as an indicator of the severity of the liver disease, it is represented as negative if the test result is within the normal lab limits else if it is greater than 1.5 times the normal it is recorded as positive. Duration, which is the number of years for which the patient had the disease is also important in determining the future progression of the disease.

---

[1] The material presented here is copied from MINING MEDICAL DATA: Predicting the stage of Hepatitis-C Using the WEKA 3.2 Data Mining System, capstone project of PADMA TATAVARTHY, SHWETHA TIPPA, KARUNASRI SEELA, CIT, CCSU, Fall 2002.

Liver Biopsy is performed to stage the severity of the disease (I, II, II, IV) and also to determine if the treatment is indicated. For example a 50 year old male acquired HepCV at the age of 20 by IVDA has normal LFT and liver biopsy (Bx) shows stage I disease no treatment is indicated.

It is generally that if the Biopsy is in Stage I and II no treatment is indicated, else if it is in stage III or IV the patient is given a regimen of Interferon and Ribavarin.

The table shown below gives the attributes, their values and descriptions.

No	Attributes	Type of Data	Values	Descriptions
1	Sex	Categorical	M, F	Gender
2	DOB	Numeric	Date	Date of Birth
3	DOT	Numeric	Date	Date of transmission of the disease
4	Route	Categorical	Coc, IV, Tx, N, NRF, Tatt, Sex	The route through which the disease was transmitted.
5	IV	Categorical	+, -	Intravenous
6	Tx	Categorical	+,-	Blood Transfusion
7	Coc	Categorical	+, -	Usage of Cocaine
8	Tatt	Categorical	+, -	Presence of Tatoo on the body of the patient
9	HBV	Categorical	+, -	Presence of Hepatitis B virus in the patient.
10	HIV	Categorical	+, -	Presence of HIV infection
11	EtOH	Categorical	+, -	Alcohol usage by the patient.
12	Obes	Categorical	+, -	Whether the patient is obese or not.
13	Rx	Categorical	+,-	Treatment, it is whether the patient has been treated.
14	Tox	Categorical	+, -	Presence of any toxic elements.
15	CLD	Categorical	+, -	Whether the patient has Chronic Liver Disease.
16	GT	Categorical	-, I, II, III	Genotype of the patient.
17	LFT	Categorical	+,-	Whether or not the Liver Function Test was done.
18	Duration	Numeric		Number years for which the patient had the disease. It is basically the difference between DOB and DOT.
19	Age	Numeric		Current age of the patient.
20	Bx	Categorical	I, II, III, IV	Biopsy result, which specifies the stage of the HepC.