Information extraction from clinical and radiology notes for liver cancer staging

Wen-wai Yim

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Meliha Yetisgen, Chair

Sharon Kwan

Lucy Vanderwende

Fei Xia

Program Authorized to Offer Degree:
Biomedical and Health Informatics

University of Washington

Abstract

Information extraction from clinical and radiology notes for liver cancer staging

Wen-wai Yim

Chair of the Supervisory Committee:
Associate Professor Meliha Yetisgen
Biomedical Informatics and Medical Education

Medical practice involves an astonishing amount of variation across individual clinicians, departments, and institutions. Adding to this condition, with the exponential pace of new discoveries in biomedical research, medical professionals, often understaffed and overworked, have little time and resources to analyze or incorporate the latest research into clinical practice. The accelerated adoption of electronic medical records (EMRs) brings about great opportunities to mitigate these issues. In computable form, large volumes of medical information can now be stored and queried, so that optimization of treatments based on patient characteristics, institutional resources, and patient preferences may be data driven. Thus, instead of relying on the skillsets of patients' support network and medical teams, patient outcomes can at least have some statistical guarantees.

In this dissertation, we focused specifically on the task of hepatocellular carcinoma (HCC) liver cancer staging using natural language processing (NLP) techniques. Staging, or categorizing cancer patients by extent of diseases, is important for normalizing over patient characteristics. Normalized stages, can then be used to facilitate research in evidence-based medicine to optimize for treatments and outcomes. NLP is necessary, as with other clinical tasks, a majority of staging information is trapped in free text clinical data.

This thesis proposes an approach to liver cancer stage phenotype classification using a mixture of rule-based and machine learning techniques for text extraction. Included in this

approach is a careful, layered design for annotation and classification. Each constituent part of our system was characterized by detailed quantitative and qualitative analysis. Two important modules in this thesis are a framework for normalizing text evidence related to specific conditions and an algorithm for tumor reference resolution.

The overall results of our system revealed an F1 performance of 0.55, 0.50, 0.43 for AJCC, BCLC, and CLIP liver cancer stages, respectively. Although outperforming baseline classifications, these accuracies are not viable for clinical use. Analysis of error suggests that performance for some constituent stage parameters would improve through additional annotation. However, one identified crippling bottleneck was the requirement of reference resolution and discourse-level reasoning to determine the number of tumors in a patient, a crucial part of cancer staging.

Still our work provides a methodology to classify a complex phenotype, whose strength includes its interpretability and modularity while maintaining ability to scale and improve with greater amounts of data. Furthermore, submodules of our system, for which perform at higher accuracies, may be used as tools to decrease annotation costs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AJCC: American Joint Committee on Cancer

BCLC: Barcelona Clinic Liver Cancer

CLIP: Cancer of the Liver Italian Program

CRF: Conditional random field

CT: Computed tomography

ECOG: Eastern Cooperative Oncology Group

EMERGE: Electronic medical records and genomics

EMR: Electronic medical record

HCC: Hepatocellular carcinoma

ICD: International classification of diseases

LI-RADS/LIRADS: Liver Imaging Reporting and Data System

MRI: Magnetic resonance imaging

MEMM: Maximum entropy markov model

NIH: United States National Institute of Health

NLM: United States National Library of Medicine

NLP: Natural language processing

SNOMED-CT: Systematized nomenclature of medicine - clinical terms

SVM: Support vector machine

ACKNOWLEDGMENTS

Completion of this thesis would not be possible without the help of the persistent pacific northwest precipitation, making indoor activities ever much more so attractive. However, perhaps someone that was even more influential these couple years than the weather was my advisor, Meliha Yetisgen. I would like to thank her for her advice, encouragement, and belief in me. I am grateful that she introduced me into the wonderful world of natural language processing. Even if I could have done this without her, I doubt it would have been as fun.

I would like to thank my committee members, Sharon Kwan, Fei Xia, Lucy Vanderwende, and Gina-Anne Levow for their eyes, ears, and insights. I am indebted to Sharon for bringing me such an exciting thesis project, but most of all for her dedication and patience. I am grateful to Fei for instilling in me the finer points of corpus annotation, and to Lucy for her frequent much-needed thoughtful suggestions.

I owe my thanks to the University of Washington Linguistics program and Computer Science and Engineering program for providing such a rich environment, resources and people included, to learn about NLP. It has really been such a pleasure. I thank the Biomedical and Health Informatics Program staff and professors for providing a warm and supportive environment.

Thank you to my BHI cohort, Alan Kalet, Logan Kendall, Amanda Lazar, Leslie Liu, Hannah Mandel, and Albert Park, and my comrade-in-arms, Prescott Klassen, for their friendship and the multitudes of help and advise. I thank my dear friends Geraldine Chan, Spring Sun, Jonathan Joe, Kelli Xu, Sarah Larsen, Silvia Wu, and my Husky Badminton friends for their kindness and encouragement when I required it the most.

Finally, I would, most of all, like to thank my family. I want to thank my parents who

worked very hard to give my sister and I the safety, security, and opportunities they were never afforded and who have always been there for me with love and support, regardless of the circumstances; my sister who, when I was unsure and bright-eyed, was my first teacher (sometimes to my unfortunate detriment); and all the rest of my family, my grandmas, aunts, uncles, and cousins who watched me grow.

Chapter 1

INTRODUCTION

1.1 Context and motivation

Medical practice involves an astonishing amount of variation. At all levels, between individual clinicians, departments, institutions, local regions, and countries, there will be differences in diagnoses for the same symptoms, different techniques for the same procedures, and different recommendations even if the same diagnosis is agreed upon. Moreover, with the exponential pace of new discoveries in biomedical research, medical professionals, often understaffed and overworked, have little time and resources to analyze or incorporate the latest research into clinical practice. Patients and their family members, on the other hand, with whom the burden of choosing clinical options and navigating the healthcare system usually falls, are ill-equipped to understand the consequences to the medical expenditures they find themselves unexpectedly needing.

The accelerated adoption of electronic medical records (EMRs) brings about great opportunities to mitigate these issues. In computable form, large volumes of medical information can now be stored, queried, and transferred across large distances. Although there are many issues to resolve regarding data sharing and interoperability, the implications of this technology is enormous, a fact that is captured in the large investments in EMRs globally. For example, the United States through the Health Information Technology for Economic and Clinical Health (HITECH) Act provision in the 2009 American Recovery Act has made large investments towards EMRs [1]. In the 2015 inaugural address, President Obama introduced the precision medicine initiative which strives to collect data for research such that personal characteristics, e.g. genes, environment, and lifestyle, can be incorporated into health decision making [11]. Elsewhere, the European Commission has laid out a roadmap for 2012-

2020 emphasizing innovations and advancements in eHealth as a means of improving disease management and prevention [5]. China, where most urban hospitals have adopted EMRs, has further made large allocations to accelerate EMR adoption in rural hospitals [186].

Here we concentrate specifically on the use of EMRs as a means of conducting evidence-based research[1], which may decrease variations in treatments by collecting data analytics. This can lead to, at least, data-driven guarantees, and less reliance on the particular skills and resources of individual clinicians, patients, and institutions to optimize courses of action. Ideally, every patient's episode of treatments should be saved and compared to others as a series of natural experiments. Eventually, it would then be possible to personalize treatments based on the best treatment outcomes for a patient type, given resource constraints, and individual patient preferences. However, as it is now, it is all too common for clinicians and patients to be unaware of the full host of previous examples available. To this end, we point to natural language processing (NLP) technology, as a means to help clinicians in shifting, aggregating, and processing medical evidence, for which a majority of data is in free text form, making it possible to "close the loop" between clinical practice, research, and education. [177]

1.2 Problem description

This thesis project utilizes EMR data, most of which are free text clinical notes, to identify and normalize information relevant for patients with liver cancer. The overall objectives are to be able to store and compile the normalized information so that, in the future, they can be combined with other data sources, e.g. procedure codes or genetic information, for evidence-based research.

The specific disease we characterize is hepatocellular carcinoma (HCC), a prevalent form of liver cancer. The information we seek to identify and normalize are factors related to

[1]We define evidence-based research as retrospective clinical research based on patient clinicl data. We make this distinction to differentiate apart from prospective clinical research, e.g. a randomized clinical trial. On the other hand, evidence-based medicine is the application of a clinician's personal experience in addition to outside systematic research studies to perform the best course of treatment.

liver function and cancer spread, also used to calculate liver cancer stages. In the following subsections, we motivate the need of evidence-based research in HCC, the importance of staging in cancer patients, and the necessity of NLP to meet these ends.

1.2.1 Need for evidence-based treatment for HCC

Hepatocellular carcinoma (HCC) is one of the leading cancer-related causes of death and the most common primary hepatic malignancy worldwide [191]. Though a majority of cases occur in Africa and East Asia, the incidence of HCC in developed countries is rising [192]. In the United States, mortality rates due to liver cancer has increased even as deaths related to other cancers in the same time frame have decreased [18]. Only 30% of patients diagnosed at an early stages[2] survive to 5 years, this rate decreases to 11% and 3% for patients with regional[3] and distal[4] metastasis [156].

Global comparisons show that HCC manifests at different ages in different countries, with younger populations in Africa and older populations in East Asia and North America. This is attributable to regional variations in hepatitis B virus (HBV) and hepatitis C virus (HCV) infection rates, which are strong causal factors for HCC [113]. Other cofactors include aflatoxin B1, a common mycotoxin that contaminates foodstuffs in Africa, and cirrhosis related to excessive alcohol consumption, diabetes, and obesity. HBV vaccination and sterilizing medical instruments practices have played key roles in combating HBV and HCV spread, and in turn HCC epidemics. For the United States, the most prevalent cofactors of HCC is shifting from HBV and HCV to obesity and diabetes. In fact, recent studies have identified NASH (non-alcoholic steatohepatitis), liver inflammation and fat accumulation not due to excessive alcohol intake and associated with obesity and diabetes, as an emerging HCC cofactor [192]. Furthermore, interestingly, overseas immigrant populations of ethnic groups with higher prevalence of HCC, have shown higher incidence rates to the surrounding popu-

[2]In early stages, tumors remain small and have not invaded nearby major blood vessels or lymph nodes.

[3]Tumors have invaded to nearby lymph nodes

[4]Tumors have invaded to other organs

lation, but lower rates compared to the populations of their place of origin [113]. Altogether, evidence suggests various genetic, environmental, and lifestyle components to disease susceptibility and progression.

The liver is a vital organism involved with multiple systems of the body, including glycogen storage, hormone production, and detoxification. Because of the conditions of HCC onset which often involve cirrhosis, HCC patients typically have multiple immediate comorbidities. In the presence of such constraints, clinicians must reconcile between the competing health risks associated with specific HCC progression and liver failure for treatment options [33].

Unfortunately, the variations in therapy differ as much as the number of conditions HCC can manifest in. Patients may be treated with surgical interventions, such as transplant and resection, as well as local regional treatments, such as radiofrequency ablation and transarterial chemoembolization [62][191]. In 2009, Sorafenib, an oral systemic therapy, was demonstrated to have significant improvements in survival of HCC patients, making chemotherapy another option. During their treatment progression, patients may consult a range of specialties including surgery, interventional radiology, and oncology [162]. However, ultimately, the final strategies depend on the institution's resources and the experience of local clinicians as well as the patient's tumor characteristics, liver function, comorbidities, and preferences. Thus, evidence-based clinical guidelines that account for the best treatments in the face of the diversity of disease manifestation and demographics is highly desirable.

1.2.2 Role of staging on HCC evidence-based research

Previous work in developing HCC guidelines typically prescribe treatment based on liver cancer stages or select liver function variables [73][33]. Staging is used to summarize the extent of disease for cancer patients. Each cancer domain may have different criteria for its stages. For example, prostate cancer stages use the Gleason score to measure likelihood of tumor spread based on morphology of prostate cancer tissue [159].

For liver cancers, patient performance status as well as liver function variables are incorporated into various staging schemes. The Barcelona Clinic Liver Cancer (BCLC) staging

system is one of the most widely used liver staging systems and links therapy plans to specific stages [33]. Recent advances in HCC treatment plans have led to more options not included in the original BCLC staging scheme. Furthermore experience at various institutions have have suggested loosening existing BCLC staging restrictions of prescribed techniques. For this reason, institutions often adapt their own guidelines to meet their needs. The Alberta HCC and the Japan Society of Hepatology (JSH) algorithms, for example, extends the BCLC staging with recognition of radiofrequency ablation for very early stages, liver transplantations for Child-Pugh class C patients, Sorafenib for Child-Pugh class A and B patients, and adds transarterial chemoembolization and transarterial radioembolization with 90 Yttrium as treatments. Other work, such as the Korean and Chinese guidelines in [73] look at individual stage parameters and codify appropriate treatments based on their own designed logic. Other HCC staging systems that are linked to survival rates exist, including Cancer of the Liver Italian Program (CLIP), Japanese Integrated Staging (JIS), and Chinese University Prognostic Index (CUPI) [136]. Each staging scheme represents distinct measurements of cancer spread and liver function.

Despite their use in research, cancer stages are not always recorded in the electronic medical record in structured or unstructured forms [112]. For example, even under mandated collection across all cancer type stages, an Ottawa Regional Cancer Center study had shown only 71.5% average completion [59]. Even when stage information is available, for reasons of data entry errors or incorrect application of staging guidelines, they are often inaccurate [99][148][193][151]. The absence of liver cancer staging in medical records reflects both the ongoing debate for which staging system to use, which are subject to revision, and the primary use of HCC staging as a research tool. The goal of our project is to facilitate HCC research by using NLP to automatically induce liver cancer staging. The work could not only go towards development of a nuanced evidence-based guideline for HCC disease but also provide useful data on HCC progression and treatment per demographics.

1.2.3 Leveraging NLP for liver cancer staging phenotype evidence-based research

Currently, a large portion of medical records, including information needed for our stage and stage parameters, are recorded in free-text narrative form. For example, liver disease symptoms such as *ascites* or *hepatic encephalopathy* may not be included in the problem list; nor may items such as tumor number or size be recorded as structured data[5]. This data is in a large part locked within free text daily nurse or physician notes, radiology reports, admit and discharge notes of the clinical record, where both minor and significant medical events may be recorded.

Nuances in findings or uncertainties are better captured in free text writing and it is infeasible for clinicians to code for every single possible metric for secondary purposes. Therefore, the advantage of using NLP for automatic information extraction is the ability to extract information into structured forms without adversely altering clinical workflow. A successful system would expedite HCC research on large diverse corpora of patient notes while circumventing the high cost of manual review by specialists. Moreover, automating liver cancer staging can accommodate new or changing staging schemes as NLP methodology then only requires new or updated gold annotations on patient records to retrain the liver cancer staging component. Importantly, automated liver cancer staging allows research over historical patient records in retrospective chart review studies, providing access to a potentially unlimited amount of outcome data for new hypotheses.

1.3 Contributions and objectives

The objectives of this project are to automatically extract and normalize text information and predict cancer stages relevant for three liver cancer stages. This project confronts several issues: (1) in-depth annotation of patient-related medical conditions, (2) identification and normalization of text evidence to symptoms and their severities, and (3) classification of patient characteristics. This work makes the following contributions: analysis of annota-

[5] Structured data is information that is collected with some assumed internal structure. Two examples are drop-down menu input or spreadsheet input.

tion issues for multi-layered phenotypes, a method for sparse annotation of radiology report findings, a statistically driven system to identify and normalize text evidence related to a clinical phenotype, experiments for tumor reference resolution classification, and experimentally driven results of patient staging.

Taking a broader view, we may additionally appreciate the research in this dissertation set in the larger backdrop of advancing medical evidence-based research and the sub-specialty of NLP in clinical text. Although we are motivated by a specific need, the methods and questions addressed here very much corroborate the growing trend of deep clinical phenotype[6] extraction. Though large scale projects which leverage NLP-extracted data and pair such data with traditionally structured information, such as ICD codes, to define complex phenotypes have become more and more ubiquitous, current methods have only scratched the surface. However, how to best integrate developing clinical NLP tools (which typically give less than ideal performances) into complex tasks of patient phenotype prediction with reasonable clinically-useful predictions remains open for exploration.

1.4 Guide for the reader

The structure of this dissertation describes the overall project with distinct chapters focuses on particular parts. However, Chapters 5, 6, and 7 describe sub-projects that include their own more specific literature reviews. A summary of each chapter is described following:

Chapter 2 : This chapter starts with a background on clinical NLP and briefs the reader on related works relevant to cancer of the liver prediction.

Chapter 3 : This section describes our corpus annotation for liver cancer stage parameter extraction and stage classification.

Chapter 4 : In this chapter, we describe the overall system architecture for our cancer stage prediction. Each component and how they fit within the entire system are briefly

[6]We define phenotype as a categorical label which characterizes a set of observable traits.

reviewed with references to later chapters. We also provide the results of a simple document baseline to which the rest of sub-patient classifications are measured against.

Chapter 5 : This chapter describes two special stage parameters: *ECOG* and *Child-Pugh*. We provide a discussion into their explicit and non-explicit representations and describe our rule-based extraction for them.

Chapter 6 : This section describes our statistical feature based classification on sentences to identify stage parameters: *ascites*, *hepatic encephalopathy*, *macrovascular invasion*, *metastasis*, and *portal hypertension*.

Chapter 7 : This chapter features a sub-system used to extract tumor characteristics. The target values include the stage parameters for tumor number, size, and morphology. This sub-project describes its own set of corpus annotations, system creation, training, testing, and evaluation used for radiology report tumor information. Tasks in this section include entity and relation extraction, reference resolution, and rule-based tumor characteristics extraction.

Chapter 8 : In this chapter, we piece together multiple submodules to create a patient level classifiers for 3 liver cancer stages and 11 stage parameters.

Chapter 9 : This chapter features an experiment comparing a trained non-expert against expert annotation and make comments compared to our system.

Chapter 10 : In the final chapter, we summarize the dissertation and describe directions for further work.

In addition to the above-mentioned chapters, the included appendices include created resources, such as annotation guidelines and word-lists, generated throughout this thesis.

There are several themes generally relevant to clinical informatics, but manifested in several specific instances in our work. We briefly describe them here as well as give a discussion touching these points at the end of this dissertation.

Structured versus unstructured input : Though structured input is preferable for research, it is highly inconvenient for the primary purposes of clinical record-keeping. This entire project is premised on needing to calculate liver cancer stages, which can be in theory coded as a structured input. Digging deeper, many individual staging variables are also sub-stages. For example, *Child-Pugh* liver disease stages and *ECOG* statuses can further be inputted as structured input. Where to achieve the balance between the two modalities in the clinic is an open question that is likely to change with the evolution of health care practices.

Medical conditions have arbitrarily complex signs and symptoms : Medical concepts are abstract but the observable signs used to diagnose them can be numerous and highly complex. For example, *portal hypertension* may be identified by any number of observable traits, e.g. splenomegaly (enlarged spleen) or unblocked blood vessels. As technology changes, the signs and symptoms tied to abstract concepts and the abstract concepts themselves may change in definition and in medical characteristics.

Information completeness : Information completeness is a difficult issue in regards to patient records. Patients may transfer care from outside hospitals, emergency room visits may not be properly connected to the correct records, etc. Problem lists or billing data may be incomplete; alternatively, certain details can be assumed to be true given context or some may simply have not been collected. Related to this idea is information consistency, as it is possible to have discrepant information from two data sources. These are factors to be considered when making judgements about patients.

Chapter 2
BACKGROUND

In this section, we first give a brief overview of natural language processing in the clinical domain, in general. Afterwards, we provide deeper discussion with closely related work. Finally, the chapter ends with a brief appreciation of where the work in thesis lies with respect to other biomedical research.

2.1 Natural language processing in the clinical domain

Unlike news or classical English literature, clinical documents are used as transitory notes among a patient's medical team rather than heavily edited information for wider communication. As a consequence, clinical documents may be ungrammatical and composed of short telegraphic phrases. They can include copy-and-paste results, misspellings, local abbreviations, and special formats [116]. As a domain, biomedicine is notorious for a variety of synonyms, acronyms, and abbreviations [118]. Thus, there are many ways to express a concept and each can have multiple abbreviations. Meanwhile, a single acronym can mean many different things. Furthermore, clinical documents have the idiosyncrasy of sections and section headings. Section headings of clinical documents demark parts of text as separate regions for organizational purposes. However, the number and type of sections are not strictly controlled vocabularies and subsections can occur spontaneously with little orthographic indication of hierarchy.

These nuances are highly variable depending on local institution, department, and clinician. Furthermore, even though some report types may include parts with implicit templates, e.g. outputted as tab delimited formats, methods extracting information using explicit templates, e.g. regular expressions looking for a particular pattern, may not generalizable to

other report types, departments, or institutions.

While a human interpreter may easily be able to navigate between these fluctuating disparities, it is much more difficult for a machine. Therefore, upstream NLP tasks that are considered solved for the larger English NLP domain such as sentence tokenization or parts-of-speech tagging become more problematic or irrelevant for some sections. Higher-level tasks such as chunking or sentence parsing, are even less developed.

2.1.1 Historical and existing medical NLP systems

Beginning in the late 1980s, computational linguistics and artificial intelligence (AI) was applied to the medical field. The ability to focus on a narrow domain, with a reasonable amount of knowledge consistency regardless of language, and the composition of medical words by Greek and Latin parts made the area attractive to NLP and AI researchers. Early systems, many written in Prolog and Lisp, were characterized by carefully constructed grammars linked with detailed knowledge representations, many in the form of concept graphs and frames. Though we do not give a detailed account of each here, we point the reader to an excellent review by Spyns et al [161].

Existing medical NLP processing systems are often developed in university hospital centers as large processing suites and may include section identification, negation detection, and modifier extraction, in addition to concept identification and some syntactic or semantic analyzer modules. Briefly, we name some important systems. Most notable is Columbia University's MedLEE [66][52], one of the oldest systems in continual usage. University of Utah has developed several iterations of medical NLP systems such as SymText, SPRUS, and MPLUS [75][40]. Other suites and their affiliated centers used today include the REgenstrief data eXtraction (REX) tool from Regenstrief Research Institute in Indianapolis, Indiana[64][63], MediClass from Kaiser Permanente Center for Health Research in Portland Oregon [76], MEDSYNDIKATE from Freiburg University Germany [72], MTERMS from Partners Healthcare system and Brighman and Woman's Hospital [200] and MedTAS/P (Medical Knowledge Text Analysis System/Pathology) from IBM, in collaboration

with Mayo Clinic [41].

Among freely available processing tools are NLM's (National Library of Medicine) MetaMap, which first identifies noun phrases then matches to Unified Medical Language System (UMLS) concepts by string matching [22], cTAKES (clinical Text Analysis and Knowledge Extraction System) developed from Mayo Clinic, which performs dictionary look-ups of UMLS terms [147], and HITEx (Health Information Text Extraction) from Brigham and Women's Hospital and Harvard Medical School, currently integrated into i2b2's NLP suite and based on the GATE platform, which maps to UMLS concepts by first attempting exact matches to noun phrases and subsequently trying to stem and truncate strings [198].

2.1.2 Clinical NLP challenges and datasets

Annotated corpora are important resources for building NLP systems, especially for providing example data for statistical systems. However, creation of clinical corpora is especially difficult due to privacy laws and the sensitive nature of personal medical information. We would be remiss if we did not devote some space to discuss the clinical text NLP challenges that have been so impactful in this domain.

One of the first efforts for providing de-identified public clinical text was realized through the 2007 Medical NLP Challenge which assigned ICD9CM (The International Classification of Diseases, Ninth Revision, Clinical Modification) codes to radiology reports [131]. Since then several datasets, releaseable under data use agreements have been made available through other challenges.

The i2b2 NLP challenges which continue once every couple of years are perhaps the most well known. Starting in 2009, the i2b2 Medication Challenge made clinical discharge summaries with annotations of medication and fields pertaining to the medication (e.g. dosage, mode, frequency, duration, reason, and associated narratives) available for public use [128]. The next 2010 i2b2/VA challenge focused on medical concept identification, assertion[1] clas-

[1] Assertion classification labels text as one of several categories: not associated with the patient, hypothetical, conditional, possible, absent, and present.

sification, and relation extraction [176]. The 2011 i2b2/VA/Cincinnati Children's Hospital Medical Center Shared-Tasks featured concept and pronoun coreference for one subtask and sentiment classification on suicide notes for another [14][175][132]. The 2012 i2b2 temporal relations challenge involved medical event concepts and linking them with temporal expressions such as dates, times, durations, or frequencies [171]. The most recent 2014 i2b2/UTHealth Shared-Task challenge focused on public health information de-identification and identification of heart disease risks [6][168][167].

The Clinical E-Science Framework (CLEF) corpus from the Royal Marsden Hospital (RMH) and the University of Sheffield, UK is annotated with clinical entities and their relations [141]. In a collaboration between the Shared Annotated Resources (ShARe) project funded by the US NIH and CLEF, the 2013 ShARe/CLEF eHealth challenge targeted the identification and mapping of acronyms, abbreviations, and disorders in clinical text [172]. The goal of the subsequent 2014 ShARe/CLEF e-health evaluation lab task 2 targeted template-filling of disorders and its attributes [89].

In growing recognition of the domain, recently, general NLP shared tasks have included clinical NLP tasks in their events. In 2011, the TExt Retrieval Conference (TREC) Challenge, hosted by the United States NIST (National Institute for Standards & Technology) annually, included a Medical Records Track which made de-identified patient records and their associated ICD-9 codes available for identification of clinical study eligibility [56]. The SemEval-2014 Task 7 shared task released the ShARe/CLEF data for concept identification and mapping of acronyms, abbrevations, and disorders in clinical text [137]. The SemEval-2015 Task 14, corresponding to the 2014 ShARe/CLEF eHealth challenge, included identification of disorder attributes in a template filling task [12][57].

The most deeply linguistically-annotated clinical text corpus is the MiPACQ clinical corpus out of Mayo Clinic and the SHARPn dataset which includes several layers of annotation, including treebank, PropBank, and UMLS semantic type annotations [16][51]. The THYME (Temporal History of Your Medical Events) corpus is also comprised of de-identified notes from the Mayo Clinic and marks events focusing on clinical occurrences and their temporal

relations [169].

2.2 Cancer stage prediction from clinical records

Automatic cancer staging from free text clinical notes has been explored for other cancers such as lung cancer [121][122][112], colorectal cancer [108], and general pathology reports [39], though work in the liver cancer staging domain specifically is relatively sparse. Like other cancers, the number of tumors, the size of the tumor, tumor metastasis and etc. are key characteristics to capture for cancer staging. Unlike other cancers, liver cancer staging additionally considers performance status and liver function variables [136]. Thus the problem of liver cancer staging involves extraction tasks of multiple different elements, including tumor characteristics and liver-related metrics, as well as an inference step to determine stage based on extracted evidence. Our overview of related work divides literature into those that predict cancer stages and those that extract cancer characteristics. Afterwards, we describe related subdomains.

2.2.1 Predicting cancer stage

One well-explored area of cancer stage prediction is for lung cancer patients from the Queensland Integrated Lung Cancer Outcomes Project data and the Queensland Health Pathology Information System (AUSLAB). In several papers, Nguyen et al and McCowan et al, progress from simpler approaches to increasingly complex setups to predict TNM cancer staging[2]. In [121], Nguyen et al perform support vector machine (SVM) document classification with concept normalization, negation detection, and term frequency weighting using different hierarchial set-ups for multi-class classification. Their accuracies for T and N sub-stages were 64% and 82% respectively. In [112], the cancer stage document classification problem was

[2]In TNM staging, a stage is made up of 3 sub-stages each with its own values. For example, one stage classification may look like T1 N0 M0. T, N, and M values signal tumor size, lymph node spread, and distant metastasis, respectively, and the values accompanying them signal different levels of severity. For example, T0 means no signs of tumor while T1, T2, T3, and T4 refers to increasing size or extension of a tumor.

subdivided into a number of sentence level classifications each with two-steps. First, a sentence is classified to be relevant or not to a particular TX, T0, Nx, N0, etc. Second, if relevant, the sentence is classified to be positive or negative.[3] The final stage document level stage was determined by starting from the lowest sub-stages and performing the relevant classifications until reading the highest sub-stage classification. Their T and N accuracy was improved to 74% and 87% respectively. Finally in [122], the group used a symbolic logic approach. Rules were developed with handling of conjunction with concept-normalization, negation, and normalization through the SNOMED-CT hierarchy. Their accuracies using these methods were 72% and 78% for T and N respectively.

Martinez and Li [108] similarly predicts staging (T, N, and Australian Clinico-Pathological (ACPS) Staging system), number of tumors, as a document classification task for nominal stage parameters and a sentence classifier for numerical parameters. Their dataset included colorectal cancer pathology reports from the Royal Melbourne Hospital. In this case, for T and ASCPS sub-stages, a sentence is classified as relevant or irrelevant. Subsequently, they extracted the numerical values in the sentence, using the value closest to the median from the training set. Two strategies of N sub-stage classification were compared. One strategy, involved document classification, the other involved using lymph node knowledge from a previous classification with a set of heuristics. They tested a variety of machine learning, e.g. SVM and naive Bayes (NB), and rule-based methods along with different SNOMED, UMLS, and lemma features with best F1 scores at 82%, 81%, and 75% for T, N, and ACPS staging respectively. Martinez et al further adapted their system to notes from another hospital, Barwon Health [107].

Cheng et al [39], sought to predict the cancer progression, which was made up of a 3-tuples with *status*, *magnitude*, and *significance* fields, in free-text MRI (magnetic resonance imaging) radiology reports. An example of one class is (stable, moderate, uncertain). Their

[3]Each sub-stage classification problem could have multiple two-part classifiers. For example, to classify N1, there was one two-step classifier trained to determine "peribronchial lymph node involvement" and one two-step classifier trained to determine "hilar lymph node involvement" – both whose end result is a N1 relevant or irrelevant. For details, the reader is encouraged to read the referenced paper.

approach for statical classification of *status* began with creating subdocuments by looking up pairs of subject and status words in as adjacent sentences, e.g. "mass" and some indication of progression, dividing leftover sentences to the nearest subdocument. Stop words were removed and instances were vectorized by bag-of-words and negation features and classified using an SVM. The probability of the entire document belonging to a class was calculated by normalizing over its subdocuments. Sensitivity and specificity for *status* performed at 81% and 92%. The pattern-matching modules for *magnitude* and *significance* performed at sensitivities and specificities of 79% and 89% and 67% and 86%, respectively.

There is also some work with cancer staging using structured data. [117] and [134] predicted cervical cancer Federation Internationale de Gynecologie et d'Obstetrique (FIGO) staging using neural networks with a set pre-determined parameters, achieving accuracies of 73% and 80% respectively. Their input parameters resembled other staging information such as tumor diameters and lymph node enlargements, but also had cervical cancer specific parameters such as cervical involvement and vaginal involvement.

2.2.2 Extracting cancer characteristics

Most work on clinical NLP cancer staging emphasizes *extraction* of cancer stages or cancer stage parameters over *prediction*. We make this distinction to clarify between cases where the stage is explicitly mentioned in text, e.g. "*she is a stage 1 breast cancer patient*". In contrast to the previously described works, these works only assign a cancer stage if it is explicitly found in a document, along with other cancer-related information such as tumor number and size.

Rule-based systems for these tasks typically involve a dictionary look-up, negation handling, and heuristic algorithms to structure results. These works can be divided among those that extract tumor or cancer characteristics and those that focus on cancer case-finding. In the first set of works, systems extract multiple pieces of information from radiology and pathology reports. Scores range widely between systems as well as between distinct variables within a single system, depending heavily on the selection of extraction variables. In

[41], a collaboration between Mayo Clinic and IBM, Coden et al looked for cancer grade, stage, size, date, and etc. from pathology reports and structured the results into templates. They achieved F1 scores ranging for 0.65 to 1.00 for various categories. Ashish et al [23] from the University of California Irvine (UCI) trained and tested on pathology reports from the UCI data warehouse and looked for entities such as TNM stage, capsule invasion, lymph invasion, chronic inflammation, and vascular invasion. They achieved F1 scores ranging from 0.78 to 1.00. In [80], Imler et al from the Indianapolis Veterans Affair extracted mentions of adenomas and their location, size, and number from colonoscopy reports, among other items. From Kaiser Permanente Southern California (KPSC), Danforth et al [45], working on radiology transcripts classified findings of lung nodules with a 96% and 86% sensitivity and specificity. Another study from from KPSC, [165], analyzed breast and prostate cancer pathology reports in a predetermined multi-stage pipeline which involved concept matches of clinical findings, diagnostic information, and other information such as tumor stage and Gleason score. Gao et al from Group Health [67] classified existence and location of mammography findings, using a combination of regular expressions and heuristic rules, with high accuracy.

Cancer case-finding systems are those that concentrate on detecting cancer occurrence or reoccurrence. This classification often takes place at the document level but may be subsumed to the patient level. Friedlin et al [65] used the Regenstrief EXtraction tool (REX) to look up pancreatic cancer and its synonyms as well as context, e.g. positive, negated, historical, and family history. They further compared pancreatic cancer case detection when using ICD-9 codes versus natural language processing extraction from Indiana School of Medicine data and found both cases where pancreatic cancer mentions occurred in clinical notes and not in ICD-9 codes and vice versa. In [34], Carrell et al from the Group Health Research Institute in Seattle approached the problem similarly to identify mentions of breast cancer. Further classification of reoccurrence depended upon their manipulating previously captured cancer mentions with a set of heuristics. Wilson et al [183] from the University of Pittsburgh identified and structured cases of ancillary cancer on patients and patient family

members from multiple report types. Their strategy involved using rule-based identification of "hotspots" (relevant text areas of interest), e.g. looking for words with "-oma" suffix as in "carcinoma", and creating context-dependent dynamic windows around the "hotspots" to look for context.

