In-depth annotation for patient level liver cancer staging

Wen-wai Yim
Biomedical Informatics and Medical Education
University of Washington
wyim@uw.edu

Sharon Kwan
Radiology
University of Washington
shakwan@uw.edu

Meliha Yetisgen
Biomedical Informatics and Medical Education
University of Washington
melihay@uw.edu

Abstract

Cancer stages, which summarizes extent of cancer progression, is an important tool for evidence-based medical research. However, they are not always recorded in the electronic medical record. In this paper, we describe work for annotating a medical text corpus with the goal of predicting patient level liver cancer staging in hepatocellular carcinoma (HCC) patients.

Our annotation consisted of identifying 11 parameters, used to calculate liver cancer staging, at the text span level as well as at the patient level. Also at the patient level, we annotated stages for three commonly-used liver cancer staging schemes. Our inter-rater agreement showed text annotation consistency 0.73 F1 for partial text match and 0.91 F1 at the patient level.

After annotation, we performed several document classification experiments for the text span annotations using standard machine learning classifiers, including decision trees, maximum entropy, naive Bayes and support vector machines. Thereby, we identified baseline performances for our task at 0.63 F1 as well as strategies for future improvement.

1 Introduction

Despite their importance in research, cancer stages are not always recorded in the electronic medical record (EMR) in structured or unstructured format (Evans et al., 1998). Even when collected they are often inaccurate (Yau et al., 2002)(Sexton et al., 2006). On the other hand, review of patient notes for medical conditions is both time-consuming and expensive. One strategy to minimize these costs is to leverage natural language processing (NLP) to automate the process.

In this paper, we describe work for annotating a corpus with the goal of predicting patient level liver cancer staging in hepatocellular carcinoma (HCC) patients. Our group took a detailed annotation approach, which included text span level and patient level annotation of parameters used in staging, as well as patient level annotation of stages for three liver cancer staging schemes. In our results we present our inter-rater agreements and our analysis from studying our domain experts' annotations. Finally, in our last section we deliver preliminary information extraction baselines using several standard machine learning classifiers.

Clinically relevancy for this task is especially well exemplified by HCC for which has many competing treatment options but no universally accepted clinical guidelines (Han et al., 2011). Moreover, HCC progresses differently across various age groups, ethnicities, lifestyles, and associated co-mordidities (McGlynn and London, 2011). Automatic staging may facilitate evidence-based research for targeted disease management by leveraging the EMR for best outcomes. Its scaleable nature would allow the process to be adapted for volumes of historical data, efficiently unlocking more information than comparable prospective trial studies.

2 Background

Cancer staging is used to summarize the extent of disease for cancer patients. Each cancer domain may have different criteria for its stages. For example, ovarian cancer stages differentiates between whether one ovary is invaded, both, or the entire pelvic region (American Cancer Society, 2014).

For liver cancers, in addition to tumor morphology and spread, patient performance status as well as liver function variables are incorporated into

Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi), pages 1–11,
Lisbon, Portugal, 17 September 2015.

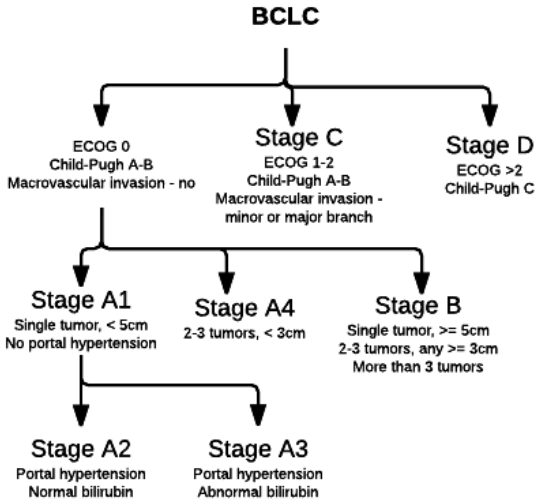

Figure 1: BCLC staging logic

various staging schemes. However, because there are various measures of tumor growth, liver failure, and overall patient well-being, over six different international liver cancer staging schemes exist (Sirivatanauksorn and Tovikkai, 2011). For our project we focus on capturing the parameters and classifications for three commonly used staging schemes: the American Joint Committee on Cancer (AJCC), the Barcelona Clinic Liver Cancer (BCLC), and the Cancer of the Liver Italian Program (CLIP) staging schemes (França et al., 2004). Figure 1 shows an example of the stage parameters, e.g. ***ECOG***, and the decision logic for classifying BCLC stages, e.g. ***Stage A1***.

In all, there are a total of 11 text parameters and 4 structured data laboratory parameters among the 3 staging schemes. Because Child-Pugh, one of the text parameters, is itself a classification system for severity of liver disease, when necessary, it must be calculated according to Table 1 logic.

The purpose of annotating stage parameters, in addition to overall stages, is two-fold. Firstly, more detailed annotation can presumably help with performance. Secondly, stage parameters may be used in more than one staging scheme, or may be re-used if a staging classification algorithm changes, given some little additional annotation.

