Project Proposal

Jejo Kola

Mentor: Dr. Michael Matheny

Using Natural Language Processing to Identify Cirrhosis Associated Features

1. Cirrhosis
   1. Prevalence
   2. As of 2008, the prevalence of chronic liver disease in the US reached 15% overall, of which 0.3% was Hepatitis B viral infection, 1.7% was Hepatitis C (HCV) viral infection, 2.1% was alcoholic liver disease, and 11% was nonalcoholic fatty liver disease (NAFLD).[27](#_ENREF_27) It was estimated that more than 44,000 deaths each year can be attributed to the complications of cirrhosis, nearly as many as attributable to diabetes and more than attributable to kidney diseases.[21](#_ENREF_21)
   3. Main risk factors (substance use, viral hepatitis, NASH)
   4. Impact on VA
      1. Substance use
         1. For FY2011, 484,785 patients were diagnosed with SUD, of whom 379,342 were isolated alcohol use disorder.[47](#_ENREF_47)
      2. Viral hepatitis
         1. The VA is the largest single provider of HCV care in the US, and has approximately 186,000 patients with chronic active HCV.
      3. NASH
         1. Overall, 90-100% of alcoholics develop liver steatosis, 10-35% develop alcoholic fibrosis and/or hepatitis, and 10% develop cirrhosis.[48](#_ENREF_48)
2. Difficult to treat
3. Liver disease is under-diagnosed
   1. One study evaluated the barriers to receiving recommended care by manual chart review from a random sample of patients with cirrhosis diagnosed with hepatocellular carcinoma (HCC) and discovered that 18% of patients had unrecognized liver disease and an additional 21% had known liver disease but unrecognized cirrhosis in the year preceding HCC diagnosis.[25](#_ENREF_25)
4. Barriers to managing cirrhosis
   1. Furthermore, among VA patients hospitalized with cirrhosis, a large proportion do not receive recommended medical treatments.26 This could be contributing to the 37% 30-day readmission rates noted in a study of hospitalized decompensated cirrhotic patients, of which 22% were deemed preventable.23
5. Risk models don’t include narrative data
   1. MELD, Madrey’s, NASH, CTP
6. NLP
   1. First, examining the utility of employing NLP in near real-time for clinical care, whether at the bedside, clinic, or in between episodes of care, is a novel area of research because it has not yet been shown to be useful outside of retrospective environments. Interactive Machine Learning (IML) has not been used for document classification, and may provide accurate and computationally inexpensive information extraction once trained. Additional innovation will be pursued by developing the IML methods for this use. As a domain, development of NLP methods for the extraction of social history has been lacking due to its complexity of representation. This is a particularly useful target for a number of clinical domains, as social history greatly impacts mental health and other chronic disease management.
7. Methods
   1. Subjects/notes
      1. Notes Initial Gastroenterology Consultation to produce enriched sample
      2. 500 total notes
      3. 1 note per patient
      4. Patients seen by inpatient or outpatient GI service from during 2011.
         1. In 2011, there were a total of 104,907 cirrhotic VA patients as determined from outpatient and inpatient diagnostic codes, with a total of 1352 at the TVHS VA and 1072 at the San Diego VA. Using only inpatient diagnostic codes, there was a total of 29,932 national, 502 TVHS VA, and 330 San Diego VA cirrhotic patient hospitalizations.
      5. Unstructured narrative text data from clinical progress notes, radiology reports, and pathology reports, will also be collected from the CDW. The data is currently updated quarterly, and efforts are underway to narrow this window.
   2. NLP
      1. YTEX is a full UIMA-based pipeline based on cTAKES. The application includes modules for detecting and annotating different sections in a document (“document sectionizer”) negated phrases, the UMLS mappings of phrases. YTEX also includes document level classification using rule-based and statistical machine learning methods.[58](#_ENREF_58),[59](#_ENREF_59)
      2. Statistical text mining: using RapidMiner version Community Edition 5.3.015x32 (www.rapidminer.com)
         1. Term by document matrix
         2. Dimension reduction
            1. Latent Semantic Analysis: SVD (single value decomposition?)
         3. Use output as factors for our machine learning
   3. Machine learning
      1. We used data among 8091 documents and 236 patients from the Veterans Aging Cohort Study, which conducted a manual review upon patient enrollment of radiographic evidence of hepatic decompensation, specifically ascites, varices, and liver masses.[58](#_ENREF_58) This data, *which included document level but not concept level annotation*, was used to train the YTEX document classifier for these conditions, and found that a decision tree algorithm overall performed the best with precision, recall and F-measures of 0.93, 0.94, and 0.93 for ascites, 1.0, 0.89, and 0.94 for varices, and 0.98, 0.78, and 0.80 for liver masses, respectively.
      2. Methods
         1. Decision tree
         2. Naïve bayes
         3. SVMs
         4. Logistic regression
      3. Validation
         1. 10-fold cross validation
   4. Statistics
      1. ML stats
         1. Precision, recall, F-measure
         2. Using R version 3.0.1