Machine learning and hybrid systems also exist. Similar to the rule-based systems, these can be divided to works that extract cancer characteristics and those that focus on case-finding. As before, scores vary tremendously based on the extraction variable and the types of documents. Kavuluru et al [88] worked on pathology reports from the Kentucky Cancer Registry to find primary site of neoplasm with macro-F1 0.72 and micro-F1 0.90 performance testing several machine learning algorithms on n-gram and medical concept features. In [87], Jouhet et al, working with French pathology reports from the Poitou-Charentes region, focused on extracting generic anatomic site, generic histology, and ICD-O3 (third edition of the International Classification of Diseases for Oncology) named entities with individual parameter performances ranging from 0.66 to 0.999 F1 for various topographic classes, e.g. prostate, skin, colon, etc. They tested between NB and SVM on bag-of-words term-weighted features. Hassanpour and Langlotz [74] experimented with named entity recognition in CT radiology reports, comparing dictionary methods, conditional random fields (CRFs), and maxiumum entropy markov models (MEMMs) with a performance of 0.85 F1. Ou et al [127] processed pathology reports into sentences, then used CRFs to extract cancer-related entities. Afterwards, entities were structured into templates using rules. Specific information of interests included diagnosis, metastasis, site, size, and specimen type, with end-to-end performance of 0.85 F1. University of Pittsburgh's Cancer Tissue Information Extraction System (caTIES) uses MMTx (MetaMap) and a series of rule-based steps to identify tumor grade, stage, other concepts, and negation [42]. Afterwards entities are arranged into a hierarchy using a simple nearest neighbor algorithm.

Cancer case finding using machine-learning systems may be evaluated at the document or patient level as well. In [46] and [145] D'Avolio et al and Sada et al, use the ARC framework to test CRF and MEMM classifiers in identifying positive cases of colorectal, prostate,

lung, and HCC cancer in pathology and radiology reports. Xu et al [187] detected colorectal CRC cancer (CRC) cases from multiple documents. They first identified CRC concepts using MedLEE, with the addition of hand-crafted keyword lookups, and performed assertion classification on those terms before testing a patient level machine-learning or rule-based final step.

Two systems were especially relevant to our project. One was a 2013 rule-based system from National Taiwan University [135] that extracted elements of liver cancer diagnosis, tumor characteristics, staging (BCLC and Child-Pugh), comorbidities, and treatments using regular expression rules to capture concepts and relations. In their study, Ping et al used a diverse set of report types include radiology, ultrasound, discharge, pathology, operation, and admission reports from 152 liver cancer patients receiving ultrasound (US) radiofrequency (RF) abalation. Relevant extraction performances are shown in Table 2.1.

The other study was a 2014 hybrid system from Wang et al, Fudan University [181]. Hepatic carcinoma information was extracted from 115 operation notes. The group first identified sentences of interest with keyword look-ups with 0.95 F1. In a second step, they tested using a rule-based versus a CRF algorithm to structure information. Their CRF setup yielded the best performances on the test set with 69.6% precision, 58.3% recall, and 63.5% F1. Relevant performances are summarized in Table 2.2. While extraction variables are similar, the difference in language (Chinese vs. English) and the difference in note types (operation notes vs. clinical and radiology notes) makes this and our work not directly comparable.

Category	**Example Expressions**	**Precision**	**Recall**	**F1**
Comorbidity diagnosis	Child-Pugh: Child's B Comorbidity: liver cirrhosis Diagnostic status: suspicious	0.99	0.98	0.99
Staging(BCLC)	BCLC stage A1, BCLC A1	0.99	0.98	0.98
Tumor	Tumor object: tumor Size: 1cm Quantifier: two Location: liver, breast	0.96	0.96	0.96

Table 2.1: Relevant Ping 2013 et al [135] results. Categories reflect several parameters pooled together.

Category	**Precision**	**Recall**	**F1**
Ascites (Q5)	0.75	0.82	0.78
Lymph node enlargement (Q6)	0.92	0.92	0.92
Tumor location (Q8)	0.37	0.50	0.42
Hepatic cirrhosis (Q9)	0.43	0.79	0.56
Tumor size (Q11)	0.33	0.45	0.38
Portal vein blocking (Portal vein thrombosis) (Q12)	0.23	0.38	0.29

Table 2.2: Relevant Wang 2014 et al [181] results. (Paranthetical items, e.g. Q5, are the corresponding markers in the paper's chart)

2.3 Relation to other clinical NLP tasks

Because every work is part of a rich mosaic of exciting research, in this section we give some details of related sub-domains for the benefit of the reader.

2.3.1 Phenotype extraction from clinical documents using NLP

We define a phenotype as a label characterized by a pre-determined set of observable values. Thus, liver cancer stage classification may be considered a special case of phenotype extraction; and part of a larger set of medically motivated problems that require classifications of a disease state using multiple data sources, often with complex definitional criteria.

In recognition of the complexity of defining phenotypes and the intrinsic value in having reference standards to them for future studies, the Phenotype KnowledgeBase (PheKB) was created within the eMERGE Network[4] starting from 2012 to support sharing and building phenotype algorithms [91]. Currently, there are 30 phenotype algorithms and more than 60 in development, of which NLP is a leading extraction factor behind ICD codes and medication data.

Previous work in phenotype identification cases have shown successful implementations of both rule-based and machine learning algorithms to aggregate multiple or disparate values in several documents per patient [154][98][130][95][96]. In the cancer domain, patient level predictions are often for cancer case identification. Ping et al performed a patient level classification of HCC occurence and reoccurence using regular expression and crafted rules [135]. Xu et al identified cases of colorectal cancer (CRC) on the entity level and then classified document level and patient level CRC status testing rule-based and machine learning methods [187].

Some common themes among phenotype problems, are: (1) multiple sources of data, (2) multiple levels of aggregation (e.g. document level, episode level, patient level), and (3) complementary data sources

[4]The mission of the eMERGE Network, funded by the National Human Genome Research Institute, is to combine DNA repositories with electronic health records (EHRs) for large-scale research [69].

2.3.2 Identification of medical topics and concepts in social media

While clinicians listen to patients and transcribe their observations to relevant medical conditions in clinical reports, patients' own personal accounts of symptoms can be another source of information as technologies evolve. Indeed, there is a growing interest in text-mining health-related topics in social media such as in community message boards and in twitter-like services. While work in this genre generally have different NLP challenges, such as internet abbreviations and emojis, capture of some stage parameters may be very similar to our tasks. For example, a forum discussing an individual's bouts of confusion or bloating may have very similarly language cues as *hepatic encephalopathy* or *ascites*. Here we identify several related examples in social media.

Most relevantly, Jha and Elhadad [83] predicted patient breast cancer stage of online forum contributors using a network model, with both text and metadata features. Brennan [32] worked on automatically detecting UMLS medical terms in emails. MacLean and Heer [103] investigated the creation of a labeled medical term set from online health forums logs using crowdsourcing and compared several ML systems for the extraction task. In twitter-like social media, there are works identifying ailment categories such as allergies, depression, aches/pains, cancer, obesity and etc [129] [48].

2.3.3 Identification of medical concepts in biomedical literature

Less patient-centered and more formal than either personal social media text or clinical text, biomedical literature may convey clinical concepts in the scope of population-based studies to contribute to growing medical knowledge. Particularly, the same sign and symptoms in clinical text may also be described in academic papers regarding specific medical conditions. There is a large body of work in this domain with various targets and techniques. We reference a few works here for further reading, such as identification of signs and symptoms in MEDLINE/PubMed abstracts [106], identification of disease mentions from PubMed

abstracts [55], and entity and relation extraction from MEDLINE articles [15] [58][199].

2.4 Summary

This chapter gave a brief overview about the challenges of NLP in the clinical domain, describe existing systems, and challenge sets. Furthermore, we discussed related works to our staging tasks, including prior works in cancer stage prediction and in cancer characteristics extraction. In the final subsection, we touched upon related sub-domains of clinical phenotype extraction and biomedically motivated extraction tasks in social media and in biomedical literature.

Chapter 3

IN-DEPTH ANNOTATION FOR PATIENT LIVER CANCER STAGING

This chapter describes the creation of our HCC liver cancer patient dataset, which will be used to train and build our staging system. Because we needed a way to evaluate our performance and have training data to build our algorithms, we performed detailed annotation. In the following sections, we provide our process in dataset collection, annotation guideline creation, and annotation. Afterwards, we show our annotator agreement, as well as provide quantitative and qualitative analysis to characterize our dataset.

3.1 Annotations

Every staging scheme emphasizes their own observable evaluation metrics related to cancer spread or liver function. Therefore, the choice of staging scheme affects what parameters should be summarized. In this study, we chose to annotate for BCLC and CLIP because of their demonstrated life expectancy correlations in liver cancers, as well as the AJCC (American Joint Committee on Cancer) recommended staging for its method commonality across cancers [79][70][155]. By using several popular and demonstratively predictive staging systems, we do not exclusively commit to one staging system and its particular bias. Each stage is determined by several stage parameters such as *tumor size*, *Child-Pugh* stages, or *ascites* [170]. Table 3.1 shows the 3 stages and the 11 pooled stage parameters that are component information for calculating the stages. Table 3.2 gives the stage parameters, their descriptions, and examples found in clinical notes. While *tumor size*, *tumor number*, *macrovascular invasion* and *metastasis*, are typically collected across all cancers, the other parameters, with the exception of *ECOG*, specifically allude to liver function.

Type	**Label**	**Values**
Stage	AJCC	I, II, IIIa, IIIb, IIIc, IVa, IVb
	BCLC	A1,A2,A3,A4, B, C, D
	CLIP	0, 1, 2, 3, 4, 5, 6
Stage Parameters	Ascites	None
		Mild-Suppressed on medications
		Moderate-Servere/Refractory
	Child-Pugh Class	A, B, C
	ECOG Performance Status	0, 1, 2, ≥ 3
	Extrahepatic Invasion	No, Yes
	Hepatic Encephalopathy	None
		Mild/Grade 1-2/Suppressed
		Severe/Grade 3-4/Refractory
	Macrovascular Invasion	No
		Yes - minor branch
		Yes - major branch
	Metastasis	No
		Yes - regional lymph nodes
		Yes - distal
	Portal Hypertension	No, Yes
	Tumor Morphology	Uninodular and extension <50% of liver
		Multinodular and extension <50% of liver
		Massive or extension $\geq$ 50% of liver
	Tumor Number	Single, 2-3, >3
	Tumor Size	<3 cm, 3-5 cm, >5cm

Table 3.1: Stage and stage parameters. (ECOG=Eastern Cooperative Oncology Group).

Label	Description	Example Text
Ascites	Accumulation of fluid in the peritoneal cavity.	"He denies increasing abdominal girth" "He has no problems with edema or ascites" "No free fluid in the abdomen" "Small volume ascites"
Child-Pugh class	Score that summarizes liver function	"Child-Pugh: A" "She is currently a Child's B score 7" "He is Child class A" "CTP-A6 cirrhosis"
ECOG performance status	Measure of general well-being of a patient, (0-5).	"ECOG performance 0." "He works out at a gym" "Notable for chronic fatigue." "She has been doing relatively well and has been undertaking her daily activities without any problems"
Extrahepatic invasion	Spread of cancer outside of liver	"Extrahepatic metastatic disease: None" "Lymph nodes: Scattered subcentimeter lymph nodes not pathologic by size criteria" "No evidence of extrahepatic extension"
Hepatic encephalopathy	Confusion or altered consciousness due to liver failure	"He has no significant ascites or encephalopathy cirrhosis has been complicated by hepatic encephalopathy" "Lactulose" "The patient denies any confusion, forgetfulness, or other symptoms of hepatic encephalopathy"
Macrovascular invasion	Spread of cancer to nearby blood vessels	"No evidence of portal vein thrombosis" "No obvious invasion of vessels is noted." "Portal veins are patent." "Vascular invasion: None"
Metastasis	Spread of cancer to outside-liver lymph nodes	"Lymph nodes suspicious for metastatic involvement: None" "No abnormal lymph nodes" "No evidence of extrahepatic extension or metastasis" "No other findings suggestive of extrahepatic disease"
Portal hypertension	Elevation of hepatic venous pressure gradient to $>$ 5mm Hg	"No evidence of portal HTN" "Patient had an EGD which showed small varices" "Recanalization of the umbilical vein, perigastric and peri-splenic varices compatible with portal hypertension physiology"
Tumor morphology	Size of tumor relative to the liver	"1 lesion measuring 2.1 x 1.7 cm in segment 6" "Lobulated hypovascular lesion in segment VIII." "Small segment 7 hepatic mass which enhances and demonstrates some degree of washout"
Tumor number	Number of liver tumors	"Multiple other indeterminate foci of arterial enhancement in the left and right lobe suspicious for HCC" "Two new liver lesions noted on the current examination with hypervascularity and washout suggesting hepatomas."
Tumor size	Radius size of liver tumor	"there is a segment 4A arterial enhancing lesion which shows homogeneous washout on the delayed phase measuring 1.7 x 1.5 cm" "Well defined mass lesion measuring 6.3 x 7.1 x 6.1 cm, epicentered in segment 4a suggestive of hepatocellular carcinoma"

Table 3.2: Text annotation examples

3.2 Annotation process

Data and data preparation

A cohort was drawn from patients visiting the University of Washington Medical Center (UWMC) primary liver cancer clinic from 1/2011-12/2013. Included data for each patient comprised of all clinical notes from the day of visit to the clinic, all laboratory results 30 days prior or following to the day of visit, and the CT or MRI of the abdomen or abdomen/pelvis or chest/abdomen/pelvis with contrast 3 months prior to or 1 month following the day of visit.

Patient records were manually reviewed by our clinical expert to exclude patients with more than one visit day in the time window and who had an obviously irrelevant diagnoses. Mislabeled reports were renamed to their correct report type. Irrelevant report types, e.g. Pre Anestheia notes, were removed from the annotation set. The list of inclusion and exclusion note types are shown in Table 3.3. The remaining report types are shown in Table 3.4. For our study, we focused on the subset of patients that have at least one clinical report, at least one radiology report, and the full-set of labs needed to calculate *Child-Pugh* and CLIP scores. Table 3.11 and 3.10 shows the required parameters for *Child-Pugh* and CLIP staging. The resulting dataset includes 236 patients and their associated 422 clinical notes and 309 radiology reports, which translates to an average of 1.8 clinical notes and 1.3 radiology notes per patient. Tables 3.5 and 3.6 gives the distribution of clinical notes and radiology notes respectively.

Guideline creation

Guidelines for liver cancer stage and stage parameter annotations were developed primarily by an interventionist radiologist with input from another interventional radiologist and a group of NLP scientists. Stage parameter values were discretization, according to the smallest common factor for all 3 staging methods. For example, while *macrovascular invasion* is important for all three staging schemes, AJCC distinguishes between major and

Included	Excluded
Surgery - Outpt Record	ED Note
Outpt Progress Note	Patient Instructions–Education Outpt
Interventional Radiology - Inpt Record	Pre Anesthesia
Hepatology–Hepatitis - Outpt Record	Nursing RecordNote
Admit Note	ED Clinical Summary
Madison - Outpt Record	ED Patient Summary
SCCA Outpt Record	Wound Care – Treatment(s)
Consultation	Procedure Note
History & Physical	Sleep Study Report
	Surgery Admit–Initial Consult Note
	Telephone Note
	Panel Summary

Table 3.3: Report types that were included or excluded

minor branch invasion, therefore the final discretization has 3 values: {*Yes-major_branch*, *Yes-minor_branch*, *No*} instead of two:{*Yes*, *No*}. Specifications of from which sections annotators were to find stage parameters in a report were formalized into annotation rules shown fully in Appendix A. Figure 3.1 provides a small excerpt of the text annotation guidelines for three stage parameters.

The original stage value assignments for AJCC, BCLC, and CLIP are defined by Table 3.7, 3.8, and 3.10, respectively. Underspecified or ambiguous cases from the original guidelines, were arbitrated by the domain expert annotators, with respect to the defined stage parameters. Decisions were explicitly enumerated in lookup tables of a spreadsheet. The final lookup tables, after additional programmatic augmentation, to find and disambiguate logical holes, are in Appendix B.

Report Type	**# Patients**
Hepatology–Hepatitis - Outpt Record	145
Surgery - Outpt Record	123
Consultation - Outpt Record	55
Interventional Radiology - Outpt Record	37
SCCA - Outpt Record	21
Clinic Note	9
Radiation Oncology - Outpt Record	4
Interventional Radiology - Inpt Record	4
Admit Note	4
Surgery Admit–Initial Consult Note	3
Surgery - Inpt Record	3
Outpt Progress Note - General	3
Initial Clinic–New Consult	3
Panel Summary	2
Cancer Treatment - Outpt Record	2
Transplant - Outpt Record	1
Hematology–Oncology - Inpt Record	1
GI - Outpt Record	1
Adult Medicine - Outpt Record	1

Table 3.4: Clinical note types

# Reports	**# Patients**
1	94
2	107
3	28
≥ 4	7

Table 3.5: Clinical notes per patient

# Reports	**# Patients**
1	170
2	59
≥ 3	7

Table 3.6: Radiology notes per patient

ECOG Performance Status

If clinic notes do not give specific scores, mark the text that gives clues to the score. If clinic note gives descriptive text evidence AND specific scores, mark both. If two notes conflict on score, annotate all scores and give appropriate respective scores, but when assigning final score during staging by consensus, this will be deemed unscorable. Values: 0,1,2 $\geq$ 3

Extrahepatic invasion

Defined as direct invasion of an adjacent organ other than gallbladder or perforation of visceral peritoneum. Values: No, Yes

Macrovascular invasion

Mark in Impression section, if data is available there. Otherwise, mark in Findings section. Major branch macrovascular invasion is defined as anything larger than left or right PV or left, middle, or right HV.

Values: No, Yes-minor branch, Yes-major branch

Figure 3.1: Annotation guideline examples

AJCC Stage	Description
Stage I	There is a single tumor (any size) that has not grown into any blood vessels. The cancer has not spread to nearby lymph nodes or distant sites.
Stage II	Either there is a single tumor (any size) that has grown into blood vessels, OR there are several tumors, and all are 5 cm (2 inches) or less across. The cancer has not spread to nearby lymph nodes or distant sites.
Stage IIIA	There is more than one tumor, and at least one is larger than 5 cm (2 inches) across. The cancer has not spread to nearby lymph nodes or distant sites.
Stage IIIB	At least one tumor is growing into a branch of a major vein of the liver (portal vein or hepatic vein). The cancer has not spread to nearby lymph nodes or distant sites.
Stage IIIC	A tumor is growing into a nearby organ (other than the gallbladder), OR a tumor has grown into the outer covering of the liver. The cancer has not spread to nearby lymph nodes or distant sites.
Stage IVA	Tumors in the liver can be any size or number and they may have grown into blood vessels or nearby organs. The cancer has spread to nearby lymph nodes. The cancer has not spread to distant sites.
Stage IVB	The cancer has spread to other parts of the body. (Tumors can be any size or number, and nearby lymph nodes may or may not be involved.)

Table 3.7: Guidelines for AJCC staging [156]

BCLC Stage	PST	Tumor Stage	Okuda Stage	Liver Function
A1	0	Single	I	No PH, normal bilirubin
A2	0	Single	I	PH, normal bilirubin
A3	0	Single	I	PH, abnormal bilirubin
A4	0	3 tumors $<$ 3cm	I-II	CP A-B
B	0	Large multinodular	I-II	CP A-B
C	1-2	Vascular invasion or extrahepatic spread	I-II	CP A-B
D	3-4	Any	III	CP C

Table 3.8: Guidelines for BCLC staging [170]. PST=performance status (ECOG), PH=portal hypertension, CP=Child-Pugh. Okuda stages are defined in Table 3.9.

Okuda factors
Tumor size >50% of liver
Ascites
Albumin < 3 g/dL
Bilirubin > 3 mg/dL

Table 3.9: Okuda stage definition. Stage I: no factors present. Stage II: 1-2 factors. Stage III: 3-4 factors.

Variable	Points		
	0	1	2
Child-Pugh	A	B	C
Tumor Morphology	Uninodular and extension $\leq 50\%$	Multinodular and extension $\leq 50\%$	Massive or extension $> 50\%$
AFP (ng/dL)	< 400	≥ 400	
Portal vein thrombosis	No	Yes	

Table 3.10: Guidelines for CLIP staging [170]. CLIP stage is a score assigned by adding up all the points from each variable.

Variable	Points		
	1	2	3
Albumin (g/dL)	> 3.5	2.8-3.5	< 2.8
Ascites	None	Mild/Moderate	Severe
Bilirubin (Total) (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 3.11: Child-Pugh parameters [125]. 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.

Section ontology creation

Clinical documents are typically organized by section headings, however every institution, department, and specialities, have their own common section formats (Figure 3.2). Because sections are a crucial element in clinical text processing, we created a section ontology for clinical notes from the UWMC primary liver clinic. An interventional radiology domain expert reviewed reports and created the ontology, located in Appendix C. Using the ontology, a biomedical informatics student manually labeled 100 clinical note documents randomly drawn from the data set. Radiology reports, also 100 randomly drawn documents, were also labeled using a previously published radiology report ontology [174]. Using this labeled data, an in-house section identifier [174] for both types were trained. The estimated performance, using 5-fold cross-validation, is 0.90 and 0.97 F1, for clinical and radiology notes, respectively. For future references to identified sections, which we treat as solved, we used this tool.

HISTORY OF PRESENT ILLNESS

..

Over the last 3 weeks he has developed abdominal distention consistent with ascites.

...

PHYSICAL EXAMINATION

...

ABDOMEN: soft.

Moderate to large distention but not tight.

Positive large ascites on exam.

...

ASSESSMENT AND PLAN:

...

He is showing evidence of increasing ascites most likely related to some mild hepatic impairment in conjunction with extensive vascular invasion.

Figure 3.2: Example of a report with multiple sections.

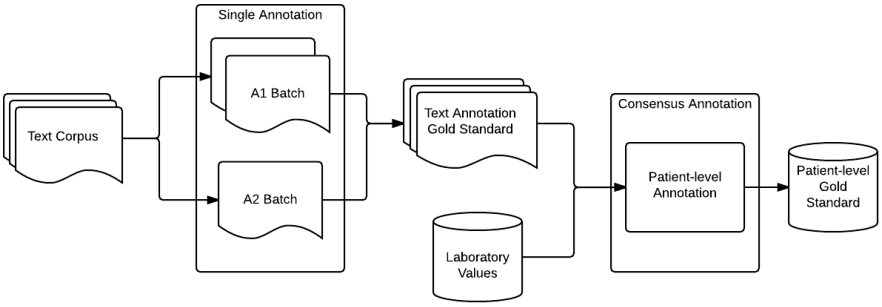


Figure 3.3: Annotation workflow. First, patient charts were divided among two annotators to annotate for text level evidence. Afterwards, patient information were consolidated and patient level annotations were annotated in consensus with access to laboratory values. The end-products are gold standards for: (1) text annotations and (2) patient level annotations.

Annotation workflow and software

Annotation occurred in two phases, as illustrated in Figure 3.3. In the first phase, relevant parts of reports were identified and associated with an annotation label and value, as shown in Table 3.1 and 3.2. During this phase our annotators marked text-annotations using brat [163], a web-based graphical annotation tool, and assigned them a label, e.g. *ECOG*, and a value, e.g. *0*, as shown in Figure 3.4 and 3.5. Irrelevant patients, e.g. patients with irrelevant diagnosis, and files, e.g. addenda, abbreviated notes, and post-treatment radiology notes, were flagged for exclusion. As annotation is relatively time-consuming for clinicians, we adopted a sparse-annotation approach. Thus not all instances of some stage parameter evidence in a document were marked. The exact algorithm on which sections to mark are discussed in our annotation guidelines in Appendix A.

During the second phase, the 3 overall stages were annotated with access to patient records and laboratory values from structured data. In addition to this, the 11 text annotation liver cancer parameters had corresponding patient level annotations that were marked

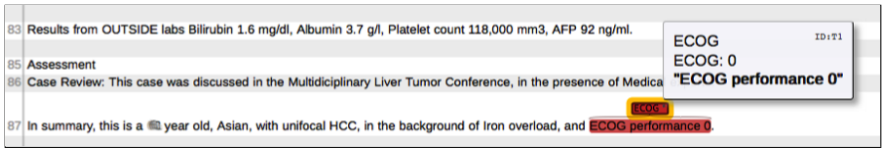


Figure 3.4: Brat text annotation. The annotator first selects the text span, then assigns a label type (ECOG) and a particular value (0).

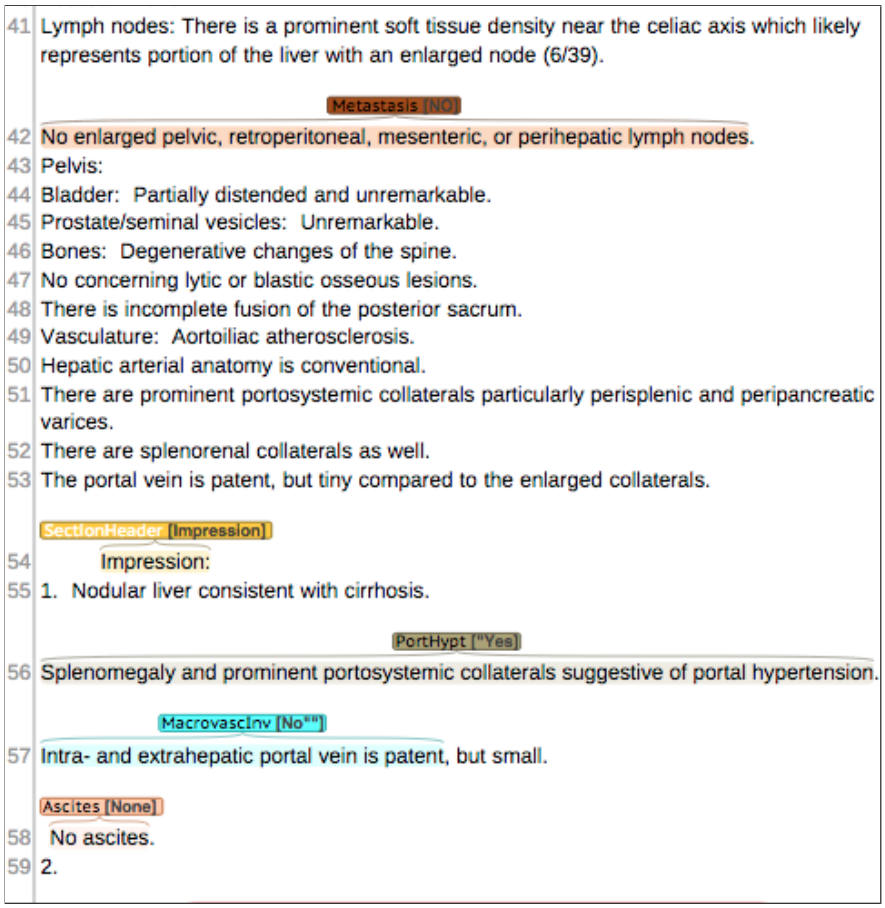


Figure 3.5: Brat text annotations for several parameters and their values. Automated section annotation, e.g. SectionHeader[Impresion], is also shown.

at this step. Part of the reason for this is to resolve missing and conflicting values from the text annotations. This phase was done as a consensus of the two domain specialists. For this task, the annotators used a specially built in-house python Tkinter-built [78] interface shown in Figure 3.6. During this phase, the annotators had access to the full marked reports as well as a summarized version of their annotations displayed in the interface. Pertinent laboratory values were displayed via the interface for *Child-Pugh* or CLIP stage calculations.

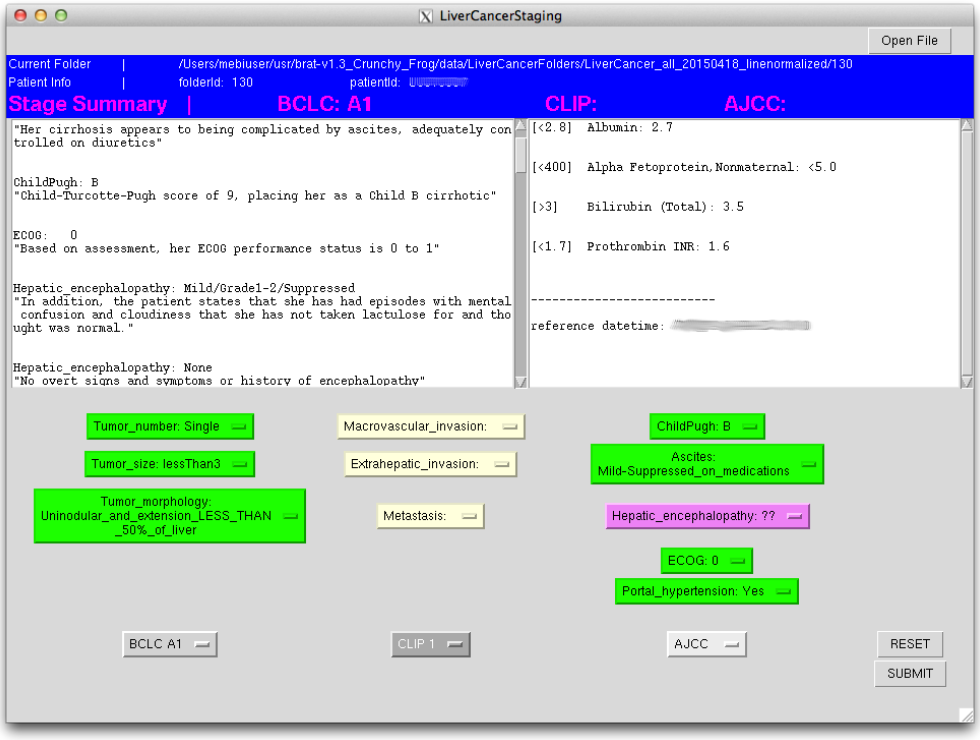


Figure 3.6: Patient level annotations. The interface loads the values for each patient annotated during the first phase of annotation. Text evidence annotations are displayed on the left panel. The button menus in the lower half of the interface facilitates annotation input. The 11 stage parameters from text evidence are automatically loaded on the button menus for patient level annotations. Parameters with a single value from text annotations are shown in green. Those parameters that have no value are left blank and colored in light yellow, leaving annotators to choose the most appropriate value. Parameters with conflicting entities are marked with value "??" and highlighted in purple for the annotators to resolve. Laboratory values are available on the right panel to calculate *Child-Pugh* if necessary as well as the CLIP score. AJCC, BCLC, and CLIP are assigned by selecting from the bottom row of button menus.

Interrater agreement

An initial set of 20 patients, with 71 documents total, were double-annotated for text level evidence and their associated values by two interventional radiologist attending physicians. We then calculated agreement levels and refined annotation guidelines to maximize agreement and minimize redundant annotation. After the inter-annotator meeting, the two radiologist re-annotated their set and agreement was re-scored.

Agreement was measured using F1-score (F1)[77] for stage parameter labels and values at a patient level, e.g. *Patient 0: Ascites None*, a document level, e.g. *Patient 0 Document 1: ascites none*, and at a partial text span level, e.g. *Patient 0: document 1: ascites none "no abdominal distention"*. Patient level and document level values were automatically generated from the text evidence according to which document or patient it belonged to. Therefore, a single patient or document may have multiple values for the same stage parameter at once, e.g. *"ascites: none"* and *"ascites: mild"*. A partial text span true positive requires some overlap in highlighted text span, as well as matching labels, e.g. *ascites*, and values, e.g. *none*. The formulas for defining F1 are shown in the following equations, where *TP=True Positive*, *FP=False Positive* and *FN=False Negative*:

$$P = \frac{TP}{TP + FP} \qquad (3.1) \qquad R = \frac{TP}{TP + FN} \qquad (3.2) \qquad F1 = \frac{2PR}{P + R} \qquad (3.3)$$

After refining guidelines, the rest of the dataset was divided among the two annotators to annotate separately, in batches of 31 patients. patient level stages and parameter values were annotated by a consensus of the two annotators and therefore could not be evaluated for interrater agreement.

Results

Table 3.12, 3.13, and 3.14 shows text evidence agreement levels generated at the patient level, document level, and partial text span level after the first round of annotations. The

degradation from patient level to partial text span levels was expected given the precision required for each level. The patient level annotation is ultimately the most clinically relevant, however measuring the document level and partial text-span level performances gives a sense of how consistent the two annotators are with each other. The performance gap between patient level and document level alludes to how often annotators find the same patient information in different files. Similarly, the difference from document level and text span levels demonstrates how often the same values in the same document come from different areas of the text.

Qualitative analysis and the low agreement across all levels showed that annotators were locating evidence in different parts of the same report, did not have consistent definitions for each parameter category, and did not consistently exclude files or patients. During the inter-annotator meeting, the annotators discussed: (1) definitions of parameters and what should be marked, (2) which files and which sections to start sparse annotation, and (3) patient exclusion criteria (one annotator excluded 3 patients whereas the other annotator did not exclude any). The annotators were then sent to re-annotate the same batch.

Table 3.15, 3.16, and 3.17 shows agreement levels after re-annotating. One annotator (A2) marked 3 patients for exclusion, whereas the other annotator (A1) marked only 2, though he did not mark other entities for that patient suggesting this was an annotation error. The higher level of agreement at the partial text-span level signaled that annotators were looking at similar areas for certain values. The four lowest-performing staging parameters at the patient level were *ascites*, *ECOG*, *extrahepatic invasion*, and *hepatic encephalopathy. Ascites*, *ECOG*, and *hepatic encephalopathy*, were difficult because they appeared in different places in a clinical note and were often repeated and in different expression formats. Additionally, one annotator marked *ascites* drugs while the other did not. Meanwhile, *extrahepatic invasion* discrepancies were due to one annotator identifying more information than the other. The overall higher agreement at all levels, 0.910 vs 0.764 for patient level, 0.854 vs. 0.599 at document level, and 0.729 vs. 0.447 at the partial text span level, showed better consistency between the annotators after the first inter-annotator meeting.