Variable	Points		
	1	2	3
Albumin (g/dL)	> 3.5	2.8-3.5	< 2.8
Ascites	None	Mild/Moderate	Severe
Bilirubin (mg/dL)	< 2	2-3	> 3
Hepatic Encephalopathy	None	Grade 1-2	Grade 3-4
Prothrombin INR	< 1.7	1.7-2.3	> 2.3

Table 1: Child-Pugh parameters. Adding up the points for all variables, stage is assigned where Child-Pugh A: 5-6 points, Child-Pugh B: 7-9 points, and Child-Pugh C: 10-15 points.

3 Related work

We describe previous work by grouping systems by those that predict a cancer stage and those that extract cancer characteristics which make up stage parameters.

3.1 Cancer stage prediction

Previous work in automatic cancer staging from clinical documents focused on TNM cancer stage classification using document classification. A brief explanation: T, N, and M represent tumor size, lymph node spread, and metastasis, respectively. Each parameter takes different values depending on spread. For example T0 means no tumor, while T1-T4 are increasingly larger sizes. An example TNM stage for a patient is ***T2 N1 M0***.

Nguyen et al. (2007) predicted patient TNM stage by using multi-class document classification of concatenated records with support vector machines (SVM). They tested various hierarchical set-ups, i.e. binary for each variable vs. all versus all, etc, achieving accuracies of 64% and 82% for T and N sub-stages. The same group, McCowan et al. (2007) divided the document classification problem into a number of sentence-level classifications, in which a sentence is first classified for a particular parameter, e.g. T2, N1, etc. After predicting a value for each sentence, using SVM or some rules, the final stages were determined by post-processing heuristics. This strategy improved T and N accuracy to 74% and 87%. In their latest work, Nguyen et al. (2010) used a symbolic logic approach. Rules leveraged concept-normalization, negation, and normalization through the SNOMED-CT hierarchy. Their accuracy using these methods improved to 72%, 78%, and 94% for T, N, and M, respectively. Martinez and Li (2011) classified report level TN and ACPS stages, testing a mixture of document classification, sentence-level extraction, and rule-based methods and arrived at best F1 scores of 82%, 81%, and 75% for T, N, and ACPS staging, respectively.

Viewed from a larger scope, patient cancer stage prediction may be framed as a special case of clinical phenotype identification, which similarly involves distilling a patient's multiple clinical data sources, free-text and structured, to identify a spe-

cific disease or set of conditions. We will not discuss this with regards to our work here, but shall point the reader to an excellent review on phenotype cohorts using EMRs (Shivade et al., 2014).

3.2 Cancer information extraction

Two previous works focused specifically on liver cancer information extraction. One was a 2013 rule-based system (Ping et al., 2013) that extracted elements of liver cancer diagnosis, tumor characteristics, staging (BCLC and Child-Pugh), co-morbidities, and treatments using regular expression and rules. They captured concepts and relations in a diverse set of report types, with performances ranging 92-99% F1.

The other study was a 2014 hybrid system, Wang et al. (2014), in which HCC information was extracted from operation notes. First, relevant sentences of interest for a parameter were identified with keyword look-ups, then information was structured using a conditional random field algorithm. They achieved a 64% F1 performance.

A plethora of general cancer information extraction systems exist, concentrating on parameters such as tumor size, number and metastasis. Many use dictionary-based methods (Coden, 2009)(Ashish et al., 2014) for extracting entities before structuring them using specific algorithms. However statistical named entity recognition methods (Ou and Patrick, 2014) and document classification methods are also used (Jouhet et al., 2012)(Kavuluru et al., 2013).

Our annotation approach combines previous methods. Similar to McCowan et al. (2007) we annotated for stage parameters at a sub-document level before making an overall staging classification. However, we additionally annotated liver cancer specific information and marked at a text span level, as in Ping et al. (2013) and Wang et al. (2014). Unlike previous information extraction approaches, we annotated stage parameters at a patient level in addition to text span levels. Unlike previous cancer stage prediction systems, we classify over various report types instead of only histology and pathology reports.

4 Corpus Creation

4.1 Data and Processing

A cohort was drawn from new patients visiting the University of Washington (UW) Medical Center primary liver cancer clinic from 1/2011-12/2013 with approval by the UW Human Subjects Committee of Institutional Review Board. Included data for each patient comprised of: (1) all clinical notes from the day of visit to the clinic, including surgery, admit notes etc, (2) all laboratory results within 30 days prior to and following the visit day, and (3) radiology reports within 3 months prior to and 1 month following the visit day.

Patient records were manually reviewed by our clinical expert to exclude patients who had been seen prior to the start of the study and patients who had an obviously irrelevant diagnoses. Previously seen patients were excluded because they most likely had already started treatment, and our population of interest were patients at first presentation. Irrelevant report types were removed from the annotation set. For our study, we focused on the subset of patients that had at least one clinical report, at least one radiology report, and the full set of labs needed for staging. The resulting dataset included 236 patients and their associated 422 clinical and 309 radiology reports.

4.2 Guideline Creation

Guidelines for liver cancer stage and stage parameter annotations were developed primarily by an interventional radiologist with input from another interventional radiologist and a group of NLP scientists. Stage parameter values were discretized according to stage guidelines. The stage and stage parameters are described following.

Stage:

AJCC has classifications (I, II, IIIA, IIIB, IIIC, IVA, IVB) and is based on the TNM stage framework that primarily addresses tumor characteristics and spread but not liver functioning statuses.

BCLC has classifications (A1, A2, A3, A4, B, C, D) and is the only staging scheme that takes into account overall performance status (ECOG).

CLIP has classifications (0, 1, 2, 3, 4, 5, 6) and is the only staging scheme that takes into account the relative size of the tumor to the liver.

Stage parameters:

Ascites : accumulation of fluid in the peritoneal cavity (e.g., "no significant ascites," "does not endorse abdominal swelling") with values (None, Mild, Moderate-Severe)

Child-Pugh : a measurement of liver cirrhosis (e.g., "Child's B," "his CTP score would be 5") with values (A, B, C)

ECOG (Eastern Cooperative Oncology

Group) Performance Status : a scaled measure of general well-being where 0 is fully active and 5 is dead (e.g., "ECOG 0," "She notes good energy") with values (0, 1, 2, $\geq$3)

Extrahepatic invasion : direct spread of cancer outside of the liver (e.g., "No evidence of extrahepatic extension," "the tumor may [...] extend from the liver to the right ribs or muscular wall") with values (No, Yes)

Hepatic encephalopathy : confusion or altered consciousness due to liver failure (e.g., "patient denies confusion, forgetfulness, or other symptoms of hepatic encephalopathy," "lactulose" in the medication list) with values (None, Mild, Severe)

Macrovascular invasion : spread of cancer to nearby blood vessels (e.g. "vascular invasion: possible involvement of middle hepatic vein branches," "no evidence of portal vein thrombosis") with values (No, Yes-minor_branch, Yes-major_branch)

Metastasis : spread of cancer to outside the liver, such as to lymph nodes (e.g., "lymph nodes suspicious for metastatic involvement: none," "no lymphadenopathy") with values (No, Yes-regional, Yes-distal)

Portal hypertension : elevation of hepatic venous pressure gradient to greater than 5 mm Hg (e.g., "no evidence of cirrhosis or portal hypertension," "patient had an EGD which showed small varices") with values (No, Yes)

Tumor morphology : number and size of tumor relative to the liver (e.g., "small segment 7 hepatic mass") with values (Massive $\geq$ 50% of liver, Multinodular $<$ 50% of liver, Uninodular $<$ 50% of liver)

Tumor number : number of liver tumors (e.g., "two new liver lesions noted [...] suggesting hepatomas") with values (Single, 2-3, >3)

Tumor size : diameter size of liver tumor (e.g., "1 lesion measuring 2.1 x 1.7 cm [...] HCC") with values (<3 cm, 3-5 cm, >5 cm).

Specifications on which sections to look for stage parameters in a report were formalized into annotation rules. Hepatic encephalopathy, Child-Pugh, and ECOG parameters were marked in clinical notes. Ascites was marked in both clinical notes and radiology notes. The remaining parameters were marked in radiology notes and only in clinical notes if they could not be found in radiology notes. For clinical notes, the annotators started at the "History of Present Illness" section before marking the rest of the note. For radiology notes, the annotators started at the "Impression" section, and if information could not be found there they would move on to "Findings" section or the rest of the report. Repeats of the same information were not annotated. The exception was for ECOG in which all descriptive mentions were also annotated. If multiple pieces of information contribute to the overall value, they were all marked.

4.3 Annotation workflow and software

Annotation occurred in two phases, carried out by two interventional radiologist. In the first phase, relevant parts of reports for the 11 stage parameters were identified by single annotation, where the corpus was divided evenly among the annotators by patient. Annotators marked text annotations using Brat (Stenetorp et al., 2012), a web-based graphical annotation tool, and assigned each a label, e.g. ***ECOG***, and a value, e.g. ***0***. Irrelevant patients, e.g. patients with irrelevant diagnosis, and files, e.g. addenda, abbreviated notes, and post-treatment radiology notes, were flagged for exclusion. Figure 2 shows example mark-ups.

During the second phase, the 3 overall stages and the 11 liver cancer parameters were annotated at the patient level by the consensus annotation of the two annotators. This stage required simultaneous review of all clinical and radiology notes as well as laboratory information related to the patient. The patient level annotation for the 11 liver cancer parameters was necessary to resolve missing and conflicting values from the phase I text annotations. Annotators used a specially built in-house python Tkinter (Hughes, 2000) interface, shown in Figure 3. Annotators had access to the full marked reports as well as a summarized version of their annotations displayed in the interface, along with pertinent laboratory values.

4.4 Text annotations inter-rater agreement

A subset of 20 patients were double annotated for phase I text annotations to calculate inter-rater agreement. After one round of annotations, the annotators met to resolve conflicts and fine-tuned annotation guidelines. We used precision, ($P = \frac{TP}{TP+FP}$), recall ($R = \frac{TP}{TP+FN}$), and F1-measure, ($F1 = \frac{2PR}{P+R}$), to measure inter-rater agreement, where TP is true positives, FP is false positives, and FN is false negatives. True positive matches were measured by label, value, and partial text

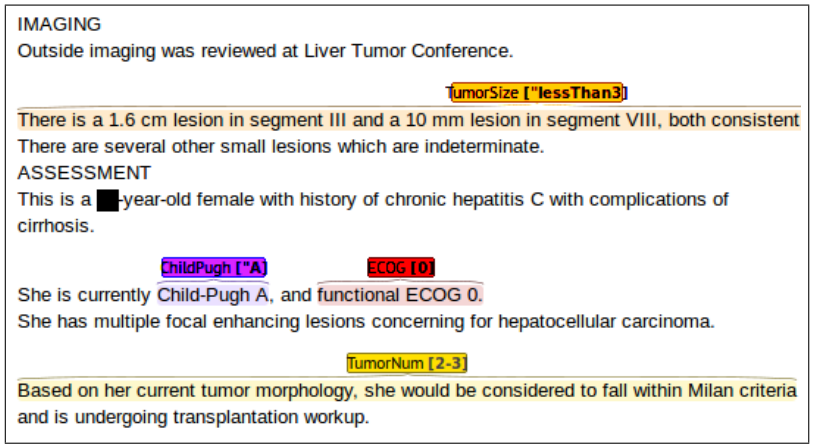

Figure 2: Example phase I text annotations using Brat

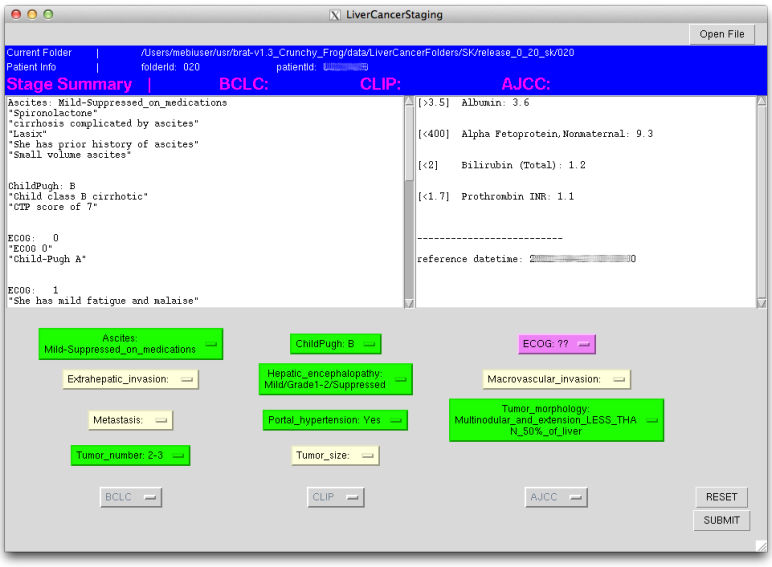

Figure 3: Example phase II patient level annotations. Display panels shows summarized text annotations (left) and lab values (right). Bottom buttons allow annotation of patient level label-values, including the 11 stage parameters (first four rows) and overall 3 stagings (bottom row).

span overlap. For example, two text annotations were considered matching if its label, e.g. ***ascites***, its value, e.g. ***none***, matches exactly and its text spans (document character offsets) overlap. It is also possible, to calculate the agreement of text annotations (still phase I annotations) resolved to the patient level, e.g. if two annotators both identify text spans in any patient 0 file with ***ascites - none***, it is a match.

After the annotator meeting, the microscore agreement for phase I text annotations improved from 0.45 to 0.73 F1. Table 2 shows the final inter-rater microscore agreement consolidated by label. Phase I patient level agreement improved from 0.76 to 0.91 F1. The final patient level agreement breakdown is shown in Table 3. Of the 20 sample patients, 3 patients were excluded due to irrelevant diagnosis.

Discrepancy between the text span and patient levels quantify how often the two annotators find the same information in separate files or different parts of the same document. The higher performance at the patient level was expected given the lower amount of precision needed for patient level agreement.

Ascites, ECOG, and hepatic encephalopathy, had lower agreements because they were often repeated in different expression formats in different report sections. Additionally, one annotator marked ascites drugs while the other did not. Extrahepatic invasion differences were due to one annotator identifying more descriptive information.

Label	TP	FP	FN	P	R	F1
Ascites	10	9	12	0.53	0.45	0.49
ChildPugh	7	0	0	1.00	1.00	1.00
ECOG	23	6	9	0.79	0.72	0.75
Extrahepatic_invasion	6	4	0	0.60	1.00	0.75
Hepatic_encephalopathy	12	3	5	0.80	0.71	0.75
Macrovascular_invasion	16	6	0	0.73	1.00	0.84
Metastasis	10	2	1	0.83	0.91	0.87
Portal_hypertension	11	7	5	0.61	0.69	0.65
Tumor_morphology	15	8	8	0.65	0.65	0.65
Tumor_number	17	6	7	0.74	0.71	0.72
Tumor_size	18	5	5	0.78	0.78	0.78
ALL	145	56	52	0.72	0.74	0.73

Table 2: Phase I inter-rater partial match of label-value per text span, consolidated by label

Label	TP	FP	FN	P	R	F1
Ascites	9	4	2	0.69	0.82	0.75
ChildPugh	6	0	0	1.00	1.00	1.00
ECOG	14	1	3	0.93	0.82	0.88
Extrahepatic_invasion	5	4	0	0.56	1.00	0.71
Hepatic_encephalopathy	8	2	2	0.80	0.80	0.80
Macrovascular_invasion	13	2	0	0.87	1.00	0.93
Metastasis	9	1	0	0.90	1.00	0.95
Portal_hypertension	11	3	0	0.79	1.00	0.88
Tumor_morphology	17	1	0	0.94	1.00	0.97
Tumor_number	17	0	0	1.00	1.00	1.00
Tumor_size	17	0	0	1.00	1.00	1.00
ALL	126	18	7	0.88	0.95	0.91

Table 3: Phase I inter-rater exact match of label-value per patient, consolidated by label

4.5 Phase I annotation statistics

A total of 36 patients and 91 documents were marked for exclusion during phase I of annotation. The total number of patients and associated documents left were 200 and 545, respectively. Of 545 documents, 303 were clinical notes and 242 were radiology notes. There was a total of 2108 text annotations. A breakdown is shown in Table 4.

4.6 Phase II annotation

At the time this paper was written, phase II patient level annotations were still under way, however the corresponding 20 patients used for inter-rater agreement had been staged. For this sample, we found cases where discrepancies in data sources or missing information led to indeterminable stage labels. This occurred for 2 out of 17 non-excluded patients in the 20 patient sample, in which BCLC staging could not be determined due to irreconcilable ECOG values.

5 Analysis of text annotation evidence

In this section, we describe the characteristics of text annotation evidence from the completed phase I of annotations, with the goal of highlighting idiosyncracies or potential challenges for building an information extraction system.

Label	Value	Freq
Ascites	Mild-Suppressed	56
	Moderate-Severe/Refractory	21
	None	189
Child-Pugh	A	73
	B	36
	C	7
ECOG	0	179
	1	102
	2	29
	≥ 3	10
Extrahepatic invasion	No	74
	Yes	3
Hepatic encephalopathy	Mild/Suppressed	48
	None	120
	Severe/Refactory	3
Macrovascular invasion	No	168
	Yes - major branch	24
	Yes - minor branch	10
Metastasis invasion	No	141
	Yes - distal	7
	Yes - regional	8
Portal hypertension	No	16
	Yes	138
Tumor morphology	Massive, $\geq 50\%$ liver	26
	Multinodular, $< 50\%$ liver	56
	Uninodular, $< 50\%$ liver	132
Tumor number	Single	139
	2-3	47
	$>$ 3cm	26
Tumor size	$<$ 3cm	100
	3-5 cm	63
	$>$ 5 cm	57

Table 4: Text annotation statistics

5.1 Data sparsity for severe conditions

Not all values for each parameter label are well populated in our dataset, as shown in Table 4. Typically the more severe cases are less represented in our data. This was probably due to the nature of our exclusion criteria (only new patients were included), as well as the rapidly declining nature of liver cancer. Five year survival rate is less than 20%, with late-stage patients having less than a year to live (American Cancer Society, 2014). Thus, patients diagnosed at more advanced stages may not be referred to the liver tumor clinic. In our system, we will have to handle these cases of class imbalance.

5.2 Overlapping evidence

Studying our annotations, we observed that related stage parameter types may be referenced by the same text evidence. For example, *"Lesion in seg-*

ment 4A measuring 3.9 x 3.6 cm" implies both that there is a single tumor number and a tumor size between 3-5 cm. Similarly, *"Extrahepatic metastatic disease: None"* suggests both that there is no extrahepatic invasion and no metastasis. Knowledge that some parameters may be grouped into the same evidence may be useful when building the system in terms of joint classification or high-level features. Table 5 gives the groupings of the various stage parameters. Parameters in the same group are more likely to have overlapping evidence, though portal hypertension and macrovascular invasion tend to have little overlap with other evidence types. Ascites and hepatic encephalopathy sometimes reference the same passage, e.g. *"no evidence of liver disease sequelae."* Meanwhile, tumor size, morphology, and number rarely do not reference the same text.

Liver/liver disease	Ascites ChildPugh Hepatic_encephalopathy Portal_hypertension
Overall health	ECOG
Tumor	Tumor_morphology Tumor_number Tumor_size
Liver/liver disease AND tumor	Extrahepatic_invasion Macrovascular_invasion Metastasis

Table 5: Stage parameter groupings

5.3 Explicit vs. non-explicit ECOG evidence

In our annotations, we observed a distinction between text annotation evidence that explicitly mentions an ECOG performance status and those that do not. We define text annotation evidence for a stage parameter as **explicit:** if ECOG (Eastern Cooperative Oncology Group) performance status or any of its abbreviations and acronyms are mentioned in the text evidence.

For example, a text annotation highlight for ECOG, e.g. *"ECOG performance 0,"* is considered explicit. Meanwhile, another ECOG reference, e.g. *"He is cachetic. He is deconditioned and needs a wheelchair to walk greater than 10 feet,"* is considered non-explicit. Other non-explicit mentions may consider patient vocation and habits, e.g. *"He continues to work full time as a security officer"* or *"He lives alone and cares for himself without difficulty."*

To get a sense about the complexity of our task, we divided our ECOG text annotation evidence into explicit and non-explicit evidence by iteratively creating rules and manually inspecting the classification. We found that 170 patients out of the 200 non-excluded patients had some mention of ECOG and as much as 23% of these patients have only non-explicit mentions. Because this division is quite dramatic, we plan to build a separate extraction system for explicit and non-explicit ECOG evidence.

5.4 Missing, ambiguous, and conflicting Child-Pugh evidence

From our 200 patient cohort, only 91 patients had some textual mention of Child-Pugh class. This will necessitate that over half of patients shall require Child-Pugh class calculated according to the logic in Table 1. Accurate Child-Pugh identification will then depend on correct extraction of ascites and hepatic encephalopathy variables. Further complicating the issue, we found cases of ambiguity, e.g. *"He has well-compensated liver disease, with Child-Pugh score of 6 or 7 [...] This puts him at a class A/B"* and cases where separate patient documents gave different Child-Pugh scores. After our final patient level annotations, we can evaluate whether calculated versions of Child-Pugh match with the notes' versions.

5.5 Tumor characteristic reference resolution

We observed for our notes, references to tumor characteristics were often equivocal. Not only were there temporal references to disambiguate, e.g. tumor information from previous readings, but also tumors were identified from radiology artifacts such as "lesions," not all of which were actual tumors. Table 6 shows an example in which 3 lesions are found but only one was suspected to be a HCC tumor. Thus, true tumor size and num-

Focal lesions:
Total number: 3:
Lesion 1: segment 4A, cm 6.3 x 7.1 X 6.4 cm ... peripherally located lesions are noted in segment 6 measuring 7 mm .. another ...

Impression:
...
One focal lesion in the segment 4a measuring 6.3 x 7.1 x 6.4 cm.
This lesion does not have a typical appearance, but is ... highly suggestive of HCC.
Exception: 2 smaller lesions noted in segment 6 peripherally, measuring 11 and 13 mm ... are likely to be arteriovenous shunting.

Table 6: Radiology report excerpt

bers will depend on resolving which lesions are actually tumors, as well as handling reference resolution and temporal factors. Tumor morphology additionally must reason about multiple tumors.

5.6 Discussion

Our inter-rater experiment showed that text annotations are being consistently captured with patient level agreement of 0.91 F1 and partial text span level agreement of 0.73 F1.

A limitation to our process is that most of our phase I text annotations were single annotated. Moreover, we assumed that specific text span passages may be attributed with a label and value assignments however some parameters may require a more patient level holistic view. Furthermore, for our study we focused on patients with available laboratory parameters in structured form. This is not always the case when patients are referred by outside organizations.

Although our annotation phase II has not been fully explored here, we have been able to characterize some of the characteristics in our text annotation evidence, which will inform our extraction task. When our patient level annotations are completed, our multi-level annotation will allow us to run several experiments, including: Given gold patient level stage parameters, how well can a system classify staging? Given gold text level stage parameters, how well can a system predict patient level stage parameters?

6 Machine learning baselines

Once our annotation phase I was completed, we wanted to gauge the complexity of our stage parameter information extraction task. To do so, we created a simple document classification baseline to identify information from phase I text annotations. We chose this baseline because of our sparse annotation approach, i.e. a single document may have several occurrences of the same value but may be annotated only once. Our findings from these experiments will be used to advise us of reasonable performance results and issues to consider for our final system.

6.1 Data

The full corpus of patients was randomly divided into a 20% test, 80% training set. The 160 patient training set included 439 documents (243 clinical, 196 radiology) and 1681 text annotations. The test set will be used in a future comparison of the full staging system against a human abstractor. The training set of patients was divided into 5 folds for training and evaluation.

6.2 Methods

Document level classification was performed for each label-value, e.g. ***ascites-none***. The gold standard document label was automatically inferred by the text annotations from annotation phase I (i.e. If **document0.txt** has been highlighted for ***ascites-none***, then **document0.txt** is marked positive for the gold standard in that classification). The classification was binary, since multiple values for the same label may appear in a single document.

Each label-value document classification only classified document types as prescribed by annotation guidelines. For example, tumor size is restricted to classifying radiology document types, since it is possible they appear in clinical notes but are not annotated due to annotation guidelines. Therefore, ascites label-values classifications occurred over all documents (439 documents), Child-Pugh, ECOG, and hepatic encephalopathy classifications were performed over clinical notes only (243 documents), and the remaining label classifications were on radiology notes only (196 documents).

The features included lower-cased unigram, bigram, and trigram counts after tokenization with punctuations removed. We tested four algorithms with default configurations: C4.5 decision tree, discrete-variable decision tree, and maximum entropy from MALLET (McCallum, 2002), and a linear kernel SVM, scaled by min/maximum values, from LibSVM (Chang and Lin, 2011).

6.3 Results

Results are shown in Table 7. The overall classification performance was 0.63 micro-F1 with the highest and lowest F1 at 0.83 and 0.00, respectively. Best performances per label typically came from the highest frequency class. The best classifiers were the two decision trees, however each classifier was the best in at least one classification.

6.4 Discussion

Analysis of the best-performing baseline models revealed some common limitations. One was the inability to capture long-range logical constructions. One example is for ***ascites - none*** and ***metastasis - no***, which often has passages with long-range negation of related terms. And, as

Label	Freq.	Value	Class.	P	R	F1
Ascites	44	Mild	C45	0.24	0.18	0.21
	20	Moderate-Severe	DT	0.50	0.30	0.38
	146	None	DT	0.77	0.36	0.49
ChildPugh	53	A	DT	0.46	0.49	0.47
	25	B	C45	0.84	0.64	0.73
	7	C	DT	0.50	0.14	0.22
ECOG	105	0	C45	0.71	0.71	0.71
	65	1	DT	0.85	0.54	0.66
	18	2	C45	0.89	0.44	0.59
	8	≥ 3	DT	0.25	0.13	0.17
Extrahepatic	59	No	SVM	0.81	0.85	0.83
invasion	2	Yes	$\approx$	0.00	0.00	0.00
Hepatic	34	Mild	DT	0.70	0.76	0.73
encephalopathy	95	None	DT	0.71	0.73	0.72
	1	Severe	$\approx$	0.00	0.00	0.00
Macro-vascular	127	No	NB	0.71	0.96	0.82
invasion	20	Yes-major_branch	C45	0.50	0.55	0.52
	8	Yes-minor_branch	C45	1.00	0.50	0.67
Metastasis	108	No	DT	0.78	0.70	0.74
	6	Yes-distal	DT	0.50	0.17	0.25
	7	Yes-regional	$\approx$	0.00	0.00	0.00
Portal	5	No	$\approx$	0.00	0.00	0.00
hypertension	84	Yes	C45	0.84	0.80	0.82
Tumor	23	Massive	DT	0.37	0.30	0.33
morphology	40	Multinodular, <50%	ME	0.50	0.15	0.23
	105	Uninodular, <50%	NB	0.62	0.80	0.70
Tumor	112	Single	NB	0.64	0.84	0.73
number	32	2-3	DT	0.24	0.25	0.25
	19	>3	ME	0.67	0.11	0.18
Tumor	82	< 3	ME	0.64	0.62	0.63
size	45	3-5	C45	0.43	0.27	0.33
	46	>5	ME	0.59	0.28	0.38
ALL	1551			0.66	0.60	0.63

Table 7: Best baseline performances for training set. (Freq = frequency of positive cases, Class = classifier, C45 = C4.5 decision tree, DT = binary decision tree, ME = maximum entropy, NB = naive Bayes, SVM = support vector machine)

mentioned in Section 5.5, tumor characteristics require reasoning over several sentences.

Another problem was that these simplistic baseline models and features had difficulty normalizing variations in less frequent equivalent evidence. For example, "abdominal distension" and "abdominal girth" both define ascites but neither term is as frequent as "ascites" so did not become strong features. Similarly, many less frequent Child-Pugh acronyms and abbreviations were missed.

Our baselines also lacked the ability to incorporate outside or domain knowledge to infer information. For example, text evidence for ***ascites - none*** and ***hepatic encephalopathy - none*** can be *"he has no known liver disease,"* which requires knowledge that ascites and hepatic encephalopathy are liver disease symptoms. Non-explicit mentions of ECOG, as discussed in Section 5.3, fall under this category as well. There were also cases in which different values for the same label had very similar language, requiring domain knowledge to differentiate. For example, for macrovascular invasion, while, *"There is thrombus in the right posterior branch of the portal vein [...] possibly [...] tumor thrombus"* is considered ***yes - minor_branch***, *"There is enhancing tumor thrombus in the right portal vein."* is ***yes - major_branch***.

Label-value parameters with higher performances, often harbored strong n-gram features. For example "lactulose," a drug to treat hepatic encephalopathy, was found to be used as an early decision point for both ***mild*** and ***none*** values. For portal hypertension, besides "hypertension", "splenomegaly," spleen enlargement often due to portal hypertension, was a top feature.

Some strategies to overcome current limitations are to use medical ontologies and statistical feature selection to identify terms of interest, which can help normalize for term variations. To handle long-range within-sentence relations, we will apply assertion or negation classifiers and use dependency tree parses to build more complex features. For multi-sentence problems such as the tumors, we will use tools for coreference resolution and time parsing. Furthermore, to reduce noise, we may consider using sub-document level classifications, e.g. at the sentence level.

7 Conclusions and Future Work

In our paper, we described our detailed annotation process, carried out an inter-annotator agreement experiment, and analyzed some of the domain challenges and characteristics of our liver cancer patient data. We were further able to present document classification baselines and analyze their performance. In future work, we will improve information extraction over our current baselines by using targeted feature-rich approaches. We will also extend our system to patient level cancer staging, compare results against a human abstractor, and analyze the affects of using multi-levels of gold input. For example, we may experiment with predicting stages using document-level features vs. extracted text level parameter features.

Although we focus on liver cancer, our workflow may be generalizeable to other cancer or phenotype identification annotation tasks. Futhermore, successful liver cancer parameter identification may be useful for other liver cancer staging schemes or other phenotype cohorts.

References

Alex Vianey Callado França, Jorge Elias Junior, BLG Lima, Ana L C Martinelli, and Flair Jose Carrilho. 2004. *Diagnosis, staging and treatment of hepatocellular carcinoma*, volume 37. Brazilian Journal of Medical and Biological Research.

American Cancer Society 2014. *Cancer facts & figures 2014*. Atlanta, (www.cancer.org/research/cancerfactsstatistics/).

American Cancer Society 2014. *Ovarian Cancer*. (www.cancer.org/cancer/ovariancancer/detailedguide/ovarian-cancer-staging).

Andrew Kachites McCallum. 2002. *MALLET: A Machine Learning for Language Toolkit*. (http://mallet.cs.umass.edu).

Anni Coden, Guergana Savova, Igor Sominsky, Michael Tanenblatt, James Masanz, Karin Schuler, James Cooper, Wei Guan, and Piet C. de Groen. 2009. *Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model*, volume 42. Journal of Biomedical Informatics.

Anthony Nguyen, Darren Moore, Iain McCowan, Mary-Jane Courage. 2007. *Multi-class classification of cancer stages from free-text histology reports using support vector machines*, volume 2007. Annual International Conference of the IEEE Engineering in Medicine and Biology Society.

Anthony Nguyen, Michael J Lawley, David P Hansen, Rayleen V Bowman, Belinda E Clarke, Edwina E Duhig, and Shoni Colquist. 2010. *Symbolic rule-based classification of lung cancer stages from free-text pathology reports*, volume 17. Journal of the American Medical Informatics Association: JAMIA.

Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter Embi, Nomie Elhadad, Stephen Johnson, Albert Lai. 2014. *A review of approaches to identifying patient phenotype cohorts using electronic health records.*. Journal of the American Medical Informatics Association (JAMIA).

Chih-Chung Chang and Chih-Jen Lin. 2011. *LIBSVM : a library for support vector machines*, volume 2. ACM Transactions on Intelligent Systems and Technology.

David Martinez and Yue Li. 2011. *Information extraction from pathology reports in a hospital setting*, volume 2011. Proceedings of the 20th ACM International Conference on Information and Knowledge Management.

Hui Wang, Weide Zhang, Qiang Zeng, Zuofeng Li, Kaiyan Feng, Lei Liu. 2014. *Extracting important information from chinese operation notes with natural language processing methods types*, volume 48. Journal of Biomedical Informatics.

Iain McCowan, Darren Moore, Anthony N Nguyen, Rayleen V Bowman, Belinda E Clarke, Edwina E Duhig, and Mary-Jane Fry. 2007. *Collection of cancer stage data by classifying free-text medical reports.*, volume 14. Journal of the American Medical Informatics Association: JAMIA.

Jonathan C Yau, Arlene Chan, Tamina Eapen, Keith Oirourke, Libni Eapen. 2002. *Accuracy of the oncology patients information system in a regional cancer centre.*, volume 9. Oncology Reports.

Katherine A McGlynn, W Thomas London. 2011. *The Global Epidemiology of Hepatocellular Carcinoma: Present and Future.*, volume 15. Clinics Liver Disease.

Kwang-Hyub Han, Masatochi Kudo, Sheng-Long Ye, Jong Young Choi, Roonni Tung-Ping Poon, Jinsil Seong, Joong-Won Park, Takafumi Ichida, Jin Wook Chung, Pierce Chow, and Ann-Lii Cheng. 2011. *Asian Consensus Workshop Report: Expert Consensus Guideline for the Management of Intermediate and Advanced Hepatocellular Carcinoma in Asia.*, volume 81. Oncology.

Naveen Ashish, Lisa Dahm, Charles Boicey. 2014. *University of California, Irvine-Pathology Extraction Pipeline: the pathology extraction pipeline for information extraction from pathology reports.*, volume 20. Health Informatics Journal.

Phil Hughes 2000. *Python and Tkinter Programming*, volume 2000. Linux J.

Pontus Stenetorp, Sampo Pyysalo, Goran Topi, Tomoko Ohta, Sophia Ananiadou and Jun'ichi Tsujii. 2012. *brat: a Web-based Tool for NLP-Assisted Text Annotation*. In Proceedings of the Demonstrations Session at EACL 2012.

Ramakanth Kavuluru, Isaac Hands, Durbin Eric B, and Lisa Witt. 2013. *Automatic extraction of icd-o-3 primary sites from cancer pathology reports*. In AMIA Jt Summits Transl Sci Proc.

Tracy Sexton, George Rodrigues, Ed Brecevic, Laura Boyce, Denise Parrack, Michael Lock, David D'Souza. 2002. *Controversies in prostate cancer staging implementation at a tertiary cancer center*, volume 13. The Canadian Journal of Urology.

Vianney Jouhet, G Defossez, Anita Burgun, P le Beux, P Levillain, and Pierre Ingrand. 2012. *Automated Classification of Free-text Pathology Reports for Registration of Incident Cases of Cancer.*, volume 51. Methods of Information in Medicine.

William K Evans, Juanita M Crook, D Read, J Morriss, and DM Logan. 1998. *Capturing tumour stage in a cancer information database.*, volume 2. Cancer prevention & control.

Xiao-Ou Ping, Yi-Ju Tseng, Yufang Chung, Ya-Lin Wu, Ching-Wei Hsu, Pei-Ming Yang, Guan-Tarn

Huang, Feipei Lai, Ja-Der Liang. 2013. *Information extraction for tracking liver cancer patients statuses: from mixture of clinical narrative report types*, volume 19. Telemedicine Journal and E-Health: The Official Journal of the American Telemedicine Association.

Ying Ou and Jon Patrick. 2014. *Automatic Population of Structured Reports from Narrative Pathology Reports types*, volume 163. Proceedings of the Seventh Australasian Workshop on Health Informatics and Knowledge Management. Darlinghurst, Australia.

Yongyut Sirivatanauksorn and Chutwichai Tovikkai. 2011. *Comparison of Staging Systems of Hepatocellular Carcinoma*, volume 2011. HPB Surgery.