Label	TP	FP	FN	P	R	F1
Ascites	11	5	1	0.688	0.917	0.786
ChildPugh	5	2	0	0.714	1.000	0.833
ECOG	11	4	6	0.733	0.647	0.688
Extrahepatic_invasion	1	7	0	0.125	1.000	0.222
Hepatic_encephalopathy	8	2	3	0.800	0.727	0.762
Macrovascular_invasion	7	8	0	0.467	1.000	0.636
Metastasis	5	4	2	0.556	0.714	0.625
Portal_hypertension	11	0	3	1.000	0.786	0.880
Tumor_morphology	2	2	5	0.500	0.286	0.363
Tumor_number	17	2	0	0.895	1.000	0.944
Tumor_size	16	2	0	0.889	1.000	0.941
ALL	94	38	20	0.712	0.825	0.764

Table 3.12: Exact match of label-value per patient (First Round of Annotation)

Label	TP	FP	FN	P	R	F1
Ascites	16	9	6	0.640	0.727	0.681
ChildPugh	5	4	0	0.556	1.000	0.714
ECOG	17	4	8	0.810	0.680	0.739
Extrahepatic_invasion	1	8	0	0.111	1.000	0.200
Hepatic_encephalopathy	9	3	8	0.750	0.529	0.621
Macrovascular_invasion	7	16	0	0.304	1.000	0.467
Metastasis	5	5	2	0.500	0.714	0.588
Portal_hypertension	20	3	4	0.870	0.833	0.851
Tumor_morphology	1	5	6	0.167	0.143	0.154
Tumor_number	12	22	11	0.353	0.522	0.421
Tumor_size	20	26	1	0.435	0.953	0.597
ALL	113	105	46	0.518	0.711	0.599

Table 3.13: Exact match of label-value per document (First Round of Annotation)

Label	TP	FP	FN	P	R	F1
Ascites	10	17	17	0.370	0.370	0.370
ChildPugh	7	6	1	0.538	0.875	0.667
ECOG	20	11	18	0.645	0.526	0.580
Extrahepatic_invasion	1	9	0	0.100	1.000	0.182
Hepatic_encephalopathy	9	5	12	0.643	0.429	0.514
Macrovascular_invasion	5	25	2	0.167	0.714	0.270
Metastasis	6	6	4	0.500	0.600	0.545
Portal_hypertension	28	14	14	0.667	0.667	0.667
Tumor_morphology	0	6	7	0.000	0.000	0.000
Tumor_number	12	37	23	0.245	0.343	0.286
Tumor_size	19	53	2	0.264	0.905	0.409
ALL	117	189	100	0.382	0.539	0.447

Table 3.14: Partial match of label-value per text-span (First Round of Annotation)

Label	TP	FP	FN	P	R	F1
Ascites	9	4	2	0.692	0.818	0.750
ChildPugh	6	0	0	1.000	1.000	1.000
ECOG	14	1	3	0.933	0.824	0.875
Extrahepatic_invasion	5	4	0	0.556	1.000	0.714
Hepatic_encephalopathy	8	2	2	0.800	0.800	0.800
Macrovascular_invasion	13	2	0	0.867	1.000	0.929
Metastasis	9	1	0	0.900	1.000	0.947
Portal_hypertension	11	3	0	0.786	1.000	0.880
Tumor_morphology	17	1	0	0.944	1.000	0.971
Tumor_number	17	0	0	1.000	1.000	1.000
Tumor_size	17	0	0	1.000	1.000	1.000
ALL	126	18	7	0.875	0.947	0.910

Table 3.15: Exact match of label-value per patient (Second Round of Annotation)

Label	TP	FP	FN	P	R	F1
Ascites	11	8	7	0.579	0.611	0.595
ChildPugh	7	0	0	1.000	1.000	1.000
ECOG	20	4	4	0.833	0.833	0.833
Extrahepatic_invasion	6	4	0	0.600	1.000	0.750
Hepatic_encephalopathy	12	2	4	0.857	0.750	0.800
Macrovascular_invasion	16	5	0	0.762	1.000	0.865
Metastasis	11	1	0	0.917	1.000	0.957
Portal_hypertension	13	5	2	0.722	0.867	0.788
Tumor_morphology	21	2	2	0.913	0.913	0.913
Tumor_number	23	0	1	1.000	0.958	0.979
Tumor_size	21	2	2	0.913	0.913	0.913
ALL	161	33	22	0.830	0.880	0.854

Table 3.16: Exact match of label-value per document (Second Round of Annotation)

Label	TP	FP	FN	P	R	F1
Ascites	10	9	12	0.526	0.455	0.488
ChildPugh	7	0	0	1.000	1.000	1.000
ECOG	23	6	9	0.793	0.719	0.754
Extrahepatic_invasion	6	4	0	0.600	1.000	0.750
Hepatic_encephalopathy	12	3	5	0.800	0.706	0.750
Macrovascular_invasion	16	6	0	0.727	1.000	0.842
Metastasis	10	2	1	0.833	0.909	0.870
Portal_hypertension	11	7	5	0.611	0.688	0.647
Tumor_morphology	15	8	8	0.652	0.652	0.652
Tumor_number	17	6	7	0.739	0.708	0.723
Tumor_size	18	5	5	0.783	0.783	0.783
ALL	145	56	52	0.721	0.736	0.729

Table 3.17: Partial match of label-value per text-span (Second Round of Annotation)

Patient annotations

Table 3.18 and 3.20 shows the breakdown of the patient level consensus annotations for the entire set. The final number of non-excluded patients was 200. Stage annotations are skewed towards earlier stages for two out of the three schemes. This bias may be because of the population being treated (sicker patients are less likely to be referred to the university hospital) and our data exclusion of return patients. Similarly, for the individual patient level stage parameters, there are large class imbalances towards the least severe values. One patient was mistakenly not marked for exclusion, explaining the [EMPTY] value of 1 for many of the parameters. However there was one case of a missing *tumor morphology* and *extrahepatic invasion* in the set. The *ECOG* parameter had the most conflicts, "??", with 16 out of 200 patients with multiple conflicting values.

Label	**Value**	**Frequency**	**Label**	**Value**	**Frequency**	**Label**	**Value**	**Frequency**
	I	108		A1	27		0	66
	II	48		A2	21		1	62
	IIIA	16		A3	13		2	41
AJCC	IIIB	14	BCLC	A4	17	CLIP	3	18
	IIIC	0		B	23		4	8
	IVA	6		C	70		5	4
	IVB	7		D	14		6	0

Table 3.18: Stage annotations

Label	Value	Frequency
Ascites	Mild-Suppressed on medications	34
	Moderate-Severe/Refractory	11
	None	154
	??	0
	[EMPTY]	1
ChildPugh	A	125
	B	62
	C	12
	??	0
	[EMPTY]	1
ECOG	0	121
	1	42
	2	14
	≥ 3	6
	??	16
	[EMPTY]	1
Extrahepatic_invasion	No	196
	Yes	2
	??	0
	[EMPTY]	2
Hepatic_encephalopathy	Mild/Grade1-2/Suppressed	26
	None	170
	Severe/Grade 3-4/Refractory	3
	??	0
	[EMPTY]	1
Macrovascular_invasion	No	172
	Yes–yes-major_branch	14
	Yes–yes-major_branch	13
	??	0
	[EMPTY]	1

Table 3.19: Patient level liver cancer characteristic annotations (part 1)

Label	Value	Frequency
Metastasis	No	186
	Yes–distal	7
	Yes–regional lymph nodes	6
	??	0
	[EMPTY]	1
Portal_hypertension	No	73
	Yes	126
	??	0
	[EMPTY]	1
Tumor_morphology	Massive or extension $\geq$50% of liver	21
	Multinodular and extension <50% of liver	64
	Uninodular and extension <50% of liver	113
	??	0
	[EMPTY]	2
Tumor_number	2-3	43
	> 3	30
	Single	125
	??	1
	[EMPTY]	1
Tumor_size	< 3	87
	3-5	59
	> 5	53
	??	0
	[EMPTY]	1

Table 3.20: Patient level liver cancer characteristic annotations (part 2)

Label	A1 Acc.	A1 Extra	A2 Acc.	A2 Extra
Ascites	11/13	2	11/11	0
ChildPugh	5/5	1	5/5	0
ECOG	11/13	2	13/15	2
Extrahepatic_invasion	9/9	0	5/5	0
Hepatic_encephalopathy	9/10	1	10/10	0
Macrovascular_invasion	15/15	0	13/13	0
Metastasis	10/10	0	9/9	0
Portal_hypertension	12/12	2	11/11	0
Tumor_morphology	16/17	1	16/17	1
Tumor_number	17/17	0	17/17	0
Tumor_size	17/17	0	17/17	0

Table 3.21: Comparison of text annotations with the patient level consensus annotations. An annotator is said to be accurate in this context if at least one of their text annotation label-values are the same as the patient level consensus annotations. A1=Annotator 1, A2=Annotator 2, Acc=Accuracy.

Table 3.21 shows the comparison between each annotator's text evidence and the patient level consensus annotations in accuracy for the inter-annotator 20 patients. An annotator is said to be accurate for a particular label if one of their text annotations contains the value that the patient level annotation is assigned. For example, if the patient level says the patient is has an *ECOG* value of *0*. Then if the annotator marks at least one *ECOG: 0* label-value text annotation, then they get that point. If the patient level annotation is indeterminable, e.g. *ECOG:??* then we do not evaluate for that label. Extra values are for cases when multiple values appear. Non-accurate values are counted as extra. In general our annotators were more than 90% accurate for most labels, with the exception of one or two labels at approximately 85% (11/13) accuracy.

Discussion

In our annotations, the relevant text evidence examples had various forms. Though an actual concept may not be mentioned, there may be other observable clues. To give an example, we observe very explicit evidence giving direct information, e.g. *"no ascites"*, *"ecog 0"*, *"no metastasis"*, *"no hepatice encelophathy"*. However, there were additionally cases that give signs and symptoms information or which required some reasoning. For example, *ascites* may be observed through *"abdominal distension"*. Likewise, *macrovascular invasion* has indicators such as *"No portal vein thrombosis"*. Some cases required deep domain knowledge as well as logical inference. For example, we use the example of *tumor number: 2-3 tumors, "she would be considered to fall within Milan criteria"*. In order to understand why this passage would lead to a clinician to infer 2-3 tumors, we need to know what Milan criteria means (that the patient has to have either one tumor smaller than 5 cm or up to three tumors smaller than 3cm). For that particular patient, it was unclear how many tumors there were but it was clear that there were more than one and none were larger than 3 cm, which led to the resolution that there had to be *tumor number: 2-3 tumors*. Another example is for *hepatic encephalopathy*. A patient was assigned to have mild or suppressed *hepatic encephalopathy* because of evidence *"lactulose"* and *"rifaximin"* without any other mentions

of *hepatic encephalopathy*. These are drugs which are used to treat problems related to the liver. Meanwhile, *tumor number* may require reading radiographic information, recognizing mentions of malignant tumors, and summing over the resulting list.

3.3 Challenges in annotation

3.3.1 Text annotations

During the process, we encountered several technical annotation challenges: (1) conflicting information, (2) ambiguity, (3) connected multi-sentence information, and (4) numerous potentials for negative evidence.

The first case is the clearest problem. One part of a document may conflict with another part of the same document or a different document in the patient's records. One example concerns a patient with a document containing the statement that said he or she was *Child-Pugh A*, but in another document said *Child-Pugh B*. Our solution was to add a patient level resolution for these parameters during the second stage of the annotation workflow.

During initial data testing, we found at least one case where a single statement alluded to more than one stage parameter value, creating an ambiguous statement, (2). The example for this was an example that came up *"He has well-compensated liver disease, with Child-Pugh score of 6 or 7 [...] This puts him at a class A/B"*. Although, seemingly a subset of (1) where the conflicting information occurs in the same passage, one can argue this differs in that respect as this statement does not establish that *Child-Pugh B* or *Child-Pugh C* definitively as in (1) but is a hedge of the two values. To maximize information, we decided to have annotators mark such passages as both *Child-Pugh B* and *Child-Pugh C*.

There were also some issues with annotation based on discrete passages of text. For example, in reports, many parts of information are clarified in later sentences, thus making normalization for each stage parameter value challenging to identify with a single spot because they are referring to the same real-world entity, (3). Figure 3.7 shows an excerpt from a report with *ascites* information. Initially, there is mention of some *ascites*, but only

in the Physical Examination section does it become clear that is somewhat severe. At the end of the document, in the Assessment and Plan section, the *ascites* is mentioned, again with less detail. Our annotation decision was to consider sentences independently and to use less severe values if the sentence is unclear or not detailed. Therefore, the first sentence in the example is considered *mild*, even though as a human we understand, in fact, that all instances in the document actually do refer to the same severe case (*moderate/severe*).

HISTORY OF PRESENT ILLNESS

...

Over the last 3 weeks he has developed abdominal distention consistent with ascites.

...

PHYSICAL EXAMINATION

...

ABDOMEN: soft.

Moderate to large distention but not tight.

Positive large ascites on exam.

...

ASSESSMENT AND PLAN:

...

He is showing evidence of increasing ascites most likely related to some mild hepatic impairment in conjunction with extensive vascular invasion.

Figure 3.7: Ascites information is referenced several times in a document with fluctuating details regarding severity.

Another issue was (3), the possibility of annotating negative evidence[1]. A problem such as *hepatic encephalopathy* is defined as confusion related to liver failure. Thus, whether or not it

[1]We define negative evidence as information that at least partially supports the opposite of a sign or symptoms of focus. For example, *"No suspicious osseous lesions"* is not direct evidence for either *metastasis* values *none*, *regional lymph nodes*, or *distal*. It does, however, count as negative evidence for *metastasis - distal*. The idea is if you collect all instances of negative evidence, e.g. all other body parts being not suspicious, then you may rule out *metastasis - distal*.

is appropriate or possible to mark all mentions of any cognitive issues is a decision required in designing annotation guidelines. For example *"PSYCHIATRIC: Alert and Oriented to Person, Place and Time, normal mood, normal thought content, affect normal."* clearly gives some indication for no *hepatic encephalopathy* in either *mild* or *severe* form, but it is not exactly direct. Moreover, unfortunately, marking every clue to cognitive impairments (or disimpairments) is much more time-consuming, with little relative gain.

To raise other examples, we can consider *metastasis - yes*, for which involves tumors in any other area except the liver – so highlighting all negative cases is challenging, e.g. *"No suspicious osseous lesions"*. Similarly, for *portal hypertension - yes*, which is defined by increased blood pressure in the portal venous system, symptoms include swollen veins within the esophagus, stomach, rectum, or umbilical area. Thus, negative examples would encompass many anatomic parts. Because of the inter-relatedness of various information, we did not annotate negative evidence, however we should bear in mind that this information adds the context of the patient record.

3.3.2 Patient stage annotations

Patient stage annotations were a challenge as each staging scheme required many parameters, and each with many values. Furthermore, the actual guidelines prescribed by the previous Table 3.7, 3.8, and 3.10 do not exactly align to our stage parameter label and values (because of the pooling and discretization of values, e.g. to identify significant ranges such as *tumor number - [2-3]*), requiring some cognitive processing.

3.4 Limitations

Limitations to the annotation process are that the annotation stages and values were developed by one primary expert and another assisting expert, allowing room to overlook certain concepts or other schemas. Most of the text evidence annotations were single-annotated, leaving possibility of annotation error. The set was also sparsely annotated, and therefore it may be missing particular semantic forms of the same information.

Patient level consensus annotations were done by two experts in consensus, which may allow for less error but opens up an opportunity for the annotators to influence each other. The arbitrations for unclear stage cases were done by two experts, but may not be agreed on by the entire community. Finally, the dataset is from the University of Washington Medical Center system so these guidelines may not be generalizable towards other departments or institutions.

3.5 Summary

In this chapter, we described the creation of an HCC cohort, the development of a set of annotation guidelines, as well as provide analysis on our dataset statistics and inter annotator agreements. In the next chapters, the development of extraction and classification systems for the various HCC stages and stage parameters will be based on this dataset.

Chapter 4

STAGING SYSTEM ARCHITECTURE

In this section, we give a brief description of our system design, including explanations of how constituent parts relate to each other. We further describe a simple document classification baseline for which sub-patient stage parameter classification tasks will be measured against.

4.1 Overall system architecture

The overall system architecture consists of several items, depicted in Figure 4.1. At the top are the three liver cancer stage classifications required per patient. Below the stages, are the 11 text parameters, and 4 laboratory values. Our system first annotates for the patient level 11 parameters before classifying overall patient liver cancer labels after. Each parameter classification had specific pipelines depending on their needs.

4.1.1 Data: Training and testing

Of the 200 non-excluded patients, 160 patients were separated as training and 40 patients were kept as a test set. For the stage parameter extractions, all experiments were done using only files from the 160 patients. For all sub-system modules (Chapters 5 and 6), except the tumor characteristics extraction (Chapter 7), training and development was performed on a 5-fold cross-validation of those 160 patients. The tumor characteristics extraction module (Chapter 7) used a smaller subset of the 160 training patients.The test set was only evaluated on after the entire system was created.

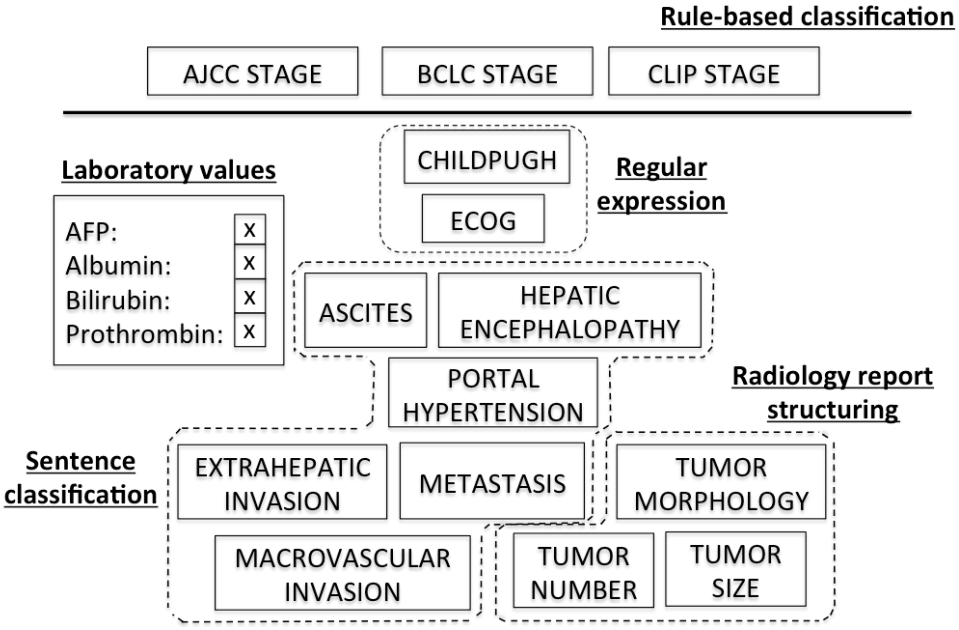


Figure 4.1: Overall system architecture

4.1.2 Stage parameter extraction

Stage parameters were determined by one of three overall strategies: (1) a two-step regular expression method, (2) a sentence classifier, and (3) a tumor characteristics extraction sub-system which classifies subdocument entities and relations and uses the information in addition to a number of rules. We compare each module to a document classification baseline described next, but also detailed in a previously published paper [194]. In the next chapters, at the completion of each module, we will refer back to these baselines.

Document classification baseline

As stage parameters text annotations were annotated sparsely, it was not feasible to do an evaluation at the subdocument level. Therefore, the various extraction modules (regular expression extraction, sentence classification, and tumor characteristics extraction) described in the next sections will be measured against a document classification baseline. As men-

tioned before, the document classification was measured in a 5-fold cross-validation, split by patient with their related documents. Specific stage parameters were restricted to certain document types following annotation guideline recommendations. Report restrictions are detailed in Table 4.1.

Label	**Document type**
Ascites	Clinical, radiology
Hepatic encephalopathy	Clinical
Extrahepatic invasion	Radiology
Macrovascular invasion	Radiology
Metastasis	Radiology

Table 4.1: Sentence classification document type restrictions

The baseline was obtained by taking the best score among several machine learning algorithms (Naive Bayes, Maximum entropy, SVM, binary decision tree, C4.5 decision tree) using simple 1-, 2-, 3- gram frequency features. The classification task was a binary decision between each label-value combination, e.g. *ascites-none*. If a label-value combination of a stage parameter appeared in the document, the document was considered positive for that class. The reason document level baseline was used was because, as per the annotation guidelines, annotators should have annotated at least one instance of a stage parameter label-value stage parameter in one report, if it appeared. We did not use multi-class classification for each stage parameter because of overlap possibilities. The performance for the baselines are shown in Table 7.28, where evaluation is for the document level, which is the same as those defined in Chapter 3.2.

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 4.2: Best baseline performances for training set.
(Freq = frequency, Class = classification method)

Regular expression extraction

Text evidence for *Child-pugh* and a portion of *ECOG* stage parameters were in a form ammenable to regular expression capture. The regular expression rules were created by studying the examples in the training set. The details of our rules along with some further description of *Child-Pugh* and *ECOG* classification challenges are discussed in Chapter 5.

Sentence classification with statistical feature selection

Several stage parameters, *ascites*, *hepatic encephalopathy*, *portal hypertension*, *extrahepatic invasion*, *metastasis*, and *macrovascular invasion*, were identified with text evidence that could be, for the most part, assumed to be identified within sentences. For these parameters, we used a sentence text classification approach using statistically selected features, including ranked n-grams and UMLS concepts, with assertion classification. To overcome our limited annotation, we enriched the data set with a subset of non-expert annotated data. Our detailed process and results are described in Chapter 6.

Tumor characteristics extraction

Tumor-related stage parameters, *tumor number*, *tumor size*, and *tumor morphology* required aggregated information over multiple sentences. For this extraction system, we built a pipeline that extracted tumor templates, performed reference resolution, and finally assigned *tumor number*, *tumor size*, and *tumor morphology* based on a rule-based algorithm. The entire system is the subject of Chapter 7.

4.1.3 Patient level classifications

Patient level classifications, described in Chapter 8, incorporates stage parameter parts of the pipeline to assign holistic values for each stage parameter. The final stage classifications take the 11 patient level stage parameters, 4 laboratory values and output the final 3 overall stages. During this chapter, we show various experiments using different levels (raw text,

text annotations, patient level stage parameters, etc) of system and gold annotations.

4.2 Summary

The development of a liver cancer staging system involved various extraction and classification modules, including a two step regular expression to extract *Child-Pugh* and *ECOG* stage mentions, a feature selected sentence classification strategy, and a pipeline of entity and relation extraction and reference resolution classification to identify tumor characteristics. In our work, we isolate each component, tune and build an appropriate module and characterize the performance individually. Patient classification performances are, of course, the most clinically relevant, which we provide in Chapter 8, along with experiments using different levels of patient gold annotations.

Chapter 5

CHILD-PUGH AND *ECOG* CLASSIFICATIONS

Child-Pugh and *ECOG* variables for our liver cancer classification are examples of the recursive nature in defining phenotypes, as they themselves are also stages. Examples of their expressions are given in Table 5.1. The stage definitions for *Child-Pugh* and *ECOG* are given in Table 5.2 and 5.3, respectively.

Given this, there is at least two ways to represent such variables: (1) the consolidated stage category, e.g. *Child-Pugh A*, and (2) the set of compositional staging factors, e.g. $\{$*ascites - none*, *hepatic encephalopathy - mild*, *bilirubin* < 2 *mg/dL*$\}$ with the algorithm in Table 5.2. In this chapter, we present our approach to extracting (1) for both stage parameters, meanwhile (2) for *Child-Pugh* is deferred to later chapters and (2) for *ECOG* will not be addressed in this thesis for reasons addressed in Section 5.5.

Here we present our text extraction performance with evaluation against our previous simple document baseline.

Label	**Description**	**Example Text**
Child-Pugh class	Stage that summarizes liver function	"Child-Pugh: A" "She is currently a Child's B score 7" "He is Child class A" "CTP-A6 cirrhosis"
ECOG performance status	Measure of general well-being of a patient, (0-5).	"ECOG performance 0." "He works out at a gym" "Notable for chronic fatigue." "She has been doing relatively well and has been undertaking her daily activities without any problems"

Table 5.1: Text annotation examples for *ECOG* and *Child-Pugh*

Variable	Points		
	1	2	3
Albumin (g/dL)	> 3.5	2.8-3.5	< 2.8
Ascites	None	Mild/Moderate	Severe
Bilirubin (Total) (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 5.2: *Child-Pugh* parameters [125]. 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.

Status	**Description**
0	Fully active, able to carry on all pre-disease performance without restriction
1	Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature, e.g., light house work, office work
2	Ambulatory and capable of all selfcare but unable to carry out any work activities; up and about more than 50% of waking hours
3	Capable of only limited selfcare; confined to bed or chair more than 50% of waking hours
4	Completely disabled; cannot carry on any selfcare; totally confined to bed or chair
5	Dead

Table 5.3: *ECOG* Guidelines [139]

5.1 Explicit vs non-explicit evidence

We refer to evidence exhibited as the consolidated stage, (1), as explicit evidence, while all other contributing factor evidence, (2), will be regarded as non-explicit evidence. Our annotation strategies for the two cases for each of *Child-Pugh* and *ECOG* were handled differently. *Child-Pugh* explicit evidence was marked and normalized, but the factors related to *Child-Pugh*, *ascites* and *hepatic encephalopathy*, were marked separately. On the other hand, *ECOG* explicit and non-explicit evidence were highlighted and normalized to an *ECOG*

value, e.g. *ECOG 0*. This discrepancy was partly due to *ECOG* factors being relatively less well-defined and criteria being more difficult to enumerate exhaustively. For example, "all-pre-disease conditions", "able to work", "50%" of the time are imprecise and subjective to define. Moreover, *Child-Pugh* factors, such as *ascites* and *hepatic encephalopathy*, are recognized medical concepts by themselves.

5.2 Related work

The closest relevant work we have found was with the rule-based regular expression extraction of *Child-Pugh* class from Ping et al [135]. *ECOG* performance status is frequently used in clinical trials criteria, however we were unable to locate any previously published extraction systems. To our knowledge there are no prior work classifying documents to non-explicit *ECOG* or *Child-Pugh*. We follow the same approach as Ping et al [135] with *Child-Pugh*, a two-step regular expression extraction, for both explicit *ECOG* or *Child-Pugh* extraction.

5.3 Two-step regular expression approach for explicit stages

We designed a two-step regular expression approach to extract explicit mentions of *Child-Pugh*. The first step identifies trigger terms. Afterwards, a second regular expression was used to capture the related value. The trigger regular expressions for *Child-Pugh* and *ECOG* with corresponding examples are shown in Table 5.4 and 5.5, respectively. We found that specifying capitalizations was necessary especially for less formal references to *Child-Pugh*, e.g. *"Child's class A"*.

In the second step, another regular expression was ran, and specific values were captured and mapped to normalized values. For example, the expressions for *Child-Pugh*, *"A"*, *"B"*, and *"C"*, would be mapped according to their respective stages. However, at times only the numerical score was available. For these cases, as according to the Table 3.11 algorithm, *Child-Pugh* scores of 5-6 are mapped to *A*, 7-9 to *B*, and 10-15 to *C*.

ECOG extraction was much simpler, as shown in Table 5.5. The value-mappings for *0*, *1*, and *2* were themselves, while values 3-5 were mapped to ≥ 3.

Regular expression	Examples
child\\s+ (class|classification|category|score)	"Child class A"
CTP	"CTP A"
ctp.*(class|classification|category|score)	"ctp score of 6"
[Cc][hH][iI][lL][dD].*?[Pp][uU][gG][hH]	"Child-Pugh A", "Child-Turcotte-Pugh A"
[Cc]hild's	"Child's A"

Table 5.4: Child-Pugh trigger regular expression

Regular expression	Examples
[Ee][Cc][Oo][Gg]	"ecog 0 to 1"
	"(ecog performance status): (0)"
	"ecog performance status 1."
	"ECOG, 1, in summary"

Table 5.5: ECOG trigger regular expression

5.3.1 Evaluation

The data set was based on the previously split, 5-fold cross-validation of 160 patients from the training set. Performance was measured in terms of precision, recall, and F1 score defined previously in Chapter 3.2, for the text span and document levels.

5.3.2 Results

Table 5.6 shows a breakdown of the microscore text extraction performance. As the dataset was annotated sparsely, compared to the gold standard, precision was low.

We manually reviewed both false positives and false negatives for both stage parameters. In the case of *Child-Pugh*, we found false positives to be all the unmarked cases not covered by our sparse annotation. From studying false negatives, on the other hand, we found 3 known classes of problems that we miss. The first are instances of *"child a"* or *"child b"*.

We did not code for this regular expression was because "child" is a common word. The second known category that we missed were later parts of hedge cases, e.g. in *"child-pugh score of 6 or 7"* only *"child-pugh 6"* is captured. As our system identifies the *"6"*, starting from that text span, it is possible to expand right-ward to identify conjunctions and add in other cases. We leave this for further work. The final class of false negative errors we miss are other abbreviations besides *"CTP"*. For example, one false negative was in a sentence *"a 65 year old man with CPA(6) HCV-related cirrhosis"*. While the *"cirrhosis"* indicates this may be *"Child-Pugh A"* with score 6, in the general medical domain, *"CPA"* has many other acronyms such as *"cardiopulmonary arrest"*, or *"childhood physical abuse"*, therefore we did not include this in our algorithm.

Category	**Pos**	**TP**	**FP**	**FN**	**P**	**R**	**F1**
A	53	49	25	4	0.662	0.925	0.772
B	25	24	9	1	0.727	0.960	0.828
C	7	7	6	0	0.538	1.00	0.700
ALL	85	80	40	5	0.667	0.941	0.780

Table 5.6: Explicit *Child-Pugh* extraction results

Table 5.7 shows the performance for *ECOG* extraction. Since we annotated both explicit and non-explicit evidence under this category, and the annotation was not comprehensive, the raw precision and recall scores are not reliable estimates of performance. On manual inspection we found that false positives were correct but unannotated. For the most part, the false negatives were due to the non-explicit text evidence. One exception, like with *Child-Pugh*, was with hedging cases, which we did not account for, e.g. *"ecog 0-1"*, where we only extract *ECOG 0*. Another exception was when the *ECOG* trigger was not in the same sentence as the value. A borderline case is when performance status is mentioned but not ECOG performance status in particular, *"Otherwise his performance status is good at 1 bordering on 0."*

Category	Pos	TP	FP	FN	P	R	F1
0	141	83	18	58	0.822	0.589	0.686
1	84	54	15	30	0.783	0.642	0.706
2	25	16	4	9	0.800	0.640	0.711
≥ 3	9	4	1	5	0.800	0.444	0.571
ALL	259	157	38	102	0.805	0.606	0.692

Table 5.7: Explicit *ECOG* extraction results

Table 5.8 shows the results resolved for the document level compared to our previous baseline. Unsurprisingly, our regular expression rule-based approach resulted in substantial improvement over the n-gram document classification baseline.

		Baseline				Regex		
Category	**Freq.**	**Class.**	**P**	**R**	**F1**	**P**	**R**	**F1**
A	53	DT	0.46	0.49	0.47	0.83	0.94	0.88
B	25	C45	0.84	0.64	0.73	0.92	0.96	0.94
C	7	DT	0.50	0.14	0.22	0.88	1.0	0.93

Table 5.8: Explicit *Child-pugh* document classification results

		Baseline				Regex		
Category	**Freq.**	**Class.**	**P**	**R**	**F1**	**P**	**R**	**F1**
0	105	C45	0.71	0.71	0.71	0.96	0.68	0.79
1	65	DT	0.85	0.54	0.66	1.00	0.72	0.84
2	18	C45	0.89	0.44	0.59	1.00	0.78	0.88
≥ 3	8	DT	0.25	0.13	0.17	1.00	0.38	0.55

Table 5.9: Explicit *ECOG* document classification results

5.4 Non-explicit Child-Pugh

In the training set, only 67 out of 160 patients had one or more explicit markings for *Child-Pugh*; our extraction method detected 73/160 patients' *Child-Pugh* values. This leaves approximately 54% of patients in need of non-explicit *Child-Pugh* calculations. The question arises of how to handle calculated *Child-Pugh* stages against known explicit stages. Furthermore, whether to resolve these issues if conflicts occur at the document level or at the patient level, is open for debate. The benefit of resolving at the document level would be to clear up any typos or internal consistency before passing incorrect information to the patient level. However, while it is possible to disambiguate at the document level, we do not have document level gold annotations for this. As a consequence, we defer this disambiguation for the patient level. How we handled multiple values of *Child-Pugh*, some coming from explicit text mentions and others from calculated text (*ascites* and *hepatic encephalopathy* values) and from laboratory data, will be explored in Chapter 8.

5.5 Non-explicit ECOG performance status

In the training set, 136 out of 160 patients had one or more markings (both explicit and non-explicit) for *ECOG*; our extraction method detected 105/160 patients' *ECOG* values. This leaves approximately 34% of patients in need of *ECOG* staging. As mentioned previously, we ultimately decided not to classify for non-explicit *ECOG*. A factor in this decision is the relative simplicity of the clinical assessment process (a couple minutes) combined with the difficulty of the classification problem. Therefore, a classification system for this would have little chance of good performance, meanwhile it would have little potential cost-savings. In the following paragraphs, we offer a discussion of the challenging aspects in classifying *ECOG* performance status as an NLP task.

ECOG performance status is one of several measures that give a global assessment of patient functional capacity, ranging from 0 (normally active) to 5 (dead). Other systems include Karnofsky Performance Status (KPS) and Palliative Performance Status (PPS). In

research settings, performance status are used as criteria for clinical trials and to assess treatment efficacy. In clinical practice, performance status is used to estimate prognosis or therapy and to assess needs for home care. [102]

Though considered subjective, these performance status are useful, as no single lab or quantifiable objective measure has the ability to capture the overall well-being of a patient. However, there are definitional difficulties [139]. For example, "able to carry on all pre-disease performance" is challenging. If *ECOG* status is used as primary or secondary measures during cancer treatments, which are known to have some adverse side-effects, should these transitory events be part of the assessment? This definition is also contingent on the patients' normal daily levels of strenuous activities as well as their own ability to cope with pain or discomfort. If drug side-effects are factored in, should weakened state due to a surgical intervention also be included?

A young healthy paraplegic may be otherwise healthy before the effects of a cancer takes a toll. Such a person may remain "fully active" but they are also technically "physically restricted" and only capable of "limited selfcare", which may be exasperated more intensely during treatments. Consider another case of an elderly patient with end-stage renal disease and some other cancer. A person with serious renal disease may not be ideal for a clinical trial or may have poor prognosis regardless of the advancing cancer. Age can also have an affect on functional status. For example, an elderly patient may be otherwise healthy, but not able to do strenuous activity due to age. Or, to provide another comparison, would a young healthy person that requires routine dialysis be rated the same or differently from a healthy elderly person with the same needs?

In clinical practice, performance statuses are assigned based on observation of the patient. *ECOG* statuses have been known to be different depending on each assessor. For example, there has been documented differences between physicians, nurses, and patients [201][20]. Typically, agreement between 0-2 and >3 stage cut-offs, which is the usual cut-off for clinical trials, is relatively high; actual kappa agreements between individual categories are more variable.

As an extraction task, the broad categorization of performance status and the conditional complex requirements renders the problem into quite a byzantine challenge. Firstly, ascertaining the patient's functional status from patient records is already a second-hand observation. Thus, there are already issues such as reporting biases and missing information. Processing information from text, not only do specialized medical terms need to be recognized and considered with their severity and the affect on the patient, their temporal attributes in regards the the past and current status needs to be considered for the latest evidence. Moreover, there are cases in which additional inferences may be required. For example, occupations are also indicators. A patient who is currently working as a airplane pilot most likely does not have any alarming physical restrictions. Another possible indicator includes institution names such as hospitals or department referrals, e.g. physical therapy. Similar to other clinical classification tasks, medication information, such as disease conditions, can provide clues to the severity of illness. This adds another burden of such a classification task, since it would require hiring highly trained medical domain expert annotators.

5.6 Summary

Child-Pugh and *ECOG* parameters are two cases of sub-stages used for liver cancer staging. For these parameters, sometimes the actual stages are recorded in text explicitly. However, when it is not recorded, the annotator must take into account algorithms used to perform the staging. Explicit *Child-Pugh* and *ECOG* stages extraction, were captured through the regular expression methods described here with reasonably success. The resolution of multiple explicit and calculated *Child-Pugh* will be discussed during patient level classifications. Because *ECOG* is difficult to capture both because of its wide scope as well as its subjective and amorphously defined nature, we decided to only extract its explicit values.

Chapter 6

SENTENCE CLASSIFICATION FOR STAGE PARAMETER NORMALIZATION USING STATISTICALLY SELECTED FEATURES

In this chapter, we describe our sentence classification approach to identify and normalize passages related to the stage parameters of *ascites*, *hepatic encephalopathy*, *extrahepatic invasion*, *macro vascular invasion*, and *metastasis*, with their respective severity values. Examples of the text evidence for each parameter are shown in Table 6.1.

For this set of stage parameters, we cast the classification task as a multi-label sentence classification for each stage parameter, using statistically selected features. Our final evaluation was again compared at the document level against our previous simple document classification baseline.

6.1 Related work

The concepts we identify here are in many ways related to the previous information extraction systems for finding cancer characteristics discussed in Chapter 2.2.2 or the sub-stages of the cancer prediction systems in Chapter 2.2.1, which requires identifying and normalization of concepts with their severity. However, of note, the dataset we have here cannot be considered a straight-forward entity and relation extraction systems, as in some previous work such as Ping et al [135], and Wang et al [181]. Firstly, annotation may consume several clauses. Secondly, our annotation for each stage parameter is annotated for its corresponding concept mentions, e.g. *"mild ascites"* as well as for related signs and symptoms text evidence, e.g. *"free fluid"*.

Two sets of methodologies from related works relevant for these parameters are those

Label	Description	Example Text
Ascites	Accumulation of fluid in the peritoneal cavity.	"He denies increasing abdominal girth" "He has no problems with edema or ascites" "No free fluid in the abdomen" "Small volume ascites"
Extrahepatic invasion	Spread of cancer outside of liver	"Extrahepatic metastatic disease: None" "Lymph nodes: Scattered subcentimeter lymph nodes not pathologic by size criteri" "No evidence of extrahepatic extension"
Hepatic encephalopathy	Confusion or altered consciousness due to liver failure	"He has no significant ascites or encephalopathy cirrhosis has been complicated by hepatic encephalopathy" "Lactulose" "The patient denies any confusion, forgetfulness, or other symptoms of hepatic encephalopathy"
Macrovascular invasion	Spread of cancer to nearby blood vessels	"No evidence of portal vein thrombosis" "No obvious invasion of vessels is noted." "Portal veins are patent." "Vascular invasion: None"
Metastasis	Spread of cancer to outside-liver lymph nodes	"Lymph nodes suspicious for metastatic involvement: None" "No abnormal lymph nodes" "No evidence of extrahepatic extension or metastasis" "No other findings suggestive of extrahepatic disease"
Portal hypertension	Elevation of hepatic venous pressure gradient to $>$ 5mm Hg	"No evidence of portal HTN" "Patient had an EGD which showed small varices" "Recanalization of the umbilical vein, perigastric and peri-splenic varices compatible with portal hypertension physiology"

Table 6.1: Text annotation examples

that exercise named entity recognition, such as Ou and Patrick [127] and Ashish et al [23], and the sub-stage sentence classifications of McCowan et al [112] and Martinez and Li [108], the latter of which also included some document classification approach. While each has its merits, because our annotation highlights were not linguistically stringent (e.g. highlighting well-defined noun phrases), we decided to use a similar approach to McCowan et al's [112] sentence classification.

Though we also use a sentence classification approach, there are some distinct differences between our approach compared to the work in McCowan et al [112]. Their classification was for TNM staging, with which each T, N, and M represents separate sub-stages each with their individual variety of characteristics. For example, the T sub-stage indicates size. Therefore, a T sub-stage value may be T0 for no signs of tumor or T1, T2, T3, and T4 for increasing sizes. Furthermore, individually, a T4 classification may require several distinct criteria such as whether any of the following occurred: "great vessel invasion", "pericardium invasion", and "separate tumor nodules in same lobe". Their strategy used a divide-and-conquer methodology. For example, the "separate tumor nodules in same lobe" criteria used a keyword lookup approach; while, the former two criteria, in addition to a keyword filtering, were treated using the same two-level sentence classifier (first level for relevance, second for identifying true positive). Using another example, for N1 classification, two separate parallel pipelines were used to identify for positive N1: one two-step sentence classifier for "hilar lymph node involvement" and another two-step sentence classifier for "mediastinal lymph node involvement". Each sub-stage pipeline was described in their paper.

In contrast to McCowan et al's [112], the stage parameters for our sentence classifications are not sub-stages. We did not subdivide each individual stage parameter and its values, e.g. using separate classifiers for different instances of *ascites - none*. Futhermore, though we include a a filter for sentences from specific document types and employ negative class instance sampling, our classification of the sentence is a simple one step classification.

In terms of the relevance of our stage parameters, our AJCC stages are in fact based on summarized TNM stages. Therefore, there are several stage parameters relevant to TNM, including *extrahepatic invasion*, *macrovascular invasion*, and *metastasis*, discussed in this chapter, as well as tumor size and number, discussed in the next chapter. Unfortunately, because our categorical divisions are not summarized the same way as McCowan et al's [112], our results cannot be directly compared.

6.2 Methods

The data set was based on the previously split, 5-fold cross-validation of 160 patients from the training set. Each stage parameter classification, was a multi-label classification problem at the sentence level. For example, if a sentence has no marked *ascites* text evidence, it is considered *NEGATIVE*. If a sentence has annotations for *ascites-none*, *ascites-none* would be its sentence tag. If there were multiple values, e.g. both *ascites-mild* and *ascites-moderate*, appeared, the tags would be a combined value, e.g. *ascites-mild_##_moderate* for a sentence *"Mild to moderate ascites."*. We used maximum entropy as the classifier, with features described in Section 6.2.2. In addition to the expert-annotated data, described previously, we add a small amount of non-expert annotated data described in Section 6.2.1. As according to our annotation, we restricted sentence classification for certain stage parameters to particular document types; they are outlined in Table 4.1. A figure of our sentence classification pipeline is shown in Figure 6.1.

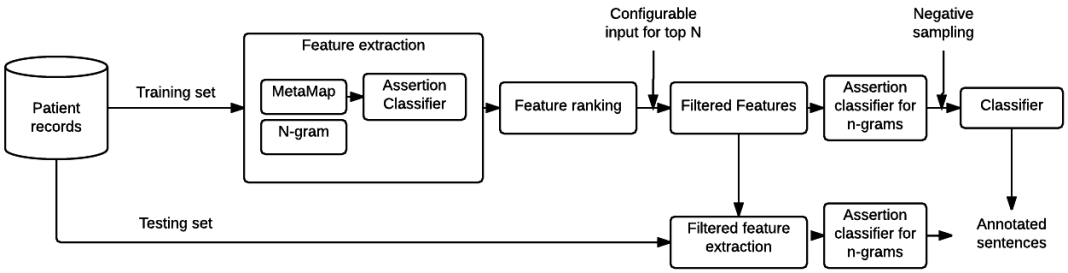


Figure 6.1: Sentence classification workflow

6.2.1 Annotation enrichment

During annotation, to consider consistency, text evidence from the same sections were marked by separate annotators and compared during inter-annotator agreement. To save time, not all sentences with the same information in the document were annotated. However, this left

positive evidence in the rest of the report sentences (some even in the same sections) that are unannotated. Moreover, text from different sections of a report have different syntactic and semantic representations despite representing the same information. For example, in the **History of Present Illness** section, *"He has no abdominal distention"*, and in the **Physical Examination** section, *"ABDOMEN: Soft, nontender, nondistended."*, both correspond to text evidence *ascites - none*.

In order to enrich the dataset, we randomly sampled 70 documents, including both radiology and clinical notes, from the training set of our HCC patient cohort and annotated these documents fully for all text evidence for the 6 stage parameters in this chapter. In total, these reports included were 5855 sentences. In contrast to previous annotations, these annotations were carried out by a non-expert. Therefore, the result for these 70 documents is fully annotated both positive and negative sentences, with mixed expert and non-expert annotations.

The benefit of this additional annotation is a greater number of training examples of the semantic variations for different parts of the document. The unmarked sentences in these documents may also be considered true negatives (whereas other unmarked sentences may either be true negatives or unmarked positives). These annotated sentences were included only during the classification training phase.

6.2.2 Features

Selected features were determined by the N-top ranked 1-, 2-, and 3- grams and asserted[1] UMLS concepts found in the original training set (excluding additional annotations). UMLS concepts were identified using MetaMap [22]. Assertion classification was performed using an in-house assertion classifier [27]. The top N-grams combined with its assertion classification for 1-, 2-, and 3- grams were also added. Another feature was a binary indicator if a categories' significance feature occurred. A final feature was turned on if no features were

[1] Assertion classification labels text as one of several categories: not associated with the patient, hypothetical, conditional, possible, absent, and present.

detected.

As an example, for the classification of *ascites*, with $N = 10$. The features would include the 1-, 2-, 3- grams (with asserted versions) and asserted UMLS concepts associated with the top 10 values for each *"none"*, *"mild/suppressed"*, and *"moderate/severe"*. For diversity, we used several significance measures, χ^2, t-test, and pointwise mutual information (PMI), defined below, and merged the resulting lists.

While this method choses N distinct significance values, the actual number of features may be much more than N because many features have the same significance values, given the small number of examples. Furthermore, since the each classification has of various values, e.g. *"none"* and *"mild/suppressed"*, each with its own significance lists, and multiple significance metrics, the set of features can become quite large even for a small N.

6.2.3 Measures of significance

Each feature, e.g. *1-gram=ascites*, and label category, e.g. *ascites-none*, was constructed into a contingency table and given a significance measure according to the following metrics.

$\underline{\chi^2}$

χ^2 is defined by the following equation for each observation type i configuration, e.g. (no occurrence of feature *1-gram=ascites* and positive occurrence *ascites-none*):

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \tag{6.1}$$

where O_i is the number of observations of type i, n is the total number of observations (number of sentences), and E_i is the expected number of observations for that configuration type. Therefore, for the example configuration, observed $O_i = frequency(nofeature, category)$ and expected $E_i = \frac{frequency(nofeature)}{n} * \frac{freq(category)}{n} * n$. This continues for other configurations, e.g. (occurrence of feature, no occurrence of category), etc. for all four combinations; the final value is the sum.

T-test

The Student's t-test is defined by the following equation:

$$t = \frac{x - x_0}{s/\sqrt{n}} \tag{6.2}$$

where we compare our observations against assuming the probabilities of a feature or a category occurring are from uniform distributions. Thus, we let the observed distribution of a positive feature, category combination to be, $x = \frac{frequency(feature,category)}{n}$ and the default to be, $x_0 = \frac{frequency(feature)}{n}\frac{frequency(category)}{n}$. The variance, s^2, was approximated by $p(1-p)$, where $p = x_0$, according to the Bernoulli distribution.

Pointwise mutual information

Pointwise mutual information (PMI) is defined by the following equation:

$$pmi(feature, category) = log\frac{p(feature, category)}{p(feature)p(category)} \tag{6.3}$$

Similar to previous calculations, the probability of $p(feature, category) = \frac{frequency(feature,category)}{n}$ and $p(feature) = \frac{frequency(feature)}{n}$, $p(category) = \frac{frequency(category)}{n}$.

6.2.4 Negative sampling

Negative sentences, defined as unannotated sentences (does not count true negatives from the annotation enrichment), were randomly sampled to accompany the annotated sentences. The number of negative sentences, x, was optimized during experimentation.

6.2.5 Evaluation

Performance was measured in terms of precision, recall, and F1 score defined previously in Chapter 3.2, for the text span and document levels. Sentence classifications were converted back to text spans and compared to gold-annotated text spans. Compound values, e.g. *ascites-mild_##_moderate*, were broken into multiple text annotations.

6.3 Results

During experimentation, we optimized for the largest N (top number of significance values for feature selection) and x (ratio to positive examples) to maximize recall performance (precision and F1 were not appropriate given the possibility of unannotated positive examples). We tested values of $N = \{1, 5, 10, 20, 50, 100\}$ and $x = \{0, 0.5, 1, 5, 10, 50, 100\}$. An example of features in the top $N = 1$ for *macrovascular_invasion-yes-minor_branch* is in Table 6.2.

1-gram	**2-gram**	**3-gram**	**asserted-UMLS concept**
associated	venous tumor	associated thrombosis of	present-abnormality
anteriorly	thrombosed ,	portal venous tumor	present-clinicaltrialbranch
portal	more completely	which is thrombosis	
	completely described	venous tumor thrombus	
	lesion extends	lesion extends to	
	associated thrombosis	to the anterior	
	.....	...	

Table 6.2: Significant features for $N = 1$ top significance values for *Macrovascular_invasion-Yes-minor_branch*

The classifier for *ascites* was optimized at $N = 50$, while *hepatic encephalopathy* was optimized at $N = 10$. All others were optimized at $N = 20$. We leave differing the N value for 1-gram, 2-gram, 3-gram, and asserted-UMLS concepts for future work. A ratio of $x = 1.0$ (equal number of positive examples) to draw unannotated sentences as negative examples was the highest ratio that maintained high recall. Table 6.3 shows the resulting performance.

Table 6.4 shows the comparison of the sentence classification resolved at the document level compared to our previous baseline. Cases with no annotation enrichment and no feature selection, and no feature selection only are included for comparison. In all but three case, *hepatic encephalopathy - mild*, *macrovascular invasion - yes-minor_branch*, and *metastasis - yes-distal*, we see that the sentence classification approach improves performance. Annotation enrichment in general helped with most of the classifications, however, *ascites - moderate-*

			No feature selection			Feature selection		
Label	**Freq.**	**Value**	**P**	**R**	**F1**	**P**	**R**	**F1**
Ascites	46	Mild	0.25	0.65	0.37	0.20	0.72	0.31
	21	Moderate-Severe	0.36	0.38	0.37	0.42	0.38	0.40
	146	None	0.32	0.91	0.47	0.29	0.90	0.43
	213	ALL	0.30	**0.80**	0.44	0.27	**0.81**	0.40
Extrahepatic	62	No	0.84	0.82	0.83	0.90	0.85	0.88
invasion	3	Yes	0.00	0.00	0.00	0.00	0.00	0.00
	65	ALL	0.84	**0.78**	0.81	0.90	**0.82**	0.85
Hepatic	41	Mild	0.41	0.29	0.34	0.34	0.56	0.42
encephalopathy	95	None	0.53	0.81	0.64	0.49	0.82	0.61
	1	Severe	0.00	0.00	0.00	0.00	0.00	0.00
	137	ALL	0.51	**0.65**	0.57	0.44	**0.74**	0.55
Macro-vascular	132	No	0.58	0.92	0.71	0.54	0.92	0.68
invasion	20	Yes-major_branch	0.36	0.60	0.45	0.22	0.75	0.34
	8	Yes-minor_branch	0.33	0.13	0.18	0.29	0.25	0.27
	160	ALL	0.55	**0.84**	0.67	0.46	**0.86**	0.60
Metastasis	95	No	0.47	0.84	0.61	0.47	0.85	0.60
	7	Yes-distal	0.00	0.00	0.00	0.00	0.00	0.00
	7	Yes-regional	0.00	0.00	0.00	0.50	0.29	0.36
	126	ALL	0.47	**0.75**	0.58	0.46	**0.77**	0.58
Portal	11	No	0.50	0.09	0.15	0.33	0.09	0.14
hypertension	106	Yes	0.44	0.69	0.54	0.40	0.72	0.52
	117	ALL	0.44	**0.63**	0.52	0.40	**0.66**	0.50

Table 6.3: Sentence classification performances

severe, *hepatic encephalopathy -mild*, *metastasis - yes-regional*, and *portal hypertension - no*, did better without more annotations. The use of statistical feature selection was successful in about half of the cases. Specifically, the minority classes benefitted more in such cases as for *ascites*, *macrovascular invasion*, and *metastasis*.

			Document baseline				No enrichment, no feature selection			No feature selection			Feature selection		
Label	**Freq.**	**Value**	**Class.**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Ascites	44	Mild	C45	0.24	0.18	0.21	0.35	0.80	0.49	0.48	0.70	0.57	**0.44**	**0.86**	**0.58**
	20	Moderate-Severe	DT	0.50	0.30	0.38	0.61	0.55	0.58	0.59	0.50	0.54	**0.67**	**0.50**	**0.59**
	146	None	DT	0.77	0.36	0.49	0.53	0.97	0.67	**0.59**	**0.98**	**0.73**	0.55	0.95	0.67
Extrahepatic	59	No	SVM	0.81	0.85	0.83	0.55	0.98	0.70	0.87	0.90	0.88	**0.90**	**0.90**	**0.90**
invasion	2	Yes	≈	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Hepatic	34	Mild	**DT**	**0.70**	**0.76**	**0.73**	0.57	0.85	0.68	0.74	0.50	0.60	0.65	0.82	0.72
encephalopathy	95	None	DT	0.71	0.73	0.72	0.44	0.96	0.60	**0.64**	**0.85**	**0.73**	**0.62**	**0.87**	**0.73**
	1	Severe	≈	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Macro-vascular	127	No	NB	0.71	0.96	0.82	0.76	0.98	0.86	**0.80**	**0.96**	**0.87**	0.78	0.97	0.86
invasion	20	Yes-major_branch	C45	0.50	0.55	0.52	0.56	0.90	0.69	**0.67**	**0.80**	**0.73**	0.46	0.90	0.61
	8	Yes-minor_branch	C45	**1.00**	**0.50**	**0.67**	1.00	0.25	0.40	1.00	0.25	0.40	0.80	0.50	0.62
Metastasis	108	No	DT	0.78	0.70	0.74	0.65	0.97	0.78	0.74	0.93	0.82	**0.75**	**0.94**	**0.83**
	6	Yes-distal	**DT**	**0.50**	**0.17**	**0.25**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	7	Yes-regional	≈	0.00	0.00	0.00	1.00	0.14	0.25	0.00	0.00	0.00	**0.50**	**0.29**	**0.36**
Portal	5	No	≈	0.00	0.00	0.00	**0.50**	**0.40**	**0.44**	0.50	0.20	0.29	0.33	0.20	0.25
hypertension	84	Yes	C45	0.84	0.80	0.82	0.69	1.00	0.82	0.92	0.92	0.92	**0.89**	**0.98**	**0.93**

Table 6.4: System evaluated at document level compared to document classification baseline. Bolded rows shows the best F1 performances for each stage parameter value.

6.4 Discussion and error analysis

We analyzed the test set for one fold of the cross validation to identify the accuracy of the sentence classifications and give some brief characterizations below.

6.4.1 Sentence classification

Ascites

We found that many false positives (98% for *none*, 48% *mild*, and 100% *moderate/severe*) were correct. Much of the misclassifications can be attributable to the confusion between *moderate-severe* category for some *mild* category cases. This was most likely due to the smaller amount of *moderate-severe* cases as well as the higher variety at which it was found, e.g. *"gross ascites"*, *"extensive ascites"*, *"refractory ascites"*, or *"ascites [...] required multi-*

ple large volume paracenteses". Other false positives were due to cirrhosis-related sentences.

Extrahepatic invasion

This stage parameter was very imbalanced. For the majority category, it was highly accurate. The affirmative, *Extrahepatic invasion - yes*, would most likely occur for patients at the later stages of liver cancer which our population does not well represent.

Hepatic encephalopathy

For false positives, 78% *none* and 64% *mild* were accurate. This classification was one in which drug information was used as text evidence. However, the use of the particular drugs might have been a source of confusion. For example, if a patient was taking *"lactulose"* they would be considered to at least have *hepatic encephalopathy - mild*. However, there were passages such as *"supposed to be on lactulose but not taking"*. Other challenges include some evidence that relies on more complex inference, e.g. *"he only complains of mild, occasional confusion"* or *"has well-compensated cirrhosis to date. he has not had any significant complications"*. The *severe* category was under-represented.

Macrovascular invasion

About 2/3 of the false positives for the *yes-major_branch* category were correct while the rest were misclassifiifed *yes-minor_branch*. The confusion between the two categories is understandable as there are many variations and word-orders to express the major blood vessels, e.g. *"the portal vein"*, in contrast to the subparts of the major blood vessels, e.g. *"superior mesenteric vein"* or *"anterior branches of the right portal vein"*. For the majority category, *no*, 84% of false positives were correct, e.g. *"vascular invasion: none"* or *"hepatic arteries, veins, and portal veins are patent"*. False negatives for all categories included different semantic variations expressing blockages of blood vessels using different types of terms, e.g. *"involvement"*, *"infiltrated"*, *"distending"* or *"occluded"*.

Metastasis

Negative findings of *metastasis*, also the majority category for this stage parameter, *metastasis - no*, had many instances of straight-forward text clues, e.g. *"extrahepatic metastatic disease: none"*, leading to a high number of false positives which were correct. However, there was some confounding features related to lymph nodes. This was due to the inclusion of lymph nodes as evidence of *metastasis - yes_regional*, but only if they were pathologically related to the cancer, e.g. *"Enlarged peripancreatic and porta hepatis lymph nodes may be reactive or due to nodal metastases."* would be marked, but *"Enlarged peripancreatic and porta hepatis lymph nodes"* alone would not be. If there were *"no enlarged lymph nodes"*, *"no lymphadenopathy"*, the lymph nodes were *"stable"* or did not *"demonstrate enhancement or washout to suggest a represent metastatic disease"*, these would be considered to support *metastasis - none*. For distant metastasis, *metastasis - yes_distal*, any malignant tumor findings in other anatomic locations outside of the liver would affirmatively support this category. However, again, these types of patients were not well-represented in our dataset.

Portal hypertension

While there were instances in the majority class text evidence that mention *portal hypertension*, a significant portion of the evidence regarded *portal hypertension* signs, e.g. *"large gastroesophageal varicose"*, *"spleen: enlarged"*, or *"splenomegaly"*. Interestingly the *portal hypertension - no* category consistently missed some relatively simple text evidence, that our features should have captured, e.g. *"No, there are no signs of portal hypertension"*, *"no evidence of portal htn"*, or *"No, liver cirrhosis with stigmata of portal hypertension"*. The false negatives for *portal hypertension - yes*, similarly had very unequivocal mentions of *portal hypertension*, e.g. *"the patient is a child's a cirrhotic with known portal hypertension"*. While normalization involving the signs and symptoms related to the portal venous system would help, e.g. grouping all vessels or organs related to the portal venous systems, merely annotating more would at least resolve these more obvious cases without need for extra fea-

ture extraction techniques. We found 95% of false positives correct for the majority, *yes*, class.

Analyzing the effect of feature selection on the feature space size, while feature selection approximately halved the number of features for *ascites*, *macrovascular invasion*, and *metastasis*, the other classifications had approximately the same number of features. This is attributable to the mechanism of feature selection which relies on some spread of significance values for each label-value stage parameter category. For example, since *extrahepatic invasion - yes* had very few examples, because of few significance value brackets, the top 20 values may turn up all the features.

These problems may be mitigated by careful tuning of N, individualized by each classification category, e.g. the N for *ascites-none* should not be the same as the N for *ascites-mild*. Even the N for 1-, 2-, 3-, and asserted UMLS concepts can be individualized and tuned to receive a more reliable performance. A key factor is the need for more expert training data. In our experiments, a small amount of non-expert enriched data, in general, facilitated improved performances. However, it is reasonable to conclude that additional expert-annotated data would be of higher quality, thereby yielding better performance and generalizability.

We used relatively simple features, N-grams and UMLS concepts, that are sufficient for variables with fewer semantic variations (such as for *ascites*). Normalization for drugs in *hepatic encephalopathy*, e.g. grouping similar drug classes into one feature, normalization for larger and smaller blood vessels in the liver *macrovascular invasion*, e.g. grouping large blood vessels under one feature, and normalization for organs and vessels that are a component of the portal venous system for *portal hypertension* would allow, theoretically, for better upstream features during feature selection. Normalization for findings associated with liver and non-liver locations would help for *extrahepatic invasion* and *metastasis*.

Finally, we note that we assumed individual sentences may characterize a patient, this idea is somewhat murky. For example, for *metastasis - none*, information regarding enlarged lymph nodes is conveyed in multiple sentences *"lymph nodes: numerous retroperitoneal and*

peri portal lymph nodes are evident. none of these demonstrate enhancement or washout to suggest a represent metastatic disease". Figures 6.2, 6.3, and 6.4, also provide examples in which outside sentence information makes a difference.

He reports drinking heavily up until his HCV diagnosis, but became abstinent since due to concern for the health of his liver.
In the past few months Mr. Xxxxxxx endorses sometimes getting confused, and has had others tell him that he is not acting like himself and not making sense.
Mr. Xxxxxxxx denies any other symptoms or complications related to his cirrhosis including nausea, vomiting, ascites, jaundice, fatigue, edema or bleeding tendency.

Figure 6.2: The meaning of overall evidence may change over multiple lines. Here we understand confusion to be related to *hepatic encephalopathy* only through the surrounding sentence context.

Lymph nodes: Prominent epiphrenic paracardiac node measures 0.8 x 0.9 cm (4/133).
Prominent porta hepatis nodes measure 1.2 cm (4/181) and 1.1 cm (4/186) .
Enlarged peripancreatic node measures 1.7 cm in short axis (4/176).
There are multiple scattered retroperitoneal and messentric lymph nodes which are not enlarged by size criteria.
Bones: No concerning lytic or blastic osseous lesions.
...
3. Enlarged peripancreatic and porta hepatis lymph nodes may be reactive or due to nodal metastases.

Figure 6.3: Example in which one sentence does not have enough information (first bolded sentence), but another similar sentence referring the same clinical phenomenon (second bolded sentence), has additional information.

HISTORY OF PRESENT ILLNESS

..

Over the last 3 weeks he has developed abdominal distention consistent with ascites.

...

PHYSICAL EXAMINATION

...

ABDOMEN: soft.

Moderate to large distention but not tight.

Positive large ascites on exam.

...

ASSESSMENT AND PLAN:

...

He is showing evidence of increasing ascites most likely related to some mild hepatic impairment in conjunction with extensive vascular invasion.

Figure 6.4: Without the text evidence in the Physical Examination section, the other mentions of *ascites* is much more vague. With less specific information, as with the first and last sentences, it makes more sense to assume the *mild* case.

6.4.2 Document level evaluation

Although in most cases, the annotation enrichment benefitted the classification, surprisingly, addition of more annotated data caused performance to drop in 4 categories, e.g. *ascites - moderate-severe*, *hepatic encephalopathy - mild*, *metastasis - yes-regional*, and *portal hypertension - no*. This is likely due to increasing in the amount of training variations, for the relatively lower frequency labels.

Our system involving sentence classification and resolving to the document level yielded higher results over the 1-, 2-, 3- gram document classification baselines. However this may be attributable to higher-level features (assertion classification) as well as enrichment of the corpus with more annotation. The document classification baseline did outperform for the *hepatic encephalopathy - mild* and *metastasis - yes-distal* baselines. Though, this was minor for the former category.

Of note, our annotations are sub-document, while our evaluation is at the document

level. So there may be some discrepancy regarding the actual category of a document when supposedly there are multiple marked categories of evidence within it. Moreover, given that there are multiple documents per patient, there was possibly some documents with missing annotations. From a comparison with document level inter-annotator agreement, several categories with below 0.85 F1 agreements were *ascites*, *extrahepatic invasion*, and *hepatic encephalopathy* with 0.60, 0.75, 0.80 F1 agreements, respectively.

6.5 Summary

In this chapter, we describe our sentence classification method to identifying and normalizing text evidence to stage parameter labels and values. Through this, we provide a general approach to find text evidence related to a certain phenotype given some annotations. As clinical domain experts cannot be expected to undertake the detailed and stringent annotation, this loose form of annotation gives enough detail (sub-document highlights) without being overwhelming (such as requiring detailed concept identification and sense normalization).

Though our approach was overall only marginally more effective for most categories than comparative set-ups, this methodology has a lot of potential to improve with little modifications. Mainly, with some further feature tuning, and most importantly, more expert annotated data, we can confidently expect additional performance boosts. Other possible improvements include additional feature normalization techniques.

Chapter 7

TUMOR CHARACTERISTICS EXTRACTION

This chapter describes a sub-system designed to extract entities and relations related for tumor information. These entities and relations are then assembled into templates, with some attribute identification. Then, finally, the templates are processed so that tumor related characteristics in a document such as size, number, and invasion of the organ are calculated. The work here is the subject of several of our papers [194], [196], [195].

7.1 Introduction

When biopsy or resection specimens are unavailable, clinicians may rely on non-invasive imaging studies to identify and characterize malignant tumors prior to planning treatment. This is often the case for HCC, since biopsies carry a significant danger of bleeding and tumor spread; further, tumor features on CT or MRI are considered highly sensitive and specific. As in other tumor diagnostic reports such as for histology and pathology, imaging reports describe crucial information related to a tumor, including location, number, size, and spread. This information is located throughout a report in a fragmented fashion, as shown in Figure 7.1, where diagnosis appears in the impressions section with summative information of previously mentioned lesions from the findings sections.

Moreover, previous measurements may become a confounding extraction problem because radiology reports often cite past readings. For example, Figure 7.2 shows a previous measurement mentioned.

In contrast to histologic or pathologic analyses which tests directly on specific corporal samples, cross-sectional imaging covers a large volume of tissue and therefore may pick up other non-cancerous entities. Further, imaging diagnostics may be prone to uncertainty

25: Focal lesions:
26: Total number: 5
27: Lesion 1: segment 8, 2.2 x 1.4cm , image 3/8, hyper enhancing with washout on delayed phase.
28: Lesion 2: segment 5, 2.0 x 1.8cm , image 3/25, hyper enhancing with washout on delayed phase.
29: Lesion 3: segment 4A, 1.8cm , image 3/7, hyper enhancing with washout on delayed phase.
30: Lesion 4: segment 8, 1.6 x 1.1cm , image 3/15, hyper enhancing with no definite washout.
31: Lesion 5: segment 6, 0.4cm , image 3/28, hyper enhancing with no definite washout.
..
35: Impression:
36: **3 focal lesions in segment 4A, 5 and 8 are hyper enhancing with washout on delayed phase, typical for HCC.**
37: **2 focal lesions in segment 8 and 6 are hyper enhancing with no definite washout on portal venous/ delayed phase suggestive of indeterminate nodules.**

Figure 7.1: Anaphoric and split antecedent tumor references in radiology reports

The previously visualized **mass** involving segment 5 and segment 6 has increased in size (cranial caudal measuring 11 mm, **previously 8.5 mm**) and now extends to involve segment 4.

Figure 7.2: Temporal tumor references

related to limitations of technology. Detected anomalies in imaging reports may be related to various cancer types, but could also be benign entities such as hemangiomas (tumors made of cells that line blood vessels), cysts (abnormal membranous sac containing fluid), pseudomasses (from imaging anomalies), or anatomic scarring. Table 7.1 shows examples in which tumor references are determined as malignant, benign, or indeterminate.

Tumor status	Example Passage
Malignant	1.9 x 1.8 cm hyperenhancing mass on the arterial phase with enhancing pseudocapsule, corresponding washout on portal venous phase as well as T2 hyperintensity and restricted diffusion, characteristic of HCC.
Benign	There are multiple scattered hepatic hypodensities that exhibit no enhancement and likely represent cysts.
Indeterminate	In segment 4a, there is a stable hypovascular lesion which is indeterminate and could represent a regenerative nodule. Would recommend MRI with Eovist specifically to further evaluate this lesion.

Table 7.1: Examples of tumor statuses

Reference resolution is the task of identifying expressions in text that refer the same real-world entity. In order to thread together the fragmented information we must be able to disambiguate which tumor findings respond to which. For example, consider the following excerpt from a radiology report:

22: Within hepatic segment II/III there is 14 x 9 mm **hypervascular lesion**$^{(1)}$ is isotense to liver parenchyma on portal venous phase ...

24: This **lesion**$^{(1)}$ is suspicious for hepatocellular carcinoma.

.....

41: Impression:

42: **Hypervascular lesion**$^{(1)}$ in hepatic segment II/III with imaging features suspicious for hepatocellular carcinoma.

Figure 7.3: Radiology report excerpt

The three mentions of the *hypervascular lesion* appear in separate sentences, yet the reader will naturally group them as one real world entity.

The state-of-the-art in reference resolution in the general domain is still challenging; this condition is even more dire for the clinical domain, in which there is a relative scarcity of annotated corpora. Furthermore, in the clinical domain, there are still well-known unsolved text processing problems such as ill-formed, ungrammatical, telegraphic, semi-structured,

abbreviation-ridden narratives.

In the following sections, we describe our annotation and system-building for: (a) tumor information extraction, to capture and structure the scattered tumor-related information in reports, (b) tumor reference resolution, to condense the same information together, and (c) tumor characteristics, to use the structured and reference resolved tumor information to find staging-relevant characteristics.

7.2 Dataset

For deeper tumor information annotation, we randomly selected 101 radiology reports from the previously divided 160 HCC patient training set. Here we have 3 distinct levels of annotation: (a) tumor-related finding event annotations, (b) tumor reference resolution and (c) tumor characteristics. After describing our annotation, we show inter annotator agreement for all three levels.

7.3 Annotation description

7.3.1 Template annotation

Templates are pre-determined simplified representations of knowledge, here comprised of entities, spans of text with assigned label names, and relations, directed links between entities. In our task, entities captured anatomic entities, tumor references, sizes, number, cancer diagnosis, whereas relations ensured that the proper descriptions linked to the items they characterized. Our template schema was designed by a biomedical informatics graduate student and a medical student. Figure 7.2 includes example sentences annotated with entities and relations. We used Brat [163] a web-based annotation tool, for our annotation software.

Our entities had the following types: (1) Anatomy: anatomic locations in the human body (e.g., segment 5 or left lobe) with attributes (Liver, NonLiver), (2) Measurement: quantitative size in the text (e.g., 2.2 x 2.0 cm), (3) Negation: indicator to some negation of a tumor reference (e.g., no) (4) Tumor count: number of tumor references (e.g., two or

multiple), (5) Tumor reference: a radiologic artifact that may reference a tumor (e.g., lesion or focal density), and (6) Tumorhood evidence: diagnostic information regarding the tumor (e.g., characterisitcs of HCC, indeterminate, suggestive of cyst) with attributes (isCancer, isBenign, inDeterminate).

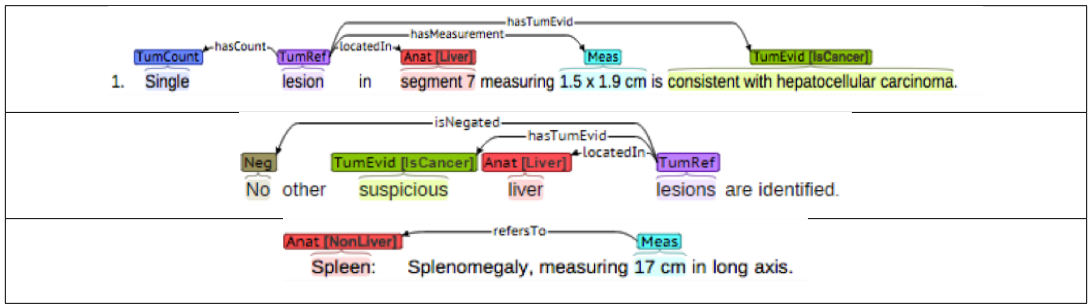


Table 7.2: Examples of radiology reports annotated with entities and relations.

Our relations were defined as directed links between the two entity types, often with either a tumor reference (preferred) or a measurement as the source, or the head, of the directed relation. They are described as follows: (1) hasCount: relation between a tumor reference to a tumor count, (2) isNegated: a negation cue, starting from the tumor reference to the negation entity, (3) locatedIn: marks in which anatomy a tumor reference or measurement was found, (4) hasMeasurement: present tense relation between a tumor reference to a measurement, (5) hadMeasurement: past tense relation between a tumor reference or measurement, to a measurement, (6) hasTumEvid: relates a tumor reference or measurement to a tumorhood evidence, (7) refersTo: relates a measurement to an anatomy indicating a measurement of anatomy. Templates were constructed by collating all connected entities and relations.

The annotation approach in this part sought to maximize information while minimizing annotation workload, as thorough entity and relation annotation is very time-consuming. Therefore we made a few important high-level annotation decisions: (1) only the "findings" and "impressions" section of the reports were annotated, (2) we annotated either a tumor

reference or a measurements (the starting point of the relation or the head) in all available lines, but only annotated other entities if they were related to our tumor reference or measurements or if it appeared in a line with annotations, (3) we annotated radiographic evidence of tumorhood evidence, e.g. "hypervascular with washout," and (4) we added extra relations from a measurement to an anatomy when they referred to different locations. Relation attachments over multiple lines were allowed, though we did not mark for co-referring information and each tumor reference was treated separately.

We decided on (1) because, we found that the "findings" and "impressions" sections comprehensively harbored the radiologic information in the report. Other parts of reports had comparatively more unimportant tumor information, e.g. in the "indication" section, "please determine size and location of tumor."

Our reason for designating tumor references and measurements as heads, in decision (2), was part of our strategy to maximize annotation simplicity. For example, we avoided a lot of excess annotation by not allowing pronouns such as *"this"*, *"these"*, *"the largest"* as a tumor reference, e.g. Figure 7.4 line 21. Measurements were allowed as heads because in absence of a nearby tumor reference, a size was the most reliable indicator of tumor information, e.g. "1. Segment VII: 2.6 x 2.4 cm, hyper enhancing with washout." By only annotating entities related to these heads, we avoided lines without any information of interest. We annotated other entities within a line, not necessarily related to an event, to provide negative example cases. For example, an anatomy entity may only be near a tumor without actually having been invaded, e.g. Figure 7.4 line 23, or instead a measurement may be measuring an anatomy entity instead, e.g. Figure 7.4.

The full annotation guidelines are provided in Appendix D.

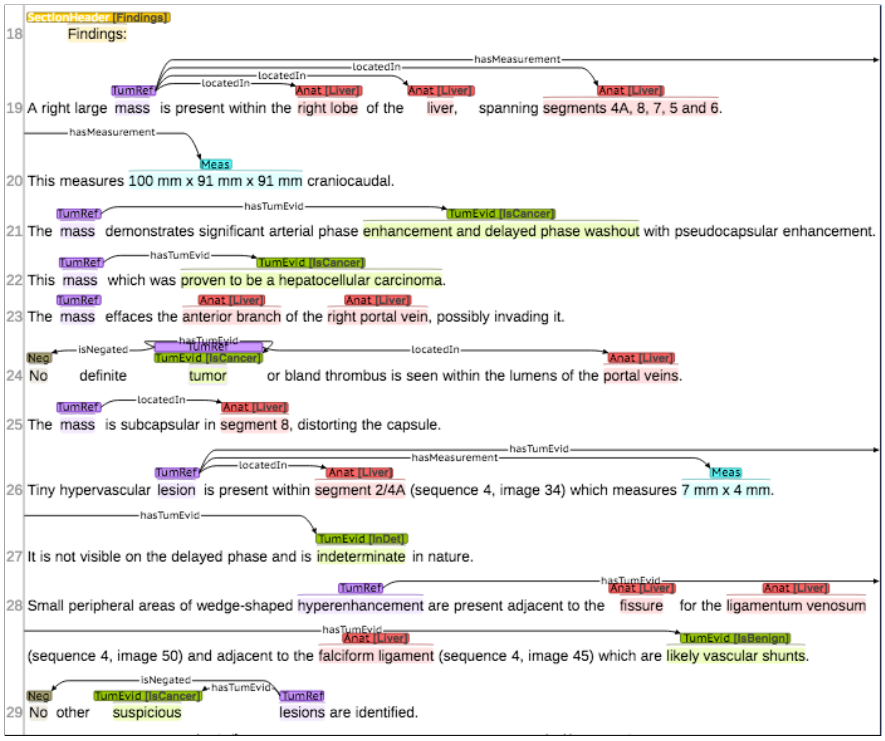


Figure 7.4: Marked up findings section of radiology report

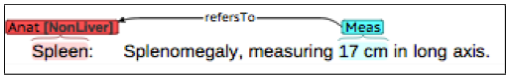


Figure 7.5: Measurement finding

7.3.2 Reference resolution annotation

Reference resolution annotation was based on two types of relations for tumor-related templates (TumorSingleton and Tumor templates) heads:

- **coreference** - equivalence relations between mentions, e.g. Figure 7.3

- **particularization** - a directed parent-child relation in which the first argument represents a set of tumor reference(s) that contains the second argument tumor reference(s),

e.g. Figure 7.6, where **Lesions**$^{(1)}$ is a reference that represents a set of items, particularized by **Lesion**$^{(2)}$, **Lesion**$^{(3)}$, and **Lesion**$^{(4)}$.

21: **Lesions**$^{(1)}$ consistent with HCC given their enhancement characteristics:
22: 1. **Lesion**$^{(2)}$ in segment 8 measuring 1.2 x 1.3 cm previously measuring 1.3 x 1.7 cm
23: 2. **Lesion**$^{(3)}$ in segment 4A measuring 1.2 x 0.9 cm not clearly seen on the previous study.
24: 3. **Lesion**$^{(4)}$ on the border between segment 4A and 8 measuring 2.9 x 3.5 cm previously measured 2.5 x 2.6 cm

Figure 7.6: Example of one reference and its particularizations

Pronominal cases, e.g. "it", "they", and "these" are unmarked.

For our annotation software, we again used brat, a web-based, annotation software for our reference resolution annotation. Since the number of coreference and particularization relations would visually render the annotations to be highly cluttered, we augmented the software to output text information regarding the clusters and particularizations annotated whenever the user selected a "show references" button, as shown in Figure 7.7.

Full annotation guidelines are provided in Appendix E.

Figure 7.7: Brat annotation with augmentations.

7.3.3 Tumor characteristics annotation

Additionally, we are interested in the real-world task of automatically categorizing patients into liver cancer staging phenotypes, to which reference resolution is only an intermediate step. To this end, we are motivated to identify three summative tumor characteristic variables important for staging: (1) largest size of a malignant tumor, (2) tumor counts, and (3)

whether 50% of the liver organ is invaded by tumors. These are relevant for three liver cancer staging algorithms: AJCC (American Joint Committee on Cancer), BCLC (Barcelona Cancer of the Liver Clinic), and CLIP (Cancer of the Liver Italian Program). Since these variables require aggregate knowledge of tumor-related attributes, an end-to-end evaluation, incorporating reference resolution, using these staging variables would provide a worthy perspective.

Tumor characteristics annotation included a spreadsheet that referenced each document name and (1) the number of tumor counts by type (benign, indeterminate, unknown, and malignant), (2) the largest size for malignant tumors, and (3) whether or not more than 50% of the liver is invaded. We decided to mark inequalities, as at times the documents do not in fact give a clear number. Meanwhile, we also collected information regarding the various tumor counts for each of the Findings and Impression sections, as well as the entire document. A sample is of this is shown in Figure 7.8.

	A	B	C	D	E	F	G	H
1	FileName	Section	Malignant	Indet	Benign	Unk	>50%	largestmalignant
2	[redacted].radiology.0.CT.A	Findings				3		
3		Impression	1		2			
4		Whole	1		2		NO	6.3 x 7.1 x 6.4 cm
5	[redacted].radiology.1.MRI	Findings	1		5			
6		Impression	1		>1			
7		Whole	1		5		NO	6.3 x 7.1 x 6.1 cm

Figure 7.8: Tumor characteristics annotation

Because the measurement for (3) is not readily quantifiable given the information in reports, we use a series of expert-created guidelines to determine the criteria for (3), as outlined in the Figure 7.9 below. The full annotation guidelines are given in Appendix F.

Tumor extension for malignant tumors are considered over 50% if ANY of the following conditions are met:

1. Tumor $\geq$ 10 cm
2. >4 segments involved
3. "majority" of "right lobe" involved
4. All right lobe segments involved
5. Entire left lobe plus some right lobe involved
6. Some description to suggest much of liver involved, e.g. massive very extensive

Figure 7.9: Logic for >50% of liver invaded

7.4 Evaluation

7.4.1 Template evaluation

Our evaluation for templates was carried out at three levels: (a) entity, (b) relation, and (c) template levels. We used precision, recall, and F-1 measure, defined previously, as our inter-annotator agreement measure (where one annotator was held as the gold standard).

Two entities were considered matching if they had the same label, attribute (if appropriate), and document offset text spans. Relations were considered matching if both of its entities matched, and the relation types both matched. Two templates were considered matching if all its entities and relations matched that of the other template. Partial entity match allowed document to be counted as matching if document text spans at least overlapped and their labels matched. Similarly relation partial matching was defined on whether the two pair of entities partially matched and if the relation type was correct. Partial template match was defined by whether all entities and relations were partially matched.

7.4.2 Reference resolution evaluation

Coreference evaluations were based on MUC [179], B-cubed [24], and CEAF [101] F1 scores. Reference resolution for particularization relations were measured using relations F1 score. Tumor characteristics were evaluated based on the label assigned to a specific document,

document section, and tumor characteristics variable combination. In the next sections, we detail the specific instance, precision, and recall definitions for each metric.

Coreference evaluation

In the following sections, we provide formulaic definitions of the coreference evaluation metrics.

MUC metric

The MUC (Message Understanding Conference) metric [179] measures the minimum amount of of correct links necessary to transform the key's (the gold) equivalence classes, partitioned by the responses's (the system) equivalence classes, back to its original equivalence classes. First the relative partition of the key's classes are identified with respect to the system's by intersecting the key's sets with the system sets that over lap with the key. For example, if $Q_i = \{A, B, C, D\}$ is one equivalence class in the key and $S = \{\{A, B\}\}$ is the system's set of equivalence classes then the partition of Q_i is $p(Q_i) = \{\{A, B\}\{C\}\{D\}\}$. After this separation, the number of links to connect back into the key's original classes can be quantified.

Recall is defined by:

$$R = \frac{\sum(|Q_i| - |p(Q_i)|)}{\sum(|Q_i| - 1)} \tag{7.1}$$

where $|p(Q_i)|$ is the cardinality of partitions in the key's i-th equivalence class, with respect to the system, and the denominator represents the minimal number of correct links necessary possible. Precision is calculated by reversing the system and the key answer sets; meanwhile F1 is computed as before in Equation 3.3.

B-cubed metric

The B-cubed metric [24] we employ here is the uniformly weighted element-wise measurement of recall for the classes containing each element.

Recall is defined by:

$$R = \sum_{e \epsilon \{\bigcup_i Q_i\}}^{n} w_i * \frac{|Q_e \bigcap S_e|}{|Q_e|} \tag{7.2}$$

where e is a mention in the key, n is the total number of mentions in the key, w_i is set to $w_i = \frac{1}{n}$ and Q_e is the equivalence class in the key that contains e and S_e is the system class that contains e.

CEAF metric

CEAF [101] precision and recall are defined based on an optimal one-to-one alignment of the system and key equivalence classes, based on a similarity function ϕ, where extraneous classes are discarded. Thus for a set of key equivalence classes $Q(d) = \{Qi : i = 1, 2, ..., |Q(d)|\}$, system equivalence classes $S(d) = \{S_i : i = 1, 2, ..., |S(d)|\}$, let G_m be the set of one-to-one maps between Q and S for $m = min\{|Q|, |S|\}$ classes, where we denote g as a function mapping Q to S. We now define the best mapping function g^*:

$$g^* = \arg\max_{g \epsilon G_m} \sum_{Q} \Phi(g) \tag{7.3}$$

where

$$\Phi(g) = \sum_{Q \epsilon Q_m} \phi(Q, g(Q)) \tag{7.4}$$

i.e. the best alignment is the one that maximizes the sum of the similarity function for the matches. Finally, precision and recall defined:

$$P = \frac{\Phi(g^*)}{\sum_i \phi(S_i, S_i)} \tag{7.5}$$

$$R = \frac{\Phi(g^*)}{\sum_i \phi(Q_i, Q_i)} \tag{7.6}$$

where F1 is defined again as in Equation 3.3. Precision is the sum of similarities between the best alignment of G to S, normalized by the sum of similarities of each item in S to itself. Recall is calculated in the similar way except using Q.

We used the relative measure of similarity $\phi = \frac{2|R \bigcap S|}{|R|+|S|}$, which is the ϕ_4 in the original definition paper. Meanwhile, the optimal assignment problem based on a given definition of ϕ may be calculated using the Kuhn-Munkres Algorithm [92].

Particularization relation evaluation

Particularization relations are labeled directed connection between two mentions. A correct relation requires the correct label (in this case *particularization*) and the correct identification of the first mention to the second mention. Precision and recall are the same as those defined in Equations 3.1 and 3.2

7.4.3 Tumor characteristics evaluation

Tumor characteristics evaluation was based on the correct label for each document and tumor characteristic variable: (1) tumor counts for benign, indeterminate, malignant, and unknown and (2) largest size for malignant tumors, and (3) whether $> 50\%$ of liver is invaded. Although we also labelled tumor counts for specific sections in a document (Findings and Impression) we only evaluate values for the entire document in this work.

We also introduced a relaxed match motivated by our specific extraction needs for liver cancer staging for AJCC, BCLC, and CLIP liver cancer algorithms. Based on staging criteria, there were certain critical thresholds that affect the score. For example, given malignant

tumor measurements all under 3cm, it does not make a difference if our algorithm cannot distinguish between 2 or 3 tumors, or if it cannot distinguish between 5 and 10 tumors; however, if the system cannot distinguish between a single tumor and multiple tumors, the cancer stage is changed drastically. The case is the same for certain sizes. Thus, our relaxed match measures based on the bins discretized from the critical values of our staging algorithms. The bin thresholds are the same as those summarizing our tumor characteristics annotation distribution in Table 7.7.

7.5 Inter-annotator agreement

7.5.1 Template agreement

After agreeing on a final annotation schema, the biomedical informatics graduate student and medical student tested inter-annotator agreement on a set of randomly selected 31 radiology documents. At the first annotator meeting, agreement was scored at 0.84, 0.73, 0.54 F1 for entities, relations, and templates, respectively. After refining guidelines further, the annotators re-annotated on the same set. The final entity, relation, and template agreements improved to 0.88, 0.78, 0.61 F1, with partial scores of 0.93, 0.90, and 0.70 F1. The full breakdown is shown in Tables 4, 5, and 6. Reported templates are broken down into categories by their constituent relations for finer-grained analysis. For example, if refersTo was a relation in the template, it is categorized as an AnatomyMeasure template; if the template has an isNegated relation, it is a Negative template. Singletons were all templates with a single entity. The remaining templates were categorized as tumor events.

The medical student annotator annotated the remaining 70 reports of the corpus. The total number of entities, relations, and templates for the 101 radiology report corpus were 3211, 2283 and 1006, respectively.

Label	TP	FP	FN	P	R	F
Anatomy	316	32	49	0.91	0.87	0.89
Measurement	159	1	2	0.99	0.98	0.99
Negation	23	0	3	1.00	0.88	0.94
Tumor count	65	2	5	0.97	0.93	0.95
Tumor reference	245	7	14	0.97	0.94	0.96
Tumorhood evidence	159	14	24	0.92	0.87	0.89
ALL	967	56	97	0.95	0.91	0.93

Table 7.3: Entity agreements (partial)

Label	TP	FP	FN	P	R	F
hadMeasurement	15	0	2	1.00	0.88	0.94
hasCount	64	3	6	0.96	0.91	0.93
hasMeasurement	94	6	8	0.94	0.92	0.93
hasTumEvid	155	23	27	0.87	0.85	0.86
isNegated	24	0	3	1.00	0.89	0.94
locatedIn	279	25	39	0.92	0.88	0.90
refersTo	24	4	7	0.86	0.77	0.81
ALL	655	61	92	0.92	0.88	0.90

Table 7.4: Relation agreement (partial)

Label	TP	FP	FN	P	R	F
AnatomyMeas	20	5	8	0.80	0.71	0.76
Negative	17	7	10	0.71	0.63	0.67
Singleton	34	23	24	0.60	0.59	0.59
TumorEvent	161	57	62	0.74	0.72	0.73
ALL	232	92	104	0.72	0.69	0.70

Table 7.5: Template agreement (partial)

7.5.2 Reference resolution agreement

Annotation for both reference resolution and tumor characteristics were performed together on all 101 reports. 20 reports were used to measure inter-annotator agreement between a medical student and a biomedical informatics graduate student. The rest of the corpus was single-annotated by the biomedical informatics student.

We measured inter-annotator agreement for coreference in terms of MUC [179], B-cubed [24], and CEAF [101] for tumor-related template heads. The agreements were at 0.956, 0.969 and 0.916 F1, respectively. For annotator 2, there were 20 clusters (no singletons), 149 clusters (with singletons), with the average size of 2.7 entities per cluster. The cluster-normalized F1 measure for particularization relations was at 0.837.

Lesion 3: Multiple satellite **lesions** for example, segment 4 measures 2 cm
This **region** is heterogeneously hyper intense with numerous **regions** of focal washout
Mild increase in size of segment 6 hyper vascular **focus** from 1.2 to 2.2 cm with central **area** of nodular hyper vascular focus

Figure 7.10: Examples of coreference relations that can be mistaken as particularizations

Some ambiguities did occur between coreference and particularization, which accounted for some of the disparity in inter-annotator agreement. Mainly, as given in the examples of Figure 7.10, some mentions are singular but may be equivalent to the plural form of another mention.

The final corpus has 210 clusters (no singletons), 479 cluster (with singletons), with an average of 2.60 mentions per cluster. Inferred particularization relations amounted to 573. The average and median number of sentences between the closest pairwise mentions in the same cluster are 10 and 6 sentences respectively. The large difference between mean and median suggests the existence of some very long-distance coreference relations. The mean proportion of mentions that are exact matches in a cluster is 37%, 43% if normalized for capi-

talizations. The average proportion of mentions found in the Findings and Impression section per cluster respectively, are 57% and 38%. The proportion of particularization relations that connect mentions in different sections is 47%.

7.5.3 Tumor characteristics agreement

The inter-annotator agreement is shown in Table 7.6. In general, results were high but there were some ambiguities in the annotation that led to some disagreements.

Label	**TP**	**F1**	**F1 (relaxed)**
Benign	17	0.85	0.95
Indet	18	0.90	0.95
Malignant	17	0.85	0.95
Unk	20	1.0	1.0
LargestSize	17	0.85	0.95
>50%	20	1.00	1.00

Table 7.6: Tumor characteristics inter-annotator agreement

Tumor characteristics annotation were subject to various gray areas. For example for tumor counts, at times there were many ambiguous statements regarding the numbers. One example of this is in the case of conjunctions, several examples of which are shown in Figure 7.11.

5-6-mm **segment 6/7 and 5** hyper vascular **foci**
A small enhancing **area** seen along the lateral aspect of **segment 6, segment 7, and segment 4b/5**

Figure 7.11: Conjunction ambiguities.

The first statement can imply either one 5-6-mm focus in segment 6/7 and one in segment 5, or one 5-6-mm touching segments 6/7 and segment 5; or multiple 5-6-mm foci in the areas

of segment 6/7 and 5. Similarly, *"enhancing area"*, in the latter statement, may be one large area inside segment 6 and 7 (which are adjacent) or separate areas in 6 and 7.

Furthermore, to gather the most accurate number bounds for tumor counts, it was at times necessary to add multiple inequalities, e.g. if there are multiple (but unspecified or only partially specified) lesions in separate areas, which added to the cognitive load.

Annotating the largest size was the least controversial, though this too has some ambiguity. For example the same lesion may have two different measurements in a single report. For example in the Findings section, the largest size might be *"2.5cm"* but the same lesion is later referred to as *"2.4cm"* in the Impression section. In another example, one measurement mentioned may be specific, e.g. *"6.3 x 6.1 x 9.8 cm"*, and but later rounded, e.g. *"6 x 6 x 10 cm"*. Moreover, the amount of text and lengths of the documents, including many possible repetitions, could make it difficult to locate the best representable sizes.

The >50% variable was at times still unclear, even with the guideline. Analyzing Figure 7.12 as an example, it is obvious that there are multiple tumors in both the right lobe and in the left lobe; however only 3 segments are specifically mentioned. It is therefore not clear if the unmentioned numerous tumors may be all over the liver or only in those specific parts.

Focal lesions:

Multifocal HCC with hypervascular washout

In the **right hepatic lobe**, the largest is in **segment 5/6**.

It measures 4.5 x 4.4 cm image 21/7

In the **left hepatic lobe**, the largest is in **segment 4a**.

It measures 3.6 x 3.5 cm

Impression:

Multifocal HCC with malignant vascular invasion of the right portal vein and IVC

Figure 7.12: Ambiguity in tumor invasion area.

The distribution for the full corpus of the tumor characteristics annotation, binned along critical threshholds, is shown in Table 7.7.

Annotation categories		
Tumor counts	Number	Freq.
Benign	0	69
	1	8
	2-3	10
	> 3	8
	[2-3, > 3]	6
Indet	0	62
	1	19
	2-3	9
	> 3	4
	[2-3, > 3]	7
Malig	0	3.0
	1	54
	2-3	25
	> 3	13
	[2-3, > 3]	6
Unk	0	89
	1	5
	2-3	2
	> 3	2
	[2-3, > 3]	2.0
	[0, 1, 2-3, > 3]	1.0
Largest size	Size (cm)	Freq.
	[0,3)	43
	[3,5]	26
	(5-10)	17
	[10,)	10
	n/a	5
>50%	Label	Freq.
	n/a	4
	no	83
	yes	14

Table 7.7: Tumor characteristics annotation distributions, binned according to crucial staging values. The value of "[0, 1, 2-3, > 3]" was for a case in which the full number of lesions was given, but it was unclear how many were malignant, resulting in an unknown lesion inequality after subtraction < 5.

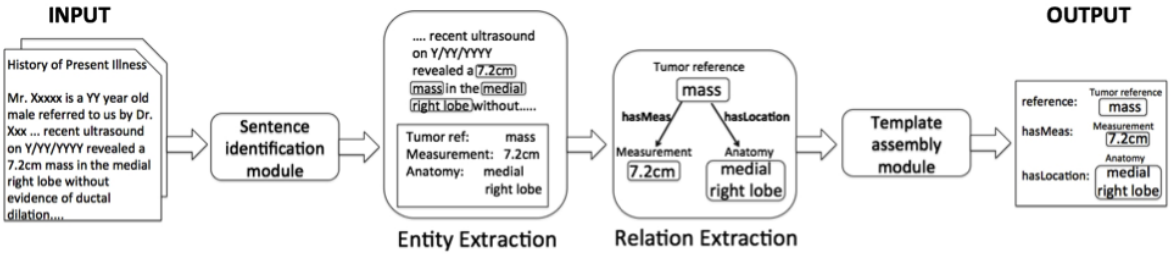


Figure 7.13: Pipeline for entity and relation extraction

7.6 Tumor template extraction : entity and relation extraction

Figure 7.13 presents the overall system architecture for template extraction. A sentence identification module identified sentences of interest. Entity types were extracted from the isolated sentences using regular expression for the measurement entity and CRFs on the remaining others. Relations were identified using a direct classification of enumerated pairwise entity to entity candidate relations. Afterwards, templates were assembled by traversing the graph of connected entities and relations. Evaluations were the same as those used for inter-annotator agreements.

7.6.1 Related work

Cancer information extraction

The challenge of tumor extraction is not new and there is much to be learned from previous work. Rule-based systems for these tasks typically involved a dictionary look-up, context and negation checking, and heuristic algorithms to structure results. Scores range widely between systems as well as between distinct variables within a single system, depending heavily on the selection of extraction variables. Coden et al [41] focused on finding hierarchical concepts such as anatomical site, grade value, date, primary tumor, etc., from pathology reports and organized results into structured classes, achieving F1 scores ranging for 0.65 to 1.0. Ashish et al [23] trained and tested on pathology reports from the University of

California Irvine data warehouse and looked for structured classes such as TNM stage, capsule invasion, lymph invasion, chronic inflammation, and vascular invasion, with per field F1 performances ranging from 0.78 to 1.0. Ping et al [135] used regular expressions and structured entities extraction using heuristic algorithms for liver cancer information, with 0.92-0.996 F1 score. The machine learning equivalent of these works used statistical methods. Hassanpour and Langlotz [74] experimented with named entity recognition in CT radiology reports, comparing dictionary methods, conditional random fields (CRFs), and maxiumum entropy markov models (MEMMs) with a performance of 0.85 F1. Ou and Patrick [127] used CRF classification to extracting cancer-related entities, such as diagnosis, metastasis, site, size, and specimen type, from processed primary cutaneous melanoma pathology reports. Afterwards, entities were populated into structured reports using rules. F1 performance for populating fields was at 0.85. Roberts et al [142] attached anatomical locations to known findings in radiology reports through statistical classifications of words around the finding and with various surface, morphological, and dependency features.

Radiology report parsing

The general task of parsing reports have been explored since the early days of biomedical informatics, with a heavy emphasis on comprehensive linguistic annotation that subsequently mapped to a separate parallel domain knowledge base. These have contributed to systems such as MedLee and others [66]. Continuing in this tradition, Taira et al [173] detailed their system that includes deep linguistic annotation with dependency parses fortified with a detailed radiology report domain ontology. Their strategy started with identifying concepts using custom dictionaries, then dependency parsing entities using statistical methods in their parser module. Once parsed, relations from their radiology ontology were constructed using their semantic interpreter module, which either used rule-based logic or a statistical maximum entropy classifier. Finally, their frame constructor bundled together their concepts and relations. They reported parsing performance of 87% recall and 88% precision. Meanwhile, their conversion of dependency parses into relations were evaluated at 79% and 87% recall

and precision, respectively.

7.6.2 Preprocessing

Radiology reports were processed to remove excess white spaces and blank lines using report-specific heuristics. Sentences were identified using NLTK punkt module [100]. Only sentences belonging to the Findings and Impressions sections, as tagged by our in house section chunker [174] were kept as per our annotation guidelines.

7.6.3 Sentence identification

To avoid classifying sentences with no annotations, we first selected sentences of interest. Based on the analysis of our corpora, we found that 90% of relations were from entities on the same line, and around 7% were from entities connected to the entities on the next line. To identify these sentences, we mimicked the annotation strategy of first finding the tumor reference or measurement before marking other entities. Sentences of interest on the first line (S1) were identified using regular expressions on measurement values, e.g. (\\d+) cm (a number before a cm word), and a word list of radiographic tumor reference terms, listed in Figure 7.14, created from the top unigrams accounting for 90% by frequency for tumor references. S1 sentences, along with the sentence following it (S2) sentences, were passed to subsequent steps. This resulted in a sentence identification recall and precision of 94% and 69%.

7.6.4 Entity extraction

Entities were extracted from the sentences identified using one of two strategies: regular expression lookup and sequential label classification. The original regular expressions used to identify measurement values in sentence identification were taken as the measurement entities. For the remaining entities, anatomy, negation, tumor reference, tumorhood evidence entities, we used CRFs classified using CRFSuite [126].

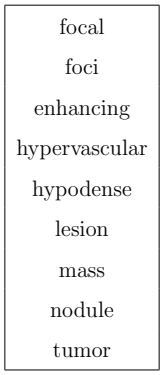


Figure 7.14: Sentence word list

We created CRF features by identifying several base features, then generatively creating the final more complex features by tuning several variables: window-size, n-gram numbers, and tag sets (for entity features) {BIOE, BIO, IOE, IO}, as illustrated in Figure 7.15. For example, suppose our base feature is unigrams. Then if we choose a window size of ±2, and n-grams of 1 and 2, then the final features would be all unigrams and bigrams within ±2 words of a word. For base features that may span over multiple words, such as tagged UMLS concepts, we additionally experimented with different tag sets. Table 8 gives a more detailed description of our base features. We also implemented two augmenting parameters, which replaces the UMLS feature with a more general term if a concept id is part of the specified list. These two lists were for liver anatomic parts and carcinoma concepts. For example, if any concept ID part of the carcinoma list is found, instead of its specific preferred name used as a feature, it will be <CARCINOMA>. Liver concepts were identified from taking all liver anatomic subparts as specified in the Foundation Model of Human anatomy. Carcinoma concepts were generated from taking sub concepts of C3263 (Neoplasm By Site) from the National Cancer Institute thesaurus. S1 and S2 sentences were trained separately.

Sentences were tokenized, tagged for parts-of-speech, dependency parsed using ClearNLP [3]. UMLS features were extracted using MetaMap [22], with word sense disambiguation turned on. During experimentation, we optimized for the feature parameters, as well as the

optimal CRF tag set for the learned labels. When tag labels overlapped, we merged tags. For example, in Figure 7, nodules would have the tag B-TumRef_I-TumEvid [isCancer].

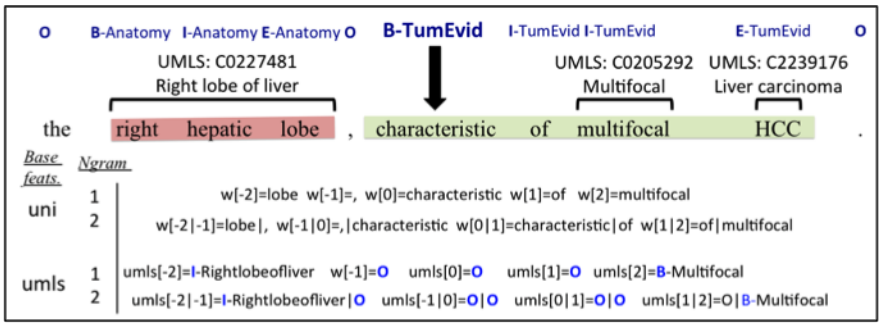


Figure 7.15: CRF features description

Feature	Feature Description
UNIGRAM	Unigram (case-sensitive)
LEMMA	Lemma (lower-cased)
PARTS-OF-SPEECH	Parts of speech
LABELLED	Label of the relation from the word to its head
DEPENDENCY	combined with the lemma of its head word
DEPENDENCY HEIGHT	Height of a word in its dependency tree
UMLS CONCEPT	{BIOE}-tagged normalized name
UMLS SEMANTIC TYPE	{BIOE}-tagged UMLS semantic type

Table 7.8: CRF features description

7.6.5 Relation extraction

Once sentences were identified with entities, they were run through a relation classifier. All possible pairwise relations between entities in S1 and corresponding S2 sentences were enumerated and classified using several machine learning algorithms. Given two entities, the direction was determined based on the entity types, e.g. the tumor reference is always the head, or the first-appearing measurement if no tumor reference is found. We experimented

with a c4.5 decision tree, maximum entropy, and a support vector machine (SVM) classifier, implemented through MALLET [110] and LibSVM [36] with default parameters. We report the classifier with the best performances. Our features were related to the entities involved, the dependency paths between them, and the words around them. They are described in detail in Table 7.16. We tuned two variables in our experiments: window size for the SURRWORDS feature and the machine learning classifier.

We report our results compared to a simple baseline. The simple baseline takes the S1 and its S2 sentence and creates a template by attaching all entities to first occurring tumor reference, or measurement if tumor references are not available. If more than one relation is possible according to our annotation guidelines, we put the highest frequency relation and do not attach a relation if no relation is possible.

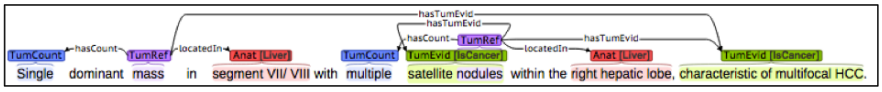


Figure 7.16: A single sentence can have multiple tumor reference subjects, with overlapping entities

Feature	Feature Description
CLOSESTREF	1 if the head entity is closest left or right tumor reference, e.g. (closestLeftRef:0, closestRightRef:1)
DIFFLINES	1 if two entities are on the same line, (e.g. sameLine:1)
ENTNUM	Number of each type of entity in corresponding line (e.g. num-l1[Anatomy]:2, num-l1[TumCount]:2)
ENTWORDS	1 for every word inside an entity, represented by its lemma (e.g. en1-nodule:1, ent2-segment:1)
POSSIBLELABELS	1 if relation label type is a possible between two entities (e.g. candidateLabel-locatedIn:1)
ONLYPOSSIBLEHEAD	1 if head entity is the only tumor reference or measurement in the lines being considered (e.g. onlyHead:0)
SHORTESTPATH	The shortest path distance between two entities through the dependency tree (e.g. minPath:3)
SHORTESTPATH.HEADS	Within the shortest path, 1 if words within path have the labels of tumor reference, measurement, or the second entity label (e.g. minPath[tumorref]:1)
SUBTREE	Minimum distance between head entity to another second entity of the same label type in its dependency subtree (not including the second entity) (e.g. subTreeNextCand[samelabel]:1)
SURRWORDS	1 for every word within the word window of the entities, (e.g. uni-ent1[-2]=multiple:1, uni-ent2[1]=with:1)

Table 7.9: Relation features description

7.6.6 Results

Table 10 and 11 shows entity extraction performances for exact and partial match, respectively, consolidated by label. Our final feature configurations included a window size of ± 1 word, 1-grams, and BIO tagging for both features and labels. Our higher performing entities, the measurement and tumor reference, were expected given the rule-based nature of measurement extraction and the strategy of sentence classification. Precision was high across all entities, which is perhaps a result of our tagging scheme and overlapping entities, which combines to very specific tags. Our entity overall extraction performance 0.87 F1 was lower compared to inter-annotator agreement 0.93 F1, which is often considered the upper bound for a task. Specifically, negation, tumor count, and tumorhood evidence were at considerably lower performances with 0.08, 0.12, and 0.19 F1 difference.

Label	TP	FP	FN	P	R	F
Anatomy	789	103	254	0.88	0.76	0.82
Measurement	472	9	17	0.98	0.97	0.97
Negation	59	6	14	0.91	0.81	0.86
Tumor count	126	6	48	0.95	0.72	0.82
Tumor reference	678	64	124	0.91	0.85	0.88
Tumorhood evidence	315	85	315	0.79	0.50	0.61
ALL	2439	273	772	0.90	0.76	0.82

Table 7.10: Entity extraction results (exact)

Label	TP	FP	FN	P	R	F
Anatomy	828	65	215	0.93	0.79	0.86
Measurement	480	1	9	1.00	0.98	0.99
Negation	59	6	14	0.91	0.81	0.86
Tumor count	127	5	47	0.96	0.73	0.83
Tumor reference	714	25	88	0.97	0.89	0.93
Tumorhood evidence	359	40	271	0.90	0.57	0.70
ALL	2567	142	644	0.95	0.80	0.87

Table 7.11: Entity extraction results (partial)

Our feature configurations for relation extraction was a window size ±3 words from each entity, using a maximum entropy classifier. The system relation and template extraction performance are shown in Table 12 and 13, with both gold and system entities. Even with gold entities, the hadMeasurement and the refersTo relations were comparatively low-performing at 0.35 and 0.67 F1 scores, respectively. Given gold entities, our system reached 0.89 F1 for relation extraction and 0.64 for tumor extraction, with 0.72 F1 for the TumorEvent template subcategory. Meanwhile, when using system entities both relation and template extraction suffered more than 10% degradation, suggesting that improvements in the entity extraction upstream task will lead to improvements in the overall system.

	Baseline		System	
Entities	**Gold**	**System**	**Gold**	**System**
hadMeasurement	0.00	0.00	0.35	0.36
hasCount	0.90	0.76	0.95	0.79
hasMeasurement	0.85	0.81	0.89	0.85
hasTumEvid	0.86	0.61	0.89	0.63
isNegated	0.97	0.86	0.98	0.87
locatedIn	0.82	0.71	0.89	0.77
refersTo	0.00	0.00	0.67	0.63
ALL	0.83	0.69	0.89	0.74

Table 7.12: Relation extraction results (partial)

	Baseline		System	
Entities	**Gold**	**System**	**Gold**	**System**
AnatomyMeas	0.00	0.00	0.49	0.26
Negative	0.84	0.57	0.82	0.57
Singleton	0.25	0.25	0.35	0.32
TumorEvent	0.69	0.42	0.72	0.44
ALL	0.60	0.38	0.64	0.42

Table 7.13: Template extraction results (partial)

7.6.7 Discussion

A significant hurdle for our entity extraction task was that, different from traditional entity extraction tasks (e.g. the i2b2 2010 challenge), our entities were not always noun phrases or even well-contained chunks of information. For example, we annotated *"hepatic"* such as in *"hepatic lesions to be an anatomy liver entity"*. Ou and Patrick [127] reported similar experiences in their extraction. Our tumorhood evidence experience entity extraction was particularly interesting in this respect. Particularly, tumorhood evidence based on radiographic evidence, such as if *"hypervascularity"* or *"enhancement"* in addition to *"washout"* cues were present, the contained text would be considered positive for cancer. These men-

tions of enhancement could occur as adjectives to other entities, e.g. *"enhancing lesion"* or *"hypervascular lesion"*, and the *"washout"* may be mentioned very far from the enhancement, resulting in long spans of identified text with spurious words. On the other hand, if both positive mentions were not met, then cues were not highlighted, e.g. enhancing with no definite washout.

Other issues included medical abbreviations. This occurred for anatomy terms, e.g. *"SMV"*, (short for superior mesenteric vein) as well as tumorhood evidence terms, e.g. LR3 (short for LI-RADS, a coding system for tumor malignancy). Overtraining on context was another problem. For example negation and tumor counts worked better in short sentences, or around words they were most often found near. Tumorhood evidence isBenign evidence were difficult to differentiate since a non cancerous entity could be a number of things, e.g. *"nonocclusive chronic thrombi"*, *"cysts"*, *"likely related to old trauma/fracture"*, *"differential includes infection"*.

Our partial match performance entity extraction performance was comparable to Ou and Patrick [127] who achieved an overall 0.84 F1 and Hassanpour and Langlotz [74] with 0.85 F1. That said, we have many opportunities with which simple adjustments can make large improvements. For example, we may condition tumorhood evidence instead as a classification on our tumor reference or measurements (e.g. *"Is lesion cancerous?"*). In so doing, we can also conveniently re-introduce outside sentence information (e.g. previous sentence unigrams), incorporate our already extracted evidence, and address the abbreviations issue for LI-RADS. This extension is the subject of Section 7.7.

Because of our definitions of S1 and S2 lines, any line with a measurement is considered its own S1 line, therefore a sentence such as *"This compared to a prior measurement of approximately 6.4 x 6.0 cm"* may not be linked to a prior tumor reference or measurement as was in annotation. Some prior measurements were also associated with past dates, which was not included in this classification. Reparameterization of this classification and inclusion of additional features will be discussed in Section 7.7.

Our relation extraction performed similarly to other statistical methods, such as Taira et

al's 79% and 87% recall and precision[173]. Not directly comparable to our relation extraction or our template evaluation definitions, but of interest to compare, we report performances of similar information extraction subcategories from related works. Coden et al [41] reported evaluations of 0.82, 0.65, and 0.93 F1 for primary tumor, metastatic tumor, and lymph nodes class structures. Ou et al [127] achieved 0.84, 0.92, 0.29, 0.92, 0.33, 0.29, 0.84, 0.93, 0.90 F1 for clinical diagnosis, diagnosis, distant metastasis, lymphovascular invasion, microsatellites, other lesions, site and laterality, size of specimen, and tumor thickness fields. Roberts et al [142] reported 0.86 F1 for extracting anatomic anatomical sites per each radiological finding.

Our template extraction performances were low, in large part because of our punishing metric. Our templates were in fact graphs that required all relations to be exact even if it could provide equivalent information, as in Figure 7.17. Moreover our annotations were fairly detailed, so that there was often repeats of the same information regarding the same information within a template. As an example, Figure 7.18 is considered incorrect because *"hypervascular [...]demonstrate washout"* should have been highlighted as tumorhood evidence isCancer. Section 7.7 will incorporate a more lenient measure.

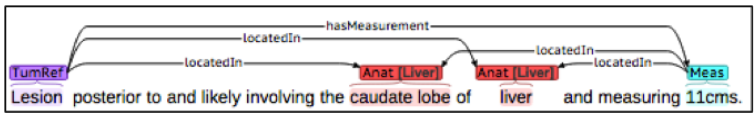


Figure 7.17: Incorrect extra relations to the measurement makes the entire template incorrect

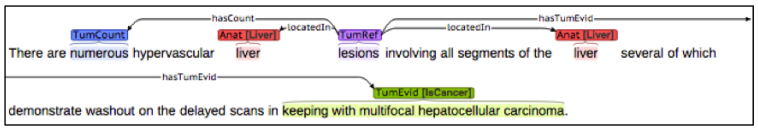


Figure 7.18: Missing extra radiographic tumor hod evidence cues causes entire template to be incorrect

7.7 Negation, temporal, and malignancy attribute classification

Realizing that not all attributes necessitate locating exact spans, or may become redundant with other evidence, we re-framed some attributes to be a classification on the entity they modify. We chose to do this for negation, temporality, and malignancy status. In the following sections, we describe our attribute classification using specialized feature engineering.

7.7.1 Adapting classifications of assertion to negation classification

For negation classification, we re-purposed our in house assertion classifier [27] to detect negation for tumor reference and measurement entities. Assertion classification is a classification of an entity, given the sentence it appears in, into one of six categories: *absent*, *conditional*, *hypothetical*, *not associated with patient*, *possible*, and *present*. To adapt assertion into a negation classification, we mapped assertion classification labels into negation labels. Thus, categories of *possible* and *present* were relabeled as PRESENT, whereas the remaining categories were mapped to NEGATED.

7.7.2 Feature-based classification on entities for temporal and malignancy

We identified temporal and malignancy attributes by classifying entities given heuristic- and domain- motivated features. Temporal attributes were only classified for measurement entities; malignancy attributes were classified for both measurement and tumor reference entities.

In addition to the categories of our previous entity and relation extraction in regards to temporal and malignancy classification task, we specified default values for unannotated tumor reference and measurement entities. Thus, if there were no annotations of *hadMeasurement* relations related to a measurement, then it is assumed to be CURRENT. Similarly, if there were no *hasTumEvid* associated with a tumor reference or measurement, it is given the default value of *UNK*, for *unknown*. If two values are associated with malignancy, they are combined, e.g. *INDET–BENIGN*, however any mention of *isCancer*, represented here as

MALIGNANT, will take precedence over combined values, e.g. *INDET–MALIGNANT* will be mapped to *MALIGNANT*.

We modeled the classification as a maximum entropy classification problem, as it has been shown to be successful with large amounts of features, such as text features. We used MALLET[110] for our machine learning implementation software. Detailed descriptions of the features used in classification are described in the following section.

7.7.3 Features

In this section, we detail the features for Section 7.7.2. Several feature parameters and feature classes were combined to produce a variety of final features. Below we give an overview of feature parameters and classes, with explanations of the features in Table 7.14 and an overview of feature extraction in Figure 7.19. During experimentation, we optimized in a greedy fashion for the best feature combinations.

Feature parameters

Our extraction had three adjustable feature parameters which affect n-gram and a couple of other features: (1) raw word vs. lemmatization, which affects all n-gram features, (2) word window, for n-gram features around an entity, and (3) sentence window, which specifies from which relative sentences to draw n-grams and rule-based entity features. The scope of features related to (2) and (3) are specified in column 2 of Table 7.14.

N-gram Features

1-, 2-, and 3- gram features were used as part of the sentence n-gram (SENTUNI) and surrounding n-gram (SURRUNI) features, each with sentence and word window restrictions, respectively. These feature either use raw words or lemma-izations for the n-grams depending on Section 7.7.3(1) configurations described above.

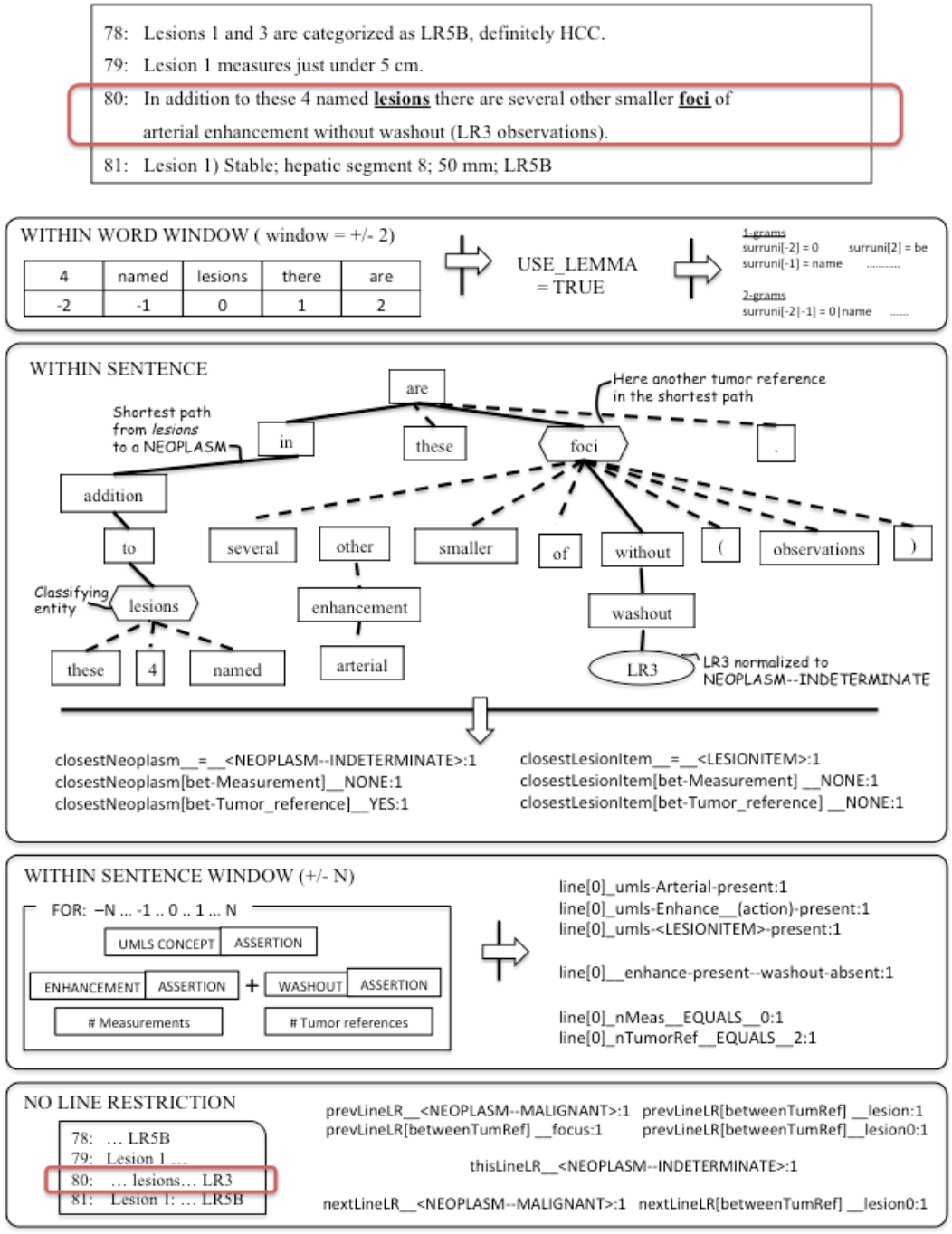


Figure 7.19: Feature extraction for various scopes

Entity Features

Features related to the entities being classified, e.g. *tumor reference* "lesion" or *measurement* "2.3 cm", were also included. Namely, these included the number of *tumor reference* or *measurement* entities in a sentence window (NUMTUMREF and NUMMEAS features).

Rule-based Entity Features

We used three classes of rule-based entity features. The first class of rule-based entities used Stanford Time Parser (SUTime) to identify temporal expressions, which were normalized to *before* or *concurrent* to the document date [35].

A second class of rule-based entities were constructed using regular expressions. These were designed to identify variations in LIRADS abbreviations as well as identify "enhancing" and "washout" terms. In Table 7.14, these are described in the LIRADS and ENHANCE-WASHOUT features. All synonyms of concepts descendant "certainty descriptor" (C0087130), e.g. "possibly," from the Taxonomy for Rehabilitation of Knee Conditions [160] were converted into regular expression and comprised of another type of entity which affected CLOSEST-CERTAINTYCUE features.

A third class of rule-based entities were first identified using MetaMap [22]. *Neoplasm*-related concepts were identified by taking all descendants of the concept *Neoplasm by Site* (C0027653) and combined by its neoplastic status, e.g. "malignant," "indeterminate," and "benign," extracted from the National Cancer Institute Thesaurus (NCI) [124]. *Lesion*-related concepts were identified by the union of "lesion" (C0221198), all descendants of "mechanical lesion" (C3872807) and all descendants of "mass" (C0577559) from the SNOMED CT 2015 ontology [7]. UMLS CUI related to certainty descriptors, generated by running MetaMap the list of synonyms from the "certainty descriptor" mentioned above, were also identified. These entities contributed to the features: ASSERT.UMLS, CLOSEST-CERTAINTYCUE, CLOSESTLESIONITEM, and CLOSEST-NEOPLASM.

Feature	Scope	Description
ASSERT.UMLS	Sentence window	UMLS concepts with assertion. If there are concepts related to certainty, lesion item, or neoplasms according to our CUI word lists, corresponding features indicating that these generalized terms appear, combined with their assertion values fire.
CLOSEST-CERTAINTYCUE	Within Sentence	Closest certainty cue through the dependency tree, combined with the closest lesion item, normalized neoplasm, or normalized LIRADS entity near the certainty cue.*
CLOSEST-DATE	Within Sentence	Closest date through the dependency tree*
CLOSEST-LESIONITEM	Within Sentence	Closest lesion item through the dependency tree*
CLOSEST-NEOPLASM	Within Sentence	Closest neoplasm with neoplastic status through the dependency tree*
DATES	No line restriction	The date normalized to before document date or document date. Outputs feature if any type of date is found within the line. The closest date mentions in previous or subsequent lines are also outputted.†
ENHANCE-WASHOUT	Within Sentence	Enhancement term with assertion, combined with washout terms with assertion.
LIRADS	No line restriction	LIRADS normalized to malignant, indeterminate, benign. Outputs feature if any type of LIRADS is found within the line. The closest LIRADS mentions in previous or subsequent lines are also outputted.†
NUMMEAS	Sentence window	number of measurement entities
NUMTUMREF	Sentence window	number of tumor references
SENTUNI	Sentence window	1-, 2-, 3- grams
SURRUNI	Word window	1-, 2-, 3- grams around the entity being classified

Table 7.14: Feature Descriptions. *Dependency path features also outputted a binary feature if a tumor reference or measurement is passed through the shortest path. †Features which looked through previous and next feature appearances, if fired, also outputted a feature if tumor references and a leading determiner were within the search path.

Shortest Dependency Path Features

Shortest dependency path features involved tracing the shortest path between tumor reference or measurement entities, e.g. *"lesion"*, to rule-based entities (Section 7.7.3) in the same sentence, e.g. *"UMLS_CONCEPT[Liver_Cancer]"*, through the dependency tree. Some of these features were related to the entities encountered through the shortest path, e.g. if another tumor reference entity is in the way of a tumor reference to a tumorhood evidence, as depicted in Figure 7.19. Other features compare competing rule-based entities with respect to the entity being classified, e.g. which tumorhood evidence is closer to a tumor reference. Several dependency path-related features are described in Table 7.14 marked with a * symbol. We used ClearNLP, trained on clinical text, to identify dependency parses [3].

For measurement entities attached to a tumor reference, dependency path features are measured from the attached tumor reference entity instead of the measurement entity itself.

Features over multiple sentences

Because dates or LIRADS abbreviations were at times mentioned far from an entity many sentences before or after, we allowed separate features to transcend the sentence windows. An example of this is shown in as in Figure 7.15, where *"LR4B"* is written many sentences after the mention of *"Lesion 2"*. For these features, the closest dates or LIRADS before or after the line of the entity would be identified and all tumor references in between would be used as additional features (DATES and LIRADS features). These features are marked with a † symbol and specified to have "No line restriction" scope in Table 7.14 column 2.

7.7.4 Evaluation

We evaluated based on individual slot attributes per each tumor reference and measurement entity classification, for negation, temporality, and malignancy, respectively. We also evaluated at a template level, where each template required correct slots for each negation, temporality, and malignancy as well as the complex relations with other entities. These can

35: Lesion 2) Increased untreated liver lesion:
36: i) Location: heaptic segment 6
37: ii) Size: 37 mm x 27.9 mm (previously 30.7 mm);
....
42: vii) LI-RADS category: LR4B

Table 7.15: Lirads malignancy beyond several lines

include complex templates, in which a tumor reference is connected to another sub-templates with a measurement and its own anatomy location, as shown in Figure 7.20. Measurements classified as PAST in templates were not counted, because they are considered obsolete information.

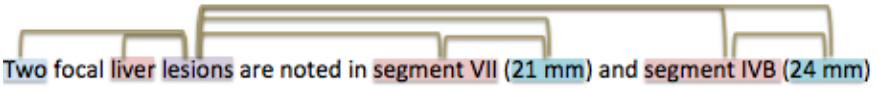


Figure 7.20: Complex template

For our evaluation metrics, we used the standard information extraction measures of precision, recall, and F1 score as defined in Chapter 3.2.

7.7.5 Results

Tables 7.16, 7.17, and 7.18 show results for classifications on gold standard tumor reference and measurements for negation, temporal, and malignancy classifications, respectively. The optimized features for malignancy included: ASSERT.UMLS, CLOSESTCERTAINTYCUE, CLOSESTLESIONITEM, CLOSESTNEOPLASM, ENHANCEWASHOUT, LIRADS, NUMMEAS, and NUMTUMREF with a sentence window of ±0 (only current sentence was used for those features restricted by the sentence window). The optimized features for temporal classification included, CLOSESTDATE, SURRUNI with a word window size of ±3.

For temporal classification, PAST instances in training were weighted to 20 times CURRENT instances to offset for the large class imbalance.

Table 7.16 shows our negation detection results after modifying our in-house assertion classifier. The NEGATED label improved with 7 more cases of true positives.

Label	Method	Pos.	TP	P	R	F1
NEGATED	75	Assertion	73	0.90	0.97	0.94
		Baseline	60	0.97	0.80	0.88
PRESENT	1216	Baseline	1214	0.99	0.998	0.99
		Assertion	1201	0.998	0.99	0.99

Table 7.16: Negation Classification Results

The baseline for the PAST category of temporal classification is expected to be low due to the constraints specifying relation candidates (which broke up potential *hasMeasurement*). However, while our new classification improves the situation, it is still at a modest performance of 0.62 F1 measure.

Label	Pos.	Method	TP	P	R	F1
CURRENT	457	Baseline	457	0.95	1.00	0.97
		Classifier	451	0.97	0.99	0.98
PAST	32	Baseline	7	1.00	0.22	0.36
		Classifier	16	0.80	0.50	0.62

Table 7.17: Temporal Classification Results

Table 7.18 shows that while the categories of INDET-BENIGN and UNK were relatively unchanged. There were sizable gains in the categories of INDET, BENIGN, and MALIGNANT categories.

Inspection of the confusion matrix, Table 7.19, shows that approximately 60% of misclassifications were due to incorrect assignment to UNK. Unsurprisingly, the low frequency category INDET-BENIGN was often mislabeled as either INDET or BENIGN.

	Pos.	Label	TP	P	R	F1
BENIGN	109	Baseline	34	0.83	0.31	0.45
		Classifier	50	0.73	0.46	0.56
INDET–BENIGN	5	Baseline	0	0.00	0.00	0.00
		Classifier	0	0.00	0.00	0.00
INDET	139	Baseline	64	0.89	0.46	0.61
		Classifier	101	0.80	0.73	0.76
MALIGNANT	456	Baseline	297	0.86	0.65	0.74
		Classifier	345	0.85	0.76	0.80
UNK	582	Baseline	552	0.66	0.94	0.78
		Classifier	463	0.76	0.80	0.78
ALL	1425	Baseline	1011	0.74	0.71	0.73
		Classifier	1060	0.80	0.74	0.77

Table 7.18: Malignancy Classification Results

		System				
		INDET	INDET–BEN	BENIGN	MALIG.	UNK
Gold	INDET	101	0	1	10	58
	INDET–BEN	6	0	4	0	2
	BENIGN	2	0	50	36	80
	MALIG.	12	0	10	345	148
	UNK	30	0	22	72	463

Table 7.19: Malignancy Confusion Matrix

Tables 7.20 reveal classifications and their influence at a template level. Improved performance at the granular level transferred well at the more integrated levels for templates. Templates are presented into subcategories as follows:

AnatomyMeas - events with *refersTo* relations

Negation - events with *isNegated* relations

TumorSingleton - events with a single entity, in which the entity is a tumor reference or measurement

OtherSingleton - events with a single entity, which are not *TumorSingleton* events
Tumor - events not part of the previous event types

We are particularly interested in the improved performance of *TumorSingleton* and *Tumor* events at 0.68 and 0.68 F1 for baseline, to 0.72 and 0.77 F1 after classification. Since we used gold standard entities for *Anatomy* and *TumorCount*, entities, *OtherSingleton* templates were expected to have a 1.00 F1 performance. Other types of template type performances were degraded depending on the classification of negation, temporality, and malignancy. Using our new classifications, tumor-related events (the combination of typed TumorSingleton and Tumor events) resulted in micro-score improvement from 0.65 to 0.72 F1.

Category	**Method**	**TP**	**P**	**R**	**F1**
AnatomyMeas	Baseline	49	0.92	0.92	0.92
	Classifier	43	0.81	0.81	0.81
Negation	Baseline	60	0.97	0.80	0.88
	Classifier	73	0.90	0.97	0.94
OtherSingleton	Baseline	70	1.00	1.00	1.00
	Classifier	70	1.00	1.00	1.00
TumorSingleton	Baseline	113	0.60	0.72	0.65
	Classifier	114	0.64	0.73	0.68
Tumor	Baseline	427	0.63	0.66	0.65
	Classifier	475	0.72	0.74	0.73

Table 7.20: Template Classification Results

7.7.6 Error Analysis and Discussion

Analysis of negation classification error analysis showed that adaption of assertion classification for negation classification worked well. One consistent error, which accounted for the majority of false positives, was for the case of long clauses, e.g. *"Stable wedge-shaped hyper vascular focus in segment 4 A with no definite washout suggesting indeterminate lesion"*

where *"lesion"* was incorrectly identified as negated. Other errors were due to annotation errors, report errors (spontaneous *"no"* appears), and unmarked hypothetical cases (considered negated).

Temporal classification errors were primarily due to false negatives. Meanwhile, manual inspection revealed false positives to be due to annotation errors. Despite our inclusion of SUTime to normalize for past dates, it appeared our classifier was not able to use this feature effectively as several false negatives had clear attached dates or references to times before the document date. Furthermore, many of the false negatives were those in which clues for past indicators were infrequent, e.g. *"prior,"* *"formerly,"* and *"prev"*. These observations, combined with the our knowledge of the optimized features for this classification, suggested that despite our efforts, although we were able to overcome the hurdle of relying on relation extraction for past classification, our model at this point does not generalize well for temporal cues. In fact, an overwhelming number of examples of past measurements have a preceding *"previously"*. That said, our training size was small and our features were relied heavily on SUTime normalization. We expect increasing data size, incorporating synonyms of *"previous"* or *"formerly"* which SUTime does not catch, relating these to the classification through the dependency tree, and merging with SUTime features, would yield much better results.

Interestingly, the optimized malignancy classification feature set did not include n-gram based features, suggesting that our specially designed high-level rule-based features captured important variables without resorting to surface level information. A significant hurdle in malignancy classification was that many mislabels were due to cues existing outside the current line, which occurred in approximately 15% of the cases. While one of our features, LIRADS, did reach outside of lines we found that increasing sentence window size to outside sentences resulted in performance degradation. As a consequence of this, it was unsurprising that many of the incorrect classifications were to UNK as, without outside line information, the state of a tumor reference or measurement would in fact be considered UNK. BENIGN classifications were a particular challenge as it encompasses a variety of things. For example,

a tumor reference or measurement would be considered BENIGN if they turned out to be imaging artifacts, scarring, or simple cysts. Though we incorporated lesion items to try to normalize these variations in our features, we could not account for all the diversity, e.g. *"perfusion abnormality"* was not specially grouped. Also we should note that according to our annotation rules, any mention of "malignant" or "suspicious" automatically classifies an entity as MALIGNANT. However, there were cases of hedging that may have created confounding factors, e.g. *"differential diagnosis includes dysplastic nodule versus atypical HCC"* where *"dysplatic nodule"* is considered to be indeterminate for malignancy (they are pre-malignant and follow-up is necessary). Given these observations, we note several ways to improve this classification. As we designed special LIRADS features that captured information in many lines before and after, we can also do this for UMLS entities related to neoplasms and lesion items and encode cues for subheading scopes. To target benign entities further, we could exploit more subsections of ontologies to enrich the feature space. To combat the effects of hedging, we could engineer features that take into account the numbers of identified malignant, indeterminate, or benign entities capture by the UMLS.

For negation and temporal attributes, false positives, e.g. mistaken classification of negated cases, are less of a concern than false negatives. A patient typically has multiple imaging files, therefore what is disregarded due to being negated or not current may be pulled in from another source. Keeping this in mind, it would be possible to change the decision boundary to optimize for the best outcome for a given dataset in such downstream applications. For malignancy classification, the importance of each category performance depends on the intended use. In general, a tumor-like finding that is found to be benign is less important than the malignant category or indeterminate category (which may turn benign). An application may only be concerned with malignant tumors such as our use for liver cancer staging, though its feasible to imagine an decision support tool that wants to track whether or not indeterminate tumors become malignant or benign at a later time state. However, as the confusion is often with the unknown category, it would be difficult to easily modify the system to get a better yield in this particular classification.

Evaluation at the template levels revealed that improved classification at the granular level reverted into improved classifications at a structured level. Templates associated with AnatomyMeas and Negation were impacted by occasional mentions of malignancies attached to those types of templates, e.g. *"Lymph nodes suspicious for metastatic involvement: Periportal greater than 1 cm"* and *"Pelvis: No suspicious lytic or sclerotic bony lesions."* Though not focused on in this work, they are nevertheless important cancer-related information.

7.8 Tumor reference classification and characteristics extraction

Some of the problems already encountered involved knowledge located in outside sentences. In general, in order the understand holistic report information, we need to be able to resolve which mentions, e.g. tumor reference entities, refer to the same real-world entity. In the following sections we describe our system to classify coreferrent and parent-child particularization relations.

7.8.1 Related work

Reference resolution is an active area of research in the natural language processing domain. General english NLP focus on reference resolution has primarily been on newswire text, with several notable information events such as the Message Understanding Conference (MUC) [71] and the Automatic Content Extraction (ACE) program [53]. Similar to our goals, one previous work that attempts to classify event, subevents, etc. using a pairwise logistic regression classifier. [21]

In the biomedical domain, the BioNLP 2011 Shared Task featured anaphoric coreference of biomedical entities, e.g. biological entities, processes, and gene expressions. [90].

In the clinical domain, annotation of a variety of concept types, e.g. person, tests, problems, for coreference, has been the focus of the 2011 i2b2/VA Cincinnati challenge. [175] Some difference between our task and that of the 2011 i2b2/VA Cincinnati challenge are the following: (a) we target very few specific mentions (tumor references instead of large classes such as person, test, or problems) and (b) our annotation is based on smaller noun

phrase chunks. For example, the i2b2 challenge puts references between long noun phrases which includes descriptors such as: *"a left facial mass"*, *"a right parietal hyper dense and heterogeneously enhancing mass"*, *"an endobronchial tumor of the right upper lobe bronchus"*, *"a 5mm linear , focal area of enhancement in the left central semiovale"*. In contrast, our references are between shorter phrases, e.g. *"hypervascular lesion"* or *"tumor"*. Similar to our task, the Ontology Development and Information Extraction (ODIE) part of the corpus has been annotated with anaphoric references, with identity, set/subset, and part/whole relations. [38]

Related works on reference resolution relevant to tumors or clinical findings have been the subject of several works. Coden et al [41] identified coreferences in pathology reports using a rule-based system. Son et al [158] classified coreferrent tumor templates between documents with a MUC score of 0.72 precision and 0.63 recall. Sevenster et al [150] paired numerical finding measurements between documents.

Actual reference resolution tasks vary widely in scope. For example, nouns, pronouns, and noun phrases are common; however, corefernce for nested noun phrases or nested named entities, (e.g. "America" in "Bank of America"), relative pronouns, and gerunds may not be annotated in a corpus. [164] Here our references are between the template heads of tumor templates. Our corpus does not include pronominal cases and nested references.

7.8.2 Reference resolution classifier

Our reference resolution classifier consists of a greedy algorithm which visits each template in the order of appearance in each document, and classifies the head of a template as EQUIV, SUBSETOF, SUPERSETOF, and NONE for each available cluster. If the template is EQUIV to one or more clusters, the template is added to the clusters and merged. For all other choices, the template forms a new cluster. As conflicts or cycles may arise at each classification round, we used simple heuristics to resolve these. If any cycles formed, all mentions would be merged. If there was a conflict between a NONE relation and another one, the other relation would take precedence. We leave more sophisticated decision-making

algorithms for future work. Relations between clusters were updated during the process.

Figure 7.21 depicts the choice of a new potential cluster being being classified with one of the relation labels for each available existing cluster. Classifications were trained using LibSVM and MALLET, using a linear kernel. Feature values are scaled by the difference between the minimum and maximum values.

7.8.3 Reference resolution features

We detail several types of features shown in Table 7.21. Some classes of these features are described in the following section.

Normalized anatomic location features

If anatomical entities are detected for a template, they are normalized to an anatomic concept. Based on this concept, we designed features based on anatomic hierarchy, e.g. "segment

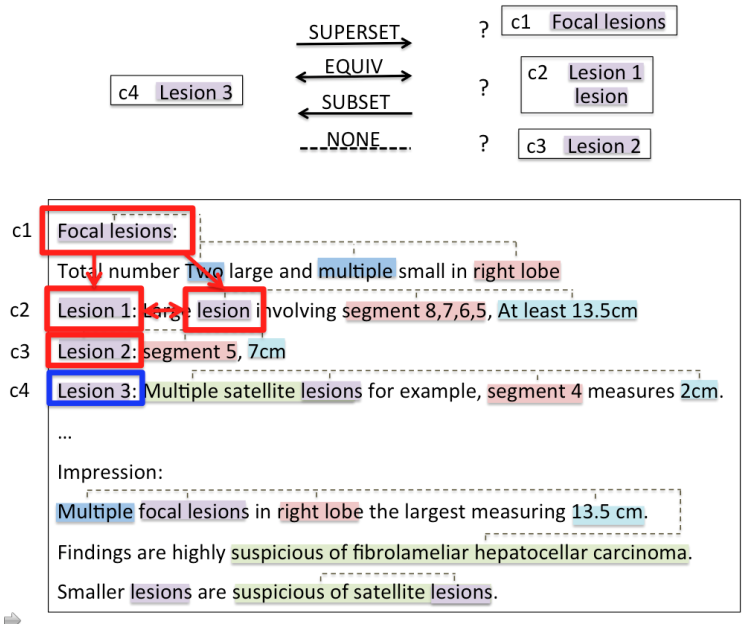


Figure 7.21: Reference resolution set up

Feature Name	Description
closestTempDist	The distance of the closest template in a candidate cluster to the current template
containedIn	If any of the anatomies in the current template are contained in the anatomy in the candidate cluster
containerOf	If any of the anatomies in the candidate cluster are contained in the current template
header	If the sentence of the template looks like a section header
isSuperset	If the candidate cluster is already a superset of another cluster
malignancy	Malignancy status of template
malignancyOfCandCluster	Malignancy status of the cluster
nextBestSim	L-2 norm of the next best similarity vector
ngrams	1-, 2-, and 3- grams (using lemma) for sentences of template and candidatecluster
ngramsMatching	Matching 1-, 2-, and 3- grams (using raw words) for sentences of template and candidatecluster
nthTemplate	The number template in the document
numOfCand	The number of candidate clusters
numOfMeas	The number of measurements
numOfTempInCluster	The number of templates in the candidate cluster
onlySameMal	The only candidate cluster with matching malignancy as template
onlySameMeas	The only candidate cluster with matching measurement malignancy as template
sameOrgan	If the organ in the sentence matches organ in a cluster
sameLocations	The matching locations of all
section	Section of the template
sim	The L2-norm of similarity vector
simvecfeats	This feature extends from the similarity vector features so that each individual similarity vector dimensions are each considered their own feature
summaryOf	If tumor reference is preceded with “the”, “this”, “these”
totalNumOfTemp	Total number of templates in the document
totalNumOfImpTemp	Total number of impression in the document
UMLS	Matching UMLS concept between the template and the cluster
Underheading	If there is a sentence in the cluster that looks like a header, and if the number leading 5 characters of variations in sentences leading up to the template is less than 4

Table 7.21: Reference resolution features

VIII” is contained in “liver”. The processing and normalization of anatomic entities is further described in Section 7.8.5. Normalization was based on Unified Medical Language System (UMLS) [31] concept names. Relevant related features are **containedIn**, **containerOf**, and **sameLocations**.

Positional features

Whether or not a template appears in the top or near the bottom of the template will affect how many options it will be clustered to and the threshold to what cluster similarity should be in order to be paired. We included several features related to the position of a template over all templates in a document. For example, **nthTemplate** gives both the absolute number and the ratio of the template position normalized to the number of all the templates.

Target	Description
sentence similarity	Jaccard proximity for sentence, word-tokenized
tumor reference similarity	Jarowinkler string proximity
number of measurements	Difference between number of measurement entities divided by the larger number of measurements
tumor count similarity	Difference in tumor count divided by the larger tumor count
matching measurement1	The number of matching measurements divided by the total number of measurements in template 1 (Measurements considered matching if within 0.1 cm)
matching measurement2	The number of matching measurements divided by the total number of measurements in template 2
anatomy1	Sum of pairwise jarowinkler proximity for all anatomy entity combinations between template 1 and 2, divided over the number of anatomy entities in template 1
anatomy2	Sum of pairwise jarowinkler proximity for all anatomy entity combinations between template 1 and 2, divided over the number of anatomy entities in template 2
malignancy1	The number of matching malignancy status (combined malignancy status' get broken up, e.g. "INDET-BENIGN" becomes "INDET" and "BENIGN"), divided by the total number of malignancy status for template 1
malignancy2	The number of matching malignancy status, divided by the total number of malignancy status for template 1

Table 7.22: Similarity features description

Relative features

Relative features identify differences between candidate clusters. For example, **onlySameMal** is in the case of if a candidate cluster is the only one of the candidate clusters which has the same malignancy status. Another exists for same measurement.

Static features

Static features includes a variety of features, such as the **section** of the template, **n-grams** in the sentence, and number of measurements (**numOfMeas**). These features remain the same regardless of what candidate cluster a template head reference is being classified with.

Similarity features

Similarity features (**simvecfeats** and **sim**) are measured from the current template head to be classified to an existing candidate cluster. The similarity with the entire cluster is measured by taking the maximum of each similarity dimension among all the templates in the existing candidate clusters. For all dimension except for 0 and 1, subset candidate templates features are combined and normalized together.

Similarity features include the sentence similarity features, tumor reference similarity, as well as similarity between template attributes. For example, tumor reference similarity, measurement similarities, anatomy similarities, and anatomy similarities. The total of all similarity features combine to form a similarity vector of 9 dimensions. Each dimension is described in the Table 7.22.

7.8.4 Tumor characteristics extraction

The tumor characteristics annotator receives grouped tumor templates and outputs (1) the number of tumor for each malignancy category, (2) the largest size malignant tumor, and (3) whether 50% of the liver is taken up by malignant tumors, using a series of heuristic rule-based algorithms. The various system components parts are shown in Figure 7.22. First

the templates are updated to a new malignancy status depending on their coreference and particularization relations to other templates, next the templates are sent through several various pipelines depending on the chosen variable. In the following sections, we describe several of the non-obvious components in the pipeline: the module for updating malignancy status, the module for classifying whether >50% of liver is invaded, and the module for consolidating referenced tumors.

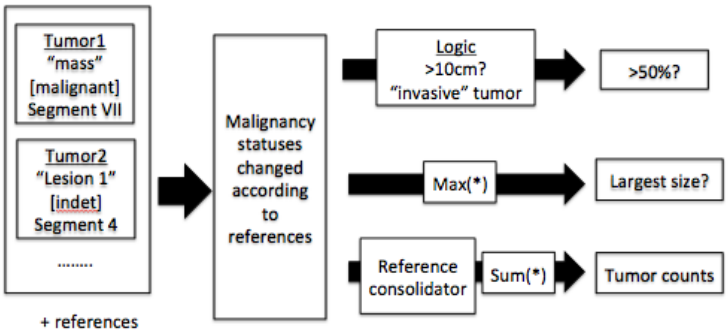


Figure 7.22: Tumor characteristics annotator

Updating malignancy status

The malignancy statuses for related tumor templates are updated in the following way. The malignancy status for coreferrent templates are updated to the most critical case. Thus, anything coreferrent to a malignant tumor template is also malignant; if the most critical status is indeterminate then all templates are updated to indeterminate. In regards to particularizations (superset/subset relations), we take a top-to-bottom approach. The status of the superset is transferred down to the templates in the subset. After this top-down-transfer, the inter-cluster malignancy status is updated once more. Extension of this forward-backward algorithm continuously is left for future work.

Invasion of >50% of liver logic

The logic for deciding whether or not >50% of the liver is invaded, as shown in Figure 7.23. The algorithm is based on the expert guidelines introduced in Figure 7.9.

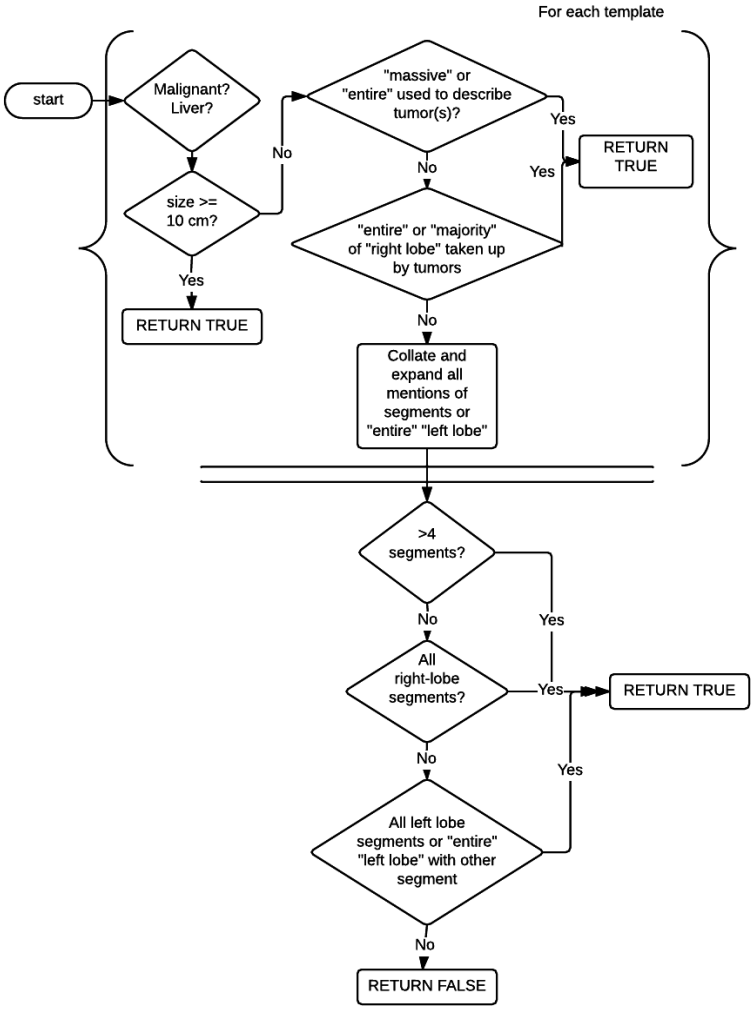


Figure 7.23: Algorithm for >50% liver is invaded

Concepts such as “right lobe”, “left lobe”, and “liver”, as well as decisions on which lines are segments are involved, are based on the anatomy normalizations from the anatomy normalization module, described in Section 7.8.5.

Reference consolidator

The reference consolidator is responsible for updating templates to the most current set of information and removing extraneous other templates. The premise is to be able to refine all the given information to a few representative templates. For example, if a reference in "Several liver lesions, suspicious for HCC" has the particularizations of "Lesion 1: segment 8, 3.0cm" and "Lesion 2: segment 5: 2.1x1.1 cm" then the template associated with the first passage will be (1) updated with measurements of "3.0cm" and "2.1x1.1 cm", (2) updated with anatomies of "segment 8" and "segment 5", and (3) updated to have a number of "2" for tumor count. Furthermore, if the particularization templates match the malignancy status of its superset template then those are deleted. The final result should yield a set of tumor templates with updated count, measurement, anatomy, and malignancy attributed that can be easily summed to determine the number of each type of tumors found in the radiology report.

Our exact algorithm includes heuristics for deciding for unambiguous cases, for example:

- If the tumor count is set to 3 what happens if there are more than 3 measurements?
- If the tumor count is not reliably determinable, how should it be decided based on the number of associated measurements?

Both coreference and particularization relations are used in the decisions.

7.8.5 Anatomy normalizer module

Even with properly marked anatomy entities, concept normalization requires both conjunction normalization as well as concept disambiguation. For example, "segment 2, 4A/B, and 5" must be normalized to "segment 2", "segment 4A", "segment 4B", and "segment 5". Furthermore, "left lobe" may refer to "lung" or "liver".

Anatomy named entity clauses are normalized to discrete concepts by first, determining the organ dictionary to use using the organ context of the sentence. Afterwards, text-spans

adjusted to account for missed endings for system entities, e.g. "**segments VIII and V**/IVb", and terms are conjunction-normalized. Finally, concepts are matched based on the lowest score of summing together the matching edit distance with any leftover substrings.

In the following sections, we describe our rule-based algorithms for how to map sentences to an organ context and how to normalized for conjunctions; as well as our automatic creation of organ-specific hierarchal dictionaries using the Foundation Model of Human Anatomy (FMA) ontology [143].

Mapping sentences to organ scope

In order to disambiguate between ambiguous anatomic locations, e.g. "left lobe" the organ context for a sentence must be understood. However, this information is not always available within a sentence, requiring external information. An example of this is shown in Figure 7.24.

Findings:
Lungs bases: There is calcification of the coronary arteries.
There is a new 1.3 x 0.9 cm a sub pleural nodule in the right base.
No pleural effusion.
Abdomen:
Liver: Nodular cirrhotic liver.

Figure 7.24: Different parts of the report have anatomical context not necessarily immediately available in the same sentence or not explicitly clear. In the third sentence, "right base" can be inferred to be part of the lungs by the reference to "Lungs bases" in the previous sentence or the mention of "pleural" in the same sentence.

Our algorithm is detailed as follows. From starting at the beginning of a document to the end, each sentence, previously tagged with UMLS concepts using MetaMap [22], was categorized as related to one or more organ concepts, if these two conditions were met: (1)

an anatomic location semantic type was found and (2) the corresponding matched string was matched to the organ dictionary. The list of semantic type abbreviations included in the anatomic location list are: **anst**, **bdsy**, **blor**, **bpoc**, **bsoj**, and **tisu**. The list of organs UMLS concept identifiers was created by recursively identifying "is-a" relations starting from the top (non-inclusive) concepts listed in Figure 7.25.

"Organ with caveated organ parts" (C0927231)
"Organ with organ parts" (C0927230)
"Nonparenchymatous organ" (C0935295)
"Lobular organ" (C0927223)
"Corticomedullary organ" (C0927224)
"Homogeneous organ" (C0927225)

Figure 7.25: Starting organ concept identifiers

Our algorithm also assigns organ context by matching to organ-related adjectives, e.g. "hepatic" refers to the liver. The mapping from a organ-related adjective to an organ was created by taking pertainyms from WordNet [30] which point to a MetaMap-matched organ. Examples of the resulting dictionary is shown in Table 7.23 with the full list included in Appendix G. If no match occurs, the previous line's organ is set for the current line. At the start of each section, the assigned organ is reset to a default state. In our case, the default organ concept was set to the liver.

A manual review of 5 randomly drawn documents (215 sentences) revealed a precision of 94% for this procedure.

Organ	Adjective forms
kidney	nephritic, renal, adrenal
liver	hepatic
lung	pulmonic, lung-like, pulmonary, pneumogastric, pneumonic, cardiopulmonary, intrapulmonary
prostate	prostatic, prostate
spleen	lienal, splenetic, splenic
tibia	tibiall

Table 7.23: Organ adjectives identified using WordNet pertainyms. As bones are considered organs in the FMA, adjective forms of specific bones are also included (tibial). A full list is given in Appendix G.

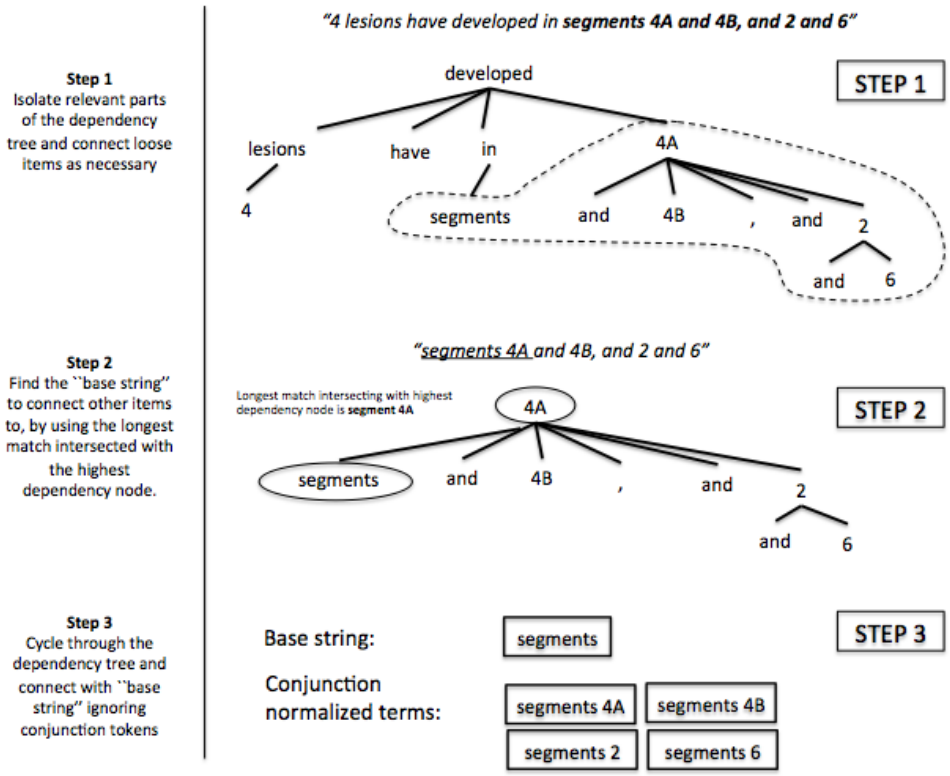


Figure 7.26: Conjunction normalization process. **Step 1**: Isolate relevant parts of the dependency tree and connect loose items as necessary. **Step 2**: Find the "base string" to connect other items to, by using the longest match intersected with the highest dependency node. **Step 3**: Cycle through the dependency tree and connect with "base string" ignoring conjunction tokens.

Normalizing for conjunctions

Conjunctions were normalized by first finding the longest match from organ-specific dictionaries. The automatic creation of these hierarchal organ-specific dictionaries is detailed in Section 7.8.5. The overlap of the longest match was then intersected with the highest node of the sentence dependency tree. The longest match was determined by finding the terms with the lowest edit distance. Starting from this match, the center-most word is popped off. Then, each unused word from the anatomy entity is paired with the match, ignoring terms such as "and", "or", "/" , "-" and "." . The construction of the pairings for "segments 4A and 4B, and 2 and 6", as shown in Figure 7.26. The same algorithm is designed to also be used for cases such as "Tumor thrombus within **main, right and proximal left portal veins"**.

Our aim here was to provide a way to capture both types of conjunction problems that we encounter for our anatomy entities, such as the right-branching conjunctions of "segments x, x, and x" as well as the left-branching conjunctions of "x, x and x portal veins" in the least assuming way possible. Thus, the generalization of this heuristic for other cases is left for future investigation.

Organ-specific hierarchal dictionary creation

Portions of each organ's hierarchal constituent structures were extracted starting from the organ concept identifiers listed in the previous section. The concepts were collected by recursively following relations: **has_regional_part**, **has_constitutional_part**, and **has_attributed_part**.

Synonym dictionaries for each concept was augmented by adding synonyms in which roman numerals were replaced with numbers (1-12), e.g. "segment II" would be duplicated with variant "segment 2". Synonyms that required mentions of the specific organ, e.g. "right lobe of the liver", were also duplicated with the removal the organ mention to allow better matching, e.g. "right lobe". The following regular expressions were used to identify portions of synonyms to be augmented: " of [organ]$", "$\hat{}$[organ] ", "$\hat{}$[organ-adjective] ". These

regular expressions were created after studying the naming conventions of the FMA. As a caveat, this may not generalizable to all ontologies and is subject to changes of the FMA terminology.

7.8.6 Results

Tables 7.24 show coreference and particularization classifications results, using gold standard tumor templates, with a simple baseline of **ngrams** and **ngrams-matching** features compared with our system with the full set of features. Though there was little improvement for particularizations, the coreference performance increased sizably.

Evaluation	**N-grams + Ngrams matching**			**All Features**		
	P	R	F1	P	R	F1
Coreference classification						
MUC	0.49	0.35	0.41	0.63	0.54	0.58
B-cubed	0.82	0.72	0.77	0.84	0.78	0.81
CEAF	0.61	0.36	0.45	0.68	0.52	0.59
$\frac{MUC+B3+CEAF}{3}$			0.54			0.66
Particularization relation						
particularization	0.51	0.39	0.44	0.42	0.44	0.43

Table 7.24: Reference resolution results. (P=precision, R=recall, F1=F1-score)

In order to quantify how well our tumor characteristics annotator works, we experiment with using no reference resolution information, using gold standard reference resolution, and, finally, system reference resolution using gold standard templates. The results are shown in Table 7.25. Given system reference resolution annotations, the tumor characteristics significantly dropped, however the performance remained high for the $> 50\%$ variable, and dropped less drastically for the largest size variable, compared to those for the tumor count variables.

We were also interested in knowing how the two components affect our entire system end-to-end. That is given, system produced templates, what is our tumor characteristics

	No Ref. Res.		Gold Ref. Res.		System Ref. Res.	
	TP	F1	TP	F1	TP	F1
>50%	94	0.93	95	0.94	95	0.94
#benign	72	0.71	80	0.79	76	0.75
#indet	67	0.66	78	0.77	76	0.75
#malig	14	0.14	70	0.69	56	0.55
#unk	12	0.12	60	0.59	52	0.51
largest size	80	0.79	94	0.93	87	0.86

Table 7.25: Tumor characteristics annotation results (gold standard templates)

annotation results? The comparison results are shown in Table 7.26. From these results we see the $> 50\%$ variable remains high, suggesting that it is a variable that is more robust to changes in reference resolution errors as well as template extraction problems. The tumor count variables for all types of malignancies are shown once again to drop substantially. However, this makes sense as even with perfect gold reference resolution, our annotation logic would not get past 0.80 exact F1; furthermore, figuring out the number of tumors requires very exact reference resolution information, making the tolerance for errors very much lower. The largest size variable was the least affected using both the system references and system templates; this accounts to the ability of the metric to absorb errors (it uses a maximum function).

	No Ref. Res.		System Ref. Res.		System Ref. Res. (relaxed)	
	TP	F1	TP	F1	TP	F1
>50%	89	0.90	90	0.90	90	0.90
#benign	67	0.67	66	0.66	68	0.68
#indet	64	0.64	68	0.68	72	0.72
#malig	18	0.18	50	0.50	62	0.62
#unk	2	0.02	34	0.34	34	0.34
largest size	63	0.63	77	0.77	79	0.79

Table 7.26: Tumor characteristics annotation results (system standard templates)

We also allowed our system to only process certain sections of the document, e.g. Findings only, Impression only, or both (default). We present our results of doing so for our three important variables in Table 7.27, with different combinations of gold and system reference resolution and template annotations.

	Section		
	Findings	**Impression**	**Both**
>50%	0.81/ 0.80/0.69	0.89 /0.89/0.78	**0.94/0.93/0.90**
#malig	0.67/0.56/0.40	**0.69/0.61/0.56**	0.69/0.50/0.50
Largest size	0.76/0.70/0.57	0.43/0.39/0.37	**0.93/0.86/0.77**

Table 7.27: Tumor characteristics annotation results restricted by section measured in accuracy (gold-templates, gold references / gold-templates, system references / system-templates, system references)

While the tumor count variable for malignant tumors did better using only the Impression section, the other two variables benefitted from having information across both the Findings and Impression section. Interestingly, the largest size variable is much lower for the Impression section compared to the Findings section, which reinforces the observation we have found that more detailed information are often kept in the Findings section, with more summary information in the Impression section.

7.8.7 Error Analysis and Discussion

Tumor reference resolution classification

Analyzing the misclassification of relations, we found that of the 358 FP particularizations, 350 were represented in the gold standard except with the opposites direction (supersetof/subsetof switch) and 27 corresponded to equivalent relations in the gold standard (the may overlap with the supersetof/subsetof switch since they are not mutually exclusive). Similarly, for the 253 FN, 217 were reversed and 34 were related to equivalent relations in

the system.

There are many areas for improvement with this classification. Firstly, the greedy merge approach for all coreference and particularization loops is simplistic. An algorithm that resolves this issue by ranking probabilities of each individual relation may do better to resolve the loop without causing large chain reactions. In general english, there are constraints such as pronoun agreement ("John" and "he" vs "her") that are used for coreference systems. We did not implement any such constraints, partly because of our small corpus. Some ideas in this vein could be constraints against different "named lesions" being in the same cluster, e.g. "Lesion 1" and "Lesion 2". Our classification for each template to all candidate clusters were done individually, though perhaps joint classification could yield marginally better results. Finally, our system aggregated clusters from top to bottom in a greedy fashion, allowing the possibilitiy of cascading errors.

Tumor characteristics annotation

Analysis of the tumor characteristics annotator using gold standard templates and referencer resolution annotations revealed some interesting phenomenon. While the tumor count errors was partly due to our system not producing inequalities (which is required in the gold standard under strict evaluation), it was also due to the heuristic rules of changing malignancy status (only if coreferrent or top-down) and in merging. Futhermore, while particularization hierarchies may go down several levels, we limited our number, measurement, and anatomy update rules to a scope to 3 levels.

In the case of $> 50\%$ invasion of the liver, there were only a handful of mistakes. One false positive was due to a possible typo in the report (listed as 24 cm in Findings but 24 mm in the Impression), one false negative in which no template was attached to a malignancy evidence finding (it was outside the Findings/Impression section), and one case which was labelled "n/a" due to no size or anatomy in the report at all. The remaining cases included one false negative in which "both segments" was not converted to mean segments 1-8 in the liver and a false positive in which "majority" was not meant to modify "liver" in the

sentence.

For the largest size variable, two errors were due to no malignancy evidence attached to templates, four errors were due to either differences in reported measurements (mistakes or simply precision differences, e.g. 2.5 cm vs. 2.4 cm). Finally, one error was due to malignancy status not being updated in a down-up fashion.

7.9 Tumor related document performance

Table 7.28 shows the result of the tumor characteristics annotator, with non-liver templates excluded, and with a rule-based transfer to the normalized tumor related liver cancer stage parameter values. For example, the number of malignant tumors are mapped to *tumor number - single*, *tumor number - 2-3*, or *tumor number* - > 3. A similar thresholding system was used for *tumor size* variables. *Tumor morphology* on the other hand used information about the number of malignant tumors as well as whether $\geq 50\%$ of the liver is invaded.

Analysis of the F1 scores shows that every category improved with this system over baseline except for *tumor number* - > 3. *Tumor number* required, entity, relation, attribute, coreference/reference classification and tumor characteristics annotation; thus, though there was some improvement over the baseline, it was not very dramatic. Manual review of the mistakes revealed that some of the *tumor morphology* gold standard had some annotation errors; otherwise the greatest confusion was between *multinodular,* $< 50\%$ and *uninodular,* $< 50\%$. The performance of the *tumor size* variables were expectedly better than the baseline as finding measurements using regular expression and normalizing them to a size is more reliable than a simple n-gram baseline.

			Baseline				Tumor char. ann.		
Label	**Freq.**	**Value**	**Class.**	**P**	**R**	**F1**	**P**	**R**	**F1**
Tumor	23	Massive	DT	0.37	0.30	0.33	0.57	0.52	0.55
morphology	40	Multinodular, <50%	ME	0.50	0.15	0.23	0.43	0.63	0.51
	105	Uninodular, <50%	NB	0.62	0.80	0.70	0.80	0.73	0.77
Tumor	112	Single	NB	0.64	0.84	0.73	0.76	0.73	0.75
number	32	2-3	DT	0.24	0.25	0.25	0.34	0.53	0.41
	19	>3	ME	0.67	0.11	0.18	0.18	0.16	0.17
Tumor	82	< 3	ME	0.64	0.62	0.63	0.89	0.79	0.84
size	45	3-5	C45	0.43	0.27	0.33	0.77	0.82	0.80
	46	>5	ME	0.59	0.28	0.38	0.79	0.59	0.68

Table 7.28: Best baseline performances for training set.

7.10 Summary

In this work, we present our annotation as well as our system design for tumor template extraction, tumor reference resolution, and tumor characteristics annotation. The tumor information extraction system here are biased towards findings that are not at first known. For example, we did not mark HCC in *"HCC in segment 8"* as a tumor reference. However through further annotation and some minor changes in the system and annotation algorithm, these cases may be augmented for.

Although our reference resolution and tumor characteristics extraction results are modest, through our experiments we can see that improvements in reference resolution will also lead to improvements in downstream tasks. Finding the number of tumors proved to be the most difficult variable, as it requires very precise reference annotations. Meanwhile the other variables, $> 50\%$ and largest size were more tolerant to errors.

Some limitations to work is that our dataset is from a single institution for a creation cohort of patients, our corpus size is small, and our annotations were mostly single-annotated. Our corpus annotations were specific towards tumors and not generalizable towards general

medical concepts. In the following chapters, we use the models built on this subset to classify the rest of the corpus.

Chapter 8

PATIENT LEVEL CLASSIFICATIONS

In this section, we incorporate previous modules and classify patient level stage parameter and liver cancer stages. We also compare with different frame-works of stage classification, e.g. classification based on raw words, on normalized concepts, or on patient level normalized concepts.

8.1 Related work

Patient classification is related to the general problem of clinical phenotype extraction. Specific cancer-related cases have had their own strategies for aggregating classifications for patient level evaluation. For example, with patient cancer identification, a patient was classified as positive for a case if at least one positive extraction in any part of the patient record was found [145][187]. In another example, the final categorizations of MRI cancer progression for Cheng et al [39] were determined by getting the category that maximized the summed probability of all a patient's subdocuments.

Of the cases related to cancer staging, we reference Nguyen et al [121], in which patient records were concatenated into a a single file and least to most severe classifications of sub-stages were determined until finding the final stage combination. In Nguyen et al [122], which used a heuristic algorithm, the most severe sub-stages were tested before moving to the least severe cases until a final stage combination was reached.

Here we describe our patient level classification for stage and stage parameters, for which we ultimately used a rule-based algorithm. In general, our strategy for stage parameter aggregation to the patient level was to use a "most severe rule" heuristic. Our conversion from stage parameters to stages used a domain expert rule-based algorithm.

8.2 Rule-based patient level stage parameter classification

To summarize, from the previous extraction systems we have the following data, described in Table 8.1

Stage parameter	**Extraction method**	**Granularity**
Laboratory data	Structured	Timestamp
Child-Pugh ECOG	Regular expression	Text highlights
Ascites Extrahepatic invasion Hepatic encephalopathy Macrovascular invasion Metastasis Portal hypertension	Sentence classification	Text highlights
Tumor morphology Tumor number Tumor size	Entity, relation extraction Reference resolution Tumor characteristics extraction	Entities and relations Document level annotation

Table 8.1: Extraction data summary

Given the automatic annotations resulting in the previous extraction systems, there are many options on how to arrive at patient level stage parameter values, e.g. take highest frequency. For our system, we use the simple heuristic of taking the most severe finding for each category in a patient. If no findings were found, the least severe of each category was given as default. Table 8.2 shows the result of the rule-based heuristic using system and gold annotations for the 160 training set patients. We also provide, for comparison, the result of a classifier that just outputs the majority class.

Besides *macrovascular invasion* and *metastasis*, our heuristic using system annotations outperformed the majority class baseline. As observed, in the results, gold annotations may not lead to a 1.00 F1 match. Part of the reason for this is that our algorithm assigns values to a patient regardless of whether or not there is evidence of it. In the annotated gold

standard, this is not the case as missing or conflicting information may lead to an unscoreable result. Lower performance with gold annotations of *ECOG* may be explained by the annotation guidelines' specifying use of the least severe case for disambiguation, opposite to our heuristic. Another stage parameter that our patient level resolution algorithm may not work best for is *Child-Pugh*. *Child-Pugh* values can either be read from text (explicit values) or calculated (non-explicit values); and it is possible to have errors in either form because of either human calculation of the explicit values, or extraction problems for the non-explicit values. This discrepancy is explored more in the next section. Which values to trust would require some logic regarding classification confidences. In the future, these decisions may be changed according to clinically motivated reasons.

	Majority class		**System annotations**		**Gold annotations**	
Label	***TP***	***F1***	***TP***	***F1***	***TP***	***F1***
Ascites	120	0.75	130	0.81	155	0.97
ChildPugh	99	0.62	139	0.87	155	0.97
ECOG	96	0.60	135	0.84	139	0.87
Extrahepatic_invasion	156	0.98	156	0.98	157	0.98
Hepatic_encephalopathy	136	0.85	145	0.91	158	0.99
Macrovascular_invasion	138	0.86	134	0.84	155	0.97
Metastasis	147	0.92	145	0.91	158	0.99
Portal_hypertension	96	0.60	131	0.82	151	0.94
Tumor_morphology	90	0.56	114	0.71	148	0.93
Tumor_number	102	0.64	106	0.66	151	0.94
Tumor_size	74	0.46	128	0.80	153	0.96

Table 8.2: Patient level stage parameters classification performance using text annotations on training set (total 160 patients)

8.2.1 Child-Pugh values comparison

With a method to resolve text annotations to a patient level stage parameter, we can compare several different versions of *Child-Pugh* from the gold and system annotations. The first is

to take the most severe *Child-Pugh* from text annotations (text), e.g. mentions such as *"Child's class A"*, the second is to use the calculated values of *Child-Pugh* with input from patient level *ascites* or *hepatic encephalopathy* (calculated), the third is to use the patient level annotations of *Child-Pugh* (patient level annotation). For patients which had at least one *Child-Pugh* text annotations (67 patients), we put the F1 agreement numbers in Table 8.3.

Annotation	**Text vs. calculated**	**Text vs. patient level annotation**	**Calculated vs. patient level annotation**
Gold	0.81	0.85	0.96
System	0.74	0.79	0.85

Table 8.3: Different comparisons of Child-Pugh

The comparison between text and patient level annotation for gold annotations reveals that the stated explicit annotations (calculated by the clinician at the time) do not always match the true *Child-Pugh* class (calculated on review by our domain experts). The difference between calculated and patient level annotations signals internal inconsistencies due to annotation errors. The disagreement between the *Child-Pugh* text values compared to the gold-standard is interesting. Some possible causes are either calculation errors of either the attending clinician of the note or our annotators, perhaps as a consequence of the cognitive load required to stage patients, or using different laboratory values for calculations.

For system annotations, the agreements quantify the consistency between taking the text annotations by themselves (text), calculating from the value using the *Child-Pugh* algorithm (calculated), and taking the most severe out of the two (patient level annotation) for the 73 patients with at least one explicit value.

8.3 Rule-based patient level stage classifications

With patient level stage parameters resolved, only the final classification for patient level stages are left. For this task, we experimented using several levels of annotation from both

the gold and the system annotations. We compared (1) a classifier that names the majority category, maximum entropy classifiers using (2) document 1-, 2-, 3- gram and normalized UMLS concepts (using MetaMap) with frequency features, (3) text annotation label values with frequency features, and (4) stage parameter label value features, and (5) a set of rules generated from the expert created lookup tables created during annotation (Appendix B).

The results using gold and system annotations on the 160 patient cross-validated training set are shown in Table 8.4 and 8.5, respectively. Table 8.5 additionally gives relaxed scores if *AJCC stage IIIA-IIIC* were merged into *AJCC stage III*, *AJCC stage IVA-IVB* were merged into *AJCC stage IV*, and if *BCLC stage A1-A4* were merged into one *BCLC stage A*.

	Majority class		**Doc. class.**		**Text annot.**		**Stage param.**		**Rules**	
Label	***TP***	***F1***	***TP***	***F1***	***TP***	***F1***	***TP***	***F1***	***TP***	***F1***
Stage_ajcc	87	0.54	86	0.54	138	0.86	140	0.88	157	0.98
Stage_bclc	57	0.36	57	0.36	112	0.70	138	0.86	155	0.97
Stage_clip	53	0.33	48	0.30	72	0.45	110	0.67	152	0.95

Table 8.4: Gold patient level stage parameters classification performance on training set (total 160 patients)

	Text annot.		**Stage param.**		**Stage param. (r)**		**Rules**		**Rules (r)**	
Label	***TP***	***F1***	***TP***	***F1***	***TP***	***F1***	***TP***	***F1***	***TP***	***F1***
Stage_ajcc	96	0.60	109	0.68	113	0.71	103	0.64	108	0.68
Stage_bclc	76	0.48	92	0.58	109	0.68	98	0.61	111	0.69
Stage_clip	115	0.28	88	0.55	88	0.55	88	0.55	88	0.55

Table 8.5: System patient level stage parameters classification performance on training set (total 160 patients), where (r) is the relaxed stage match

While each higher level of annotation using gold annotations led to better performances, with system annotations this trend did not carry through. In the end, using system annota-

tions, having more accurate staging system did not translate into better classifications due to the cascading errors from lower levels.

8.4 Patient evaluation on the test set

The results of the models trained on the entire training set and decoded for the test set are shown in Table 8.6. The drop in performance was most notice-able for the tumor-related stage parameters. Meanwhile, the general performance degradation signals overtraining on the training set.

Label	***TP***	***F1***
Ascites	29	0.73
ChildPugh	35	0.88
ECOG	32	0.80
Extrahepatic_invasion	40	1.00
Hepatic_encephalopathy	35	0.88
Macrovascular_invasion	34	0.85
Metastasis	39	0.98
Portal_hypertension	34	0.85
Tumor_morphology	23	0.58
Tumor_number	23	0.58
Tumor_size	33	0.83
Stage_ajcc	22	0.55 (0.60)
Stage_bclc	20	0.50 (0.55)
Stage_clip	17	0.43 (0.43)

Table 8.6: Patient level stage parameters classification on test set (total 40 patients), parenthesis scores are the relaxed stage matches

Comparison of corpus population to the test set revealed some differences in the overall sampling trend for *BCLC* and *CLIP* stages, which may account for some of the discrepancies between the cross-validated training set performance and the test set performance in Table 8.7. For a quantification of classification mistakes, we give the confusion matrices for the three stages in Table 9.2, 9.3, and 9.4.

Label	Value	Frequency (corpus)	Frequency (test)
AJCC	I	108	20
	II	48	15
	IIIA	16	2
	IIIB	14	3
	IIIC	0	0
	IVA	6	0
	IVB	7	0
BCLC	A1	27	2
	A2	21	4
	A3	13	7
	A4	17	4
	B	23	11
	C	70	11
	D	14	1
CLIP	0	66	10
	1	62	18
	2	41	6
	3	18	5
	4	8	0
	5	4	1
	6	0	0

Table 8.7: Stage annotations

		System							
			I	***II***	***III-a***	***III-b***	***III-c***	***IV-a***	***IV-b***
Gold		0	0	0	0	0	0	0	0
	I	0	**13**	8	0	0	0	0	0
	II	0	4	**7**	0	1	0	0	0
	III-a	0	1	0	**1**	1	0	0	0
	III-b	0	1	0	1	**1**	0	0	0
	III-c	0	0	0	0	0	**0**	0	0
	IV-a	0	1	0	0	0	0	**0**	0
	IV-b	0	0	0	0	0	0	0	**0**

Table 8.8: AJCC stage confusion matrix

		System							
			A1	***A2***	***A3***	***A4***	***B***	***C***	***D***
Gold		0	0	0	0	0	1	2	0
	A1	0	**0**	1	0	0	1	0	0
	A2	0	0	**2**	1	0	3	0	0
	A3	0	0	0	**2**	0	1	0	0
	A4	0	0	0	0	**3**	0	0	0
	B	0	1	1	2	1	**3**	0	0
	C	0	0	0	2	0	2	**9**	0
	D	0	1	0	0	0	0	0	**1**

Table 8.9: BCLC stage confusion matrix

		System							
			0	***1***	***2***	***3***	***4***	***5***	***6***
Gold		0	0	0	0	0	0	0	0
	0	0	**6**	5	1	1	0	0	0
	1	0	2	**10**	3	1	0	0	0
	2	0	2	3	**1**	2	0	0	0
	3	0	0	0	1	**0**	0	0	0
	4	0	0	0	0	1	**0**	1	0
	5	0	0	0	0	0	0	**0**	0
	6	0	0	0	0	0	0	0	**0**

Table 8.10: CLIP stage confusion matrix

8.4.1 Sensitivity analysis

To get a sense in the error propagation for each patient level stage parameter to patient level stage assignment, we performed two sets of sensitivity analysis experiments. The first set assumes system patient level stage parameter annotations for all except one gold stage parameter (Table 8.11). The second set assumes gold patient level stage parameter annotations for all except one system stage parameter. If given the correct number of tumors, there are drastic changes in performance for AJCC and CLIP, as shown with *tumor morphology* and *tumor number*. Recall, *tumor morphology* does take into account the tumor numbers as well. The BCLC stage, on the other hand, tended to be less volatile according to changes in one specific stage parameter, suggesting that it has a more equitable reliance on each of its stage parameters.

Note, the different staging schemes use different parameters, so not every change in variable will lead to changes in performance. The BCLC stage is the only one that uses *ECOG* stage parameters, for example. For comparison, we also provide the sensitivity experiments for the training set (Table 8.12), which exhibits the same trends.

	Ascites	ChildPugh	ECOG	Extrahepatic_invasion	Hepatic_encephalopathy	Macrovascular_invasion	Metastasis	Portal_hypertension	Tumor_morphology	Tumor_number	Tumor_size
Stage_ajcc	0.55	0.55	0.55	0.55	0.55	**0.60**	0.58	0.55	0.55	**0.85**	0.55
Stage_bclc	0.50	0.50	**0.63**	0.50	0.50	0.55	0.53	0.53	0.50	**0.63**	0.55
Stage_clip	0.43	**0.50**	0.43	0.43	0.43	0.43	0.43	0.43	**0.80**	0.43	0.43

Table 8.11: Sensivity analysis substituting one gold stage parameter in test set

The reverse experimental observations also corroborated the importance of *tumor morphology* and *tumor number*, where system values degraded the system much more than other

	Ascites	ChildPugh	ECOG	Extrahepatic_invasion	Hepatic_encephalopathy	Macrovascular_invasion	Metastasis	Portal_hypertension	Tumor_morphology	Tumor_number	Tumor_size
Stage_ajcc	0.64	0.64	0.64	0.65	0.64	**0.74**	0.73	0.64	0.64	**0.76**	0.67
Stage_bclc	0.61	0.64	**0.73**	0.61	0.61	0.67	0.64	0.66	0.61	**0.69**	0.64
Stage_clip	0.55	0.60	0.55	0.55	0.55	**0.64**	0.55	0.55	**0.74**	0.55	0.55

Table 8.12: Sensivity analysis substituting one gold stage parameter in train set

parameters (Table 8.13 and 8.14) for test and train sets, respectively.

	Ascites	ChildPugh	ECOG	Extrahepatic_invasion	Hepatic_encephalopathy	Macrovascular_invasion	Metastasis	Portal_hypertension	Tumor_morphology	Tumor_number	Tumor_size
Stage_ajcc	0.97	0.98	0.98	0.98	0.98	0.90	0.95	0.98	0.98	**0.63**	0.95
Stage_bclc	0.93	0.93	**0.83**	0.93	0.93	0.90	0.90	0.88	0.93	**0.75**	**0.85**
Stage_clip	1.00	0.88	1.00	1.00	1.00	0.93	1.00	1.00	**0.58**	1.00	1.00

Table 8.13: Sensivity analysis substituting one system stage parameter in test set

	Ascites	ChildPugh	ECOG	Extrahepatic_invasion	Hepatic_encephalopathy	Macrovascular_invasion	Metastasis	PortalHypertension	Tumor_morphology	Tumor_number	Tumor_size
Stage_ajcc	0.98	0.98	0.98	0.98	0.98	0.88	0.89	0.98	0.98	**0.85**	0.94
Stage_bclc	0.97	0.94	**0.87**	0.97	0.97	0.91	0.95	0.93	0.97	**0.88**	0.95
Stage_clip	0.95	**0.84**	0.95	0.95	0.95	**0.84**	0.95	0.95	**0.69**	0.95	0.95

Table 8.14: Sensivity analysis substituting one system stage parameter in train set

8.5 Summary

This chapter covered the rule-based algorithms for patient level classifications. Included, as well, were experiments using various levels of gold and system annotations. We used a simple heuristic of aggregating patient level stage parameters, however for future work, it is easy to experiment with other methods, e.g. using the highest frequency. The patient level stage classifications used a rule-based method, which was high-performing assuming its stage parameter input were accurate. For future work, in order to decrease this low level of robustness, it would be interesting to experiment with methods that take into account confidences regarding multiple stage parameter values at different granularities, e.g. making a classification based on text evidence as well as patient level values with associated uncertainties.

The final performance of the system on the test set was unfortunately not as high as estimated with the cross-validation training set. Analyzing the sensitivity of each parameter, we found that the weakest link was the tumor numbers, which, non-coincidentally, was the hardest stage parameter to correctly classify, due to requirements of discourse-level information.

Chapter 9

EXPERT VS NON-EXPERT PATIENT CLASSIFICATIONS

Because hospital human abstractors are typically staff personnel with some medical background, e.g. a nurse, but not a domain expert, e.g. an interventional radiologist, we wanted to experiment with the accuracy of a non-expert with medical background for patient level stage parameter and stage assignments.

A non-expert medical student was trained for six hours over three days. Afterwards, their task was to assign patient level annotations for stage parameters and the 3 stages, given the free text notes and the necessary laboratory values. 100 of the patients were randomly chosen from the entire set and the results are compared to the consensus expert interventional radiologist gold standard patient level annotations.

9.1 Results

The overall classifications for all patient labels (stage parameters and stages) are shown in Table 9.1. In general patient level annotation for stage parameters were quite high. Only 3 out of 11 parameters were below 0.90 F1. The lower performing categories were *ascites*, *Child-Pugh*, and *tumor morphology*. Possibly ascites was more problematic because of the number of locations it may appear in. *Child-Pugh* stage on the other hand required either calculating the new stage or finding explicit mentions of *Child-Pugh* in the text. Finally, *tumor morphology* is one of the most difficult staging parameters as it requires disambiguating malignancy in various findings and performing reference resolution over one (or more) documents. The average time spent on staging per patient was 13 ± 4 minutes. The minimum, maximum, and median time spent per patient was 6, 34, and 12 minutes, respectively.

Label	*TP*	*F1*
Ascites	86	0.86
ChildPugh	89	0.89
ECOG	88	0.88
Extrahepatic_invasion	94	0.94
Hepatic_encephalopathy	95	0.95
Macrovascular_invasion	93	0.93
Metastasis	98	0.98
Portal_hypertension	94	0.94
Tumor_morphology	86	0.86
Tumor_number	92	0.92
Tumor_size	94	0.94
Stage_ajcc	89	0.89 (0.90)
Stage_bclc	82	0.82 (0.86)
Stage_clip	75	0.75 (0.75)

Table 9.1: Expert versus non-expert patient level annotations

Below is a breakdown of the 3 stages shown as confusion matrices. While most of the mistakes were off the center for AJCC, CLIP stages and BCLC stages were more scattered. This is most likely caused by the comparative complexity of BCLC staging.

		System							
			I	***II***	***III-a***	***III-b***	***III-c***	***IV-a***	***IV-b***
Gold		0	1	0	0	0	0	0	0
	I	0	**49**	4	0	0	0	0	0
	II	0	2	**25**	1	0	0	0	0
	III-a	0	0	0	**5**	0	0	0	0
	III-b	0	0	1	0	**4**	1	0	0
	III-c	0	0	0	0	0	**0**	0	0
	IV-a	0	0	0	0	0	0	**3**	0
	IV-b	0	0	0	1	0	0	0	**3**

Table 9.2: AJCC stage confusion matrix

		System							
			A1	***A2***	***A3***	***A4***	***B***	***C***	***D***
Gold		**2**	0	1	0	0	0	2	0
	A1	0	**9**	2	0	0	1	0	0
	A2	0	0	**9**	1	0	2	0	0
	A3	0	0	1	**5**	0	0	0	0
	A4	0	0	0	0	**8**	0	0	0
	B	0	0	0	0	0	**10**	2	1
	C	0	2	1	1	0	0	**31**	1
	D	0	0	0	0	0	0	0	**8**

Table 9.3: BCLC stage confusion matrix

		System							
			0	***1***	***2***	***3***	***4***	***5***	***6***
Gold		0	1	0	0	0	0	0	0
	0	0	**25**	5	1	0	0	0	0
	1	0	2	**24**	2	3	0	0	0
	2	0	0	3	**17**	5	0	0	0
	3	0	0	0	1	**6**	0	0	0
	4	0	0	0	0	0	**1**	2	0
	5	0	0	0	0	0	0	**2**	0
	6	0	0	0	0	0	0	0	**0**

Table 9.4: CLIP stage confusion matrix

9.2 Summary

A non-expert trained to discriminate liver cancer stagings achieved high accuracies. As with the system performance, errors among the stage parameter propagated to overall stage classification performance. The annotator took an average of 13 minutes to compete a staging task for 40 patients. While, the annotator had limited training and further more experience would increase the speed of the task, simultaneously the effort required for staging more patients may cause substantial cognitive load and thus annotation fatigue, possibly resulting

in annotation errors.

Comparatively, a human annotator performs significantly better than our automated system. As described in the previous chapter, one of the main performance bottlenecks is identifying the correct tumor number for a patient. The superior performance of a human is consistent with the fact that anaphoric reference resolution is still an unsolved problem.

Chapter 10

CONCLUSIONS AND FUTURE WORK

We conclude this thesis with a summary of results, a discussion of our contributions, a review of the limitations in this thesis, and finally a look to future work.

10.1 Summary of results

Table 10.1 summarizes the annotation work in this thesis, while Table 10.2 summarizes the classification results for this work.

Annotations	**Granularity**	**Description**
Clinical note sections	Text highlights	Annotations demarking clinical note sections
Stage parameter text evidence	Text highlights	Annotations of text evidence related to 11 liver cancer stage parameters
Patient level stage parameters	Patient labels	Patient level values for 11 liver cancer stage parameters
Patient level stages	Patient labels	Patient level stage values for three liver cancer staging schemes
Tumor templates	Entities & relations	Radiology report tumor-related events
Tumor reference resolution	Relations	Reference resolution between tumor-related event heads for coreference and particularization relations
Tumor characteristics	Document labels	Document level annotations for tumor counts, sizes, and whether 50% of liver is invaded

Table 10.1: Summary of annotation work

Classification	Agreement	Baseline	System
Stage parameter text evidence	0.85	0.63	0.75
Patient level stage parameters	N/A	0.71	0.95
Patient level stages	N/A	0.41	0.97
Tumor templates	0.70	0.38	0.42
Tumor reference resolution	0.95/0.84	0.54/0.44	0.66/0.43
Tumor characteristics	0.91	N/A	0.79

Table 10.2: Summary of classification results (F1) for cross-validation sets. Each component assumes gold standard inputs. For example, patient level stage parameter classification assumes gold text annotation input; tumor reference resolution classification assumes gold tumor templates input. Stage parameter text evidence are measured at the document level. Tumor reference resolution measured for coreference (averaged MUC, B3, and CEAF) and particularization.

10.2 Contributions

In this thesis, we contribute (1) a framework for normalizing concepts and severities/attributes using data driven statistical classification, (2) a sparse annotation approach for tumor findings, an annotation methodology for tumor reference and characteristics, (3) experiments with anaphoric reference resolution using a greedy feature-rich classifier, (4) a phenotype algorithm for liver cancer stages of AJCC, BCLC, and CLIP stages, and (5) characterizations of challenges for the classification for the 11 stage parameters and 3 stages in liver cancer.

An important issue to address for this thesis is that though our system has some reasonable performances for stage parameters, they and their corresponding cancer stage classifications do not perform near human level. As our goal is to automate staging to facilitate evidence-based medicine, this result is somewhat problematic. In regards to this, we can offer two consoling points. The first is that each module in our system, using the least assumptions possible for our tasks, has great potential for growth with small configuration

changes and with more annotated data.

Secondly, though our full system does not provide clinically acceptable performances, parts of it may be used as tools for pre-annotation or in conjunction with human inputted data for stage parameters that are not achievable using current methods. For example, the sentence classifiers may be used to find important sentences with normalized values, and a human can examine whether or not they are correct. This would potentially cut annotation costs. Similarly, for tumor characteristics annotation, entity and relation classification may be used to highlight important parts of the report. Of course, whether or not pre-annotation would decrease annotation time, currently approximately 13 minutes per patient, requires validation through human-centered interaction research. Another use case is taking human input for the more problematic stages, while keeping the higher performing modules. As identifying the number of tumors is the most difficult challenge in this thesis, if this was inputted by a human, the baseline for the patient classification would increase.

Other cancer stages may benefit from using the same methods and workflow we build. Diagnosed medical conditions not captured by problem lists may also follow similar patterns as our stage prediction workflow. Furthermore, our subtask classifications for individual stage parameter identification may be applicable or relatable to other studies.

10.3 Limitations

One of the limitations of our work include that it is from small dataset of a single institution. Therefore our models may be overtrained or may not generalize well to the language of other institutions. Annotation typically involved a small number of domain experts.

The liver cancer stage parameter and stage classifications applied to 3 stages but may not collect enough information for all other liver cancer stage schemes. Our population was skewed towards early to intermediate stage patients, and we also assumed we could correctly identify irrelevant files and report types, e.g. addendum file types, etc. Our staging was also targeted towards patients before treatments. However, staging of patients after treatments start may involve different text evidence and more complex criteria for identifying relevant

symptoms.

10.4 Conclusions and future work

This thesis details our automated system for liver cancer staging. The problems that we encounter are typical for complex patient clinical phenotype classifications. We offer some of our solutions that are relevant to other similar tasks. During the course of our work, we identified a significant challenge in our classification tasks as issues of anaphoric references and reasoning over discourse.

Though our system is currently not clinically viable, several improvements may lead to reasonable performance gains. For example, with more training examples, our stage parameter extraction methods for sentence classification and tumor entity and relation extraction would increase. Adding additional rules, e.g. to handle conjunctions for *ECOG* and *Child-Pugh* stages, would increase performance with not much further effort.

Barring use of the full system, certain submodules can be used in conjunction with humans to decrease annotation costs. This is related to one of our described themes of **structured vs. unstructured** data entry. A balance must be achieved to optimize the maximum data collection strategy while minimizing the cost of data collection given resource constraints. The pragmatic goals, in regards to application of NLP and AI in clinical decision support, is to prioritize important data entry tasks for humans that cannot be easily or reliably done with a machines; meanwhile decreasing data-entry tasks for humans which can be instead be dealt with using machines. Then, as a secondary use step, this distilled data may leverage methods to facilitate evidence-based research or even as more gold standard data to improve existing classification algorithms.

We point out the strength in our approach is the interpretability and trace-ability for phenotype classification. Our system can not only predict a patient stage value, but can also provide the predicted patient level stage parameter values and the relevant text evidence. This approach is desirable given our second theme, that **medical conditions have arbitrarily complex signs and symptoms**. Configurable definitions for medical conditions

and accompanying signs and symptoms may have different criteria depending on the study. For example, a subset of the TREC 2011 Medical Records Track challenge [56], shown in Table 10.3, already includes different levels of specificifications for which a variety of text evidence may be useful for. Meanwhile, the use of statistical algorithms within the overall methodology maintains some ability to increase performance as more data is added and the ability to adapt to other datasets.

Topic number	**Description**
101	Patients with hearing loss
102	Patients with complicated GERD who receive endoscopy
104	Patients diagnosed with localized prostate cancer and treated with robotic surgery
105	Patients with dementia
112	Female patients with breast cancer with mastectomies during admission
123	Diabetic patients who received diabetic education in the hospital
133	Patients admitted for care who take herbal products for osteoarthritis
135	Cancer patients with liver metastasis treated in the hospital who underwent a procedure

Table 10.3: Subset of topics in the Text Retrieval conference (TREC) 2011 Medical Records Track challenge to retrieve patients eligible for clinical studies.

Another strength of our system is its modularity. Separate modules of the system were divided by granularities as well as by the classification category task. Connected to this was the challenging decisions of how and where to require separate layers of integration, a problem that touches upon the issues of **information completeness and consistency**. We used only some simple heuristics for integration here, e.g. taking most severe stage parameter cases, however there is room to experiment with trying different integration methods, e.g. highest frequency, or with integration methods taking in multiple levels of integration, e.g. using subdocument text evidence as well as patient level information and classification confidences.

With improvements, we hope our system could eventually be integrated with other knowledge, e.g. treatment billing codes and life expectancy, to bolster evidence-based medicine.

10.5 Final remarks

Advancing computer processing and storage has led to great strides in AI. The last several decade has heralded improvements in digital assistants (Alexa, Cortana, and Siri), self-driving vehicles, and self-balancing robots. Today, machine learning is used for a variety of tasks including fraud detection, consumer recommendations, improved internet search results, and voice-recognition.

With further research into clinical NLP and parallel integration into EMRs, it is hopeful to expect the advancements in AI to aid in healthcare. The trends for this is promising. Besides the incorporation of EMRs in the hospital setting, institutions are pooling together similar cohorts into large conglomerate databases, e.g. eMERGE. Private companies are also investing into personal patient records, e.g. Practice Fusion and CareCloud. Increasingly, hospitals are supporting health information exchange as well as new technologies that provide new ways of patient data collection, such as through emails, images, or other application tools. Furthermore, other data collection modalities powered by the patients themselves, e.g. Fitbit data and social media logs, may become increasingly visible and important for personalized medical care.

With all the advancements and conveniences of modern society, it would be welcoming to have such improvements ease unnecessary the suffering or the bottlenecks for a portion of the population that we or a family member will, at one time or another, inevitably fall under. Though there are still many technical and logistical barriers against effective streamlining of clinical encounters as natural feedback data points for evidence-based medicine, we are optimistic about the current atmosphere and hope that our work will add to the growing technology.

BIBLIOGRAPHY

[1] American recovery and reinvestment act of 2009 (2009 - h.r. 1). https://www.govtrack.us/congress/bills/111/hr1.

[2] Apache openNLP. https://opennlp.apache.org/index.html.

[3] ClearNLP. https://github.com/clir/clearnlp.

[4] Crossing the quality chasm: A new health system for the 21st century. http://www.iom.edu/Reports/2001/Crossing-the-Quality-Chasm-A-New-Health-System-for-the-21st-Century.aspx.

[5] EUR-Lex - 52012dc0736 - EN - EUR-Lex. http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52012DC0736.

[6] i2b2 2014 NLP shared task. https://www.i2b2.org/NLP/HeartDisease/Main.php.

[7] International Release of SNOMED CT. https://www.nlm.nih.gov/healthit/snomedct/international.html.

[8] MIMIC II: Clinical database overview. http://www.physionet.org/mimic2/mimic2_clinical_overview.shtml.

[9] OntoNotes Release 5.0 - Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC2013T19.

[10] Overview of the ShARe/CLEF eHealth evaluation lab 2014. http://www.academia.edu/8379281/Overview_of_the_ShARe_CLEF_eHealth_Evaluation_Lab_2014.

[11] Precision Medicine Initiative. https://www.nih.gov/precision-medicine-initiative-cohort-program.

[12] SemEval-2015 task 14: Analysis of clinical text. http://alt.qcri.org/semeval2015/task14/.

[13] Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, pages 2825–2830, November 2011.

[14] i2b2 2011 NLP Shared Task. `https://www.i2b2.org/NLP/Coreference/Call.php`, February 2015.

[15] Asma Ben Abacha and Pierre Zweigenbaum. Medical entity recognition: A comparison of semantic and statistical methods. In *Proceedings of BioNLP 2011 Workshop*, BioNLP '11, pages 56–64, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[16] Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F. Styler, Colin Warner, Jena D. Hwang, Jinho D. Choi, Dmitriy Dligach, Rodney D. Nielsen, James Martin, Wayne Ward, Martha Palmer, and Guergana K. Savova. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, 20(5):922–930, September 2013.

[17] Alias-i. LingPipe 4.1.0. `http://alias-i.com/lingpipe`, 2008.

[18] Sean F. Altekruse, S. Jane Henley, James E. Cucinelli, and Katherine A. McGlynn. Changing hepatocellular carcinoma incidence and liver cancer mortality rates in the united states. *The American Journal of Gastroenterology*, 109(4):542–553, April 2014.

[19] Sean F. Altekruse, Katherine A. McGlynn, Lois A. Dickie, and David E. Kleiner. Hepatocellular carcinoma confirmation, treatment, and survival in surveillance, epidemiology, and end results registries, 1992-2008. *Hepatology (Baltimore, Md.)*, 55(2):476–482, February 2012.

[20] M. Ando, Y. Ando, Y. Hasegawa, K. Shimokata, H. Minami, K. Wakai, Y. Ohno, and S. Sakai. Prognostic value of performance status assessed by patients themselves, nurses, and oncologists in advanced non-small cell lung cancer. *British Journal of Cancer*, 85(11):1634–1639, November 2001.

[21] Jun Araki, Zhengzhong Liu, Eduard Hovy, and Teruko Mitamura. Detecting Subevent Structure for Event Coreference Resolution.

[22] A. R. Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium*, pages 17–21, 2001.

[23] Naveen Ashish, Lisa Dahm, and Charles Boicey. University of california, irvine-pathology extraction pipeline: The pathology extraction pipeline for information extraction from pathology reports. *Health Informatics Journal*, August 2014.

[24] Amit Bagga and Breck Baldwin. Algorithms for Scoring Coreference Chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566, 1998.

[25] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.

[26] Cosmin A. Bejan, Lucy Vanderwende, Heather L. Evans, Mark M. Wurfel, and Meliha Yetisgen-Yildiz. On-time clinical phenotype prediction based on narrative reports. *Proceedings of the AMIA Symposium*, 2013:103–110, 2013.

[27] Cosmin Adrian Bejan, Lucy Vanderwende, Fei Xia, and Meliha Yetisgen-Yildiz. Assertion modeling and its role in clinical phenotype identification. *Journal of Biomedical Informatics*, 46(1):68–74, February 2013.

[28] Cosmin Adrian Bejan, Wei-Qi Wei, and Joshua C. Denny. Assessing the role of a medication-indication resource in the treatment relation extraction from clinical text. *Journal of the American Medical Informatics Association*, pages amiajnl–2014–002954, October 2014.

[29] Asma Ben Abacha and Pierre Zweigenbaum. Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of Biomedical Semantics*, 2(Suppl 5):S4, October 2011.

[30] Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. Revising the Wordnet Domains Hierarchy: Semantics, Coverage and Balancing. In *Proceedings of the Workshop on Multilingual Linguistic Ressources*, MLR '04, pages 101–108, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[31] Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–270, January 2004.

[32] Patricia Flatley Brennan and Alan R. Aronson. Towards linking patients and clinical information: detecting UMLS concepts in e-mail. *Journal of Biomedical Informatics*, 36(4-5):334–341, October 2003.

[33] Kelly W. Burak and Norman M. Kneteman. An evidence-based multidisciplinary approach to the management of hepatocellular carcinoma (HCC): the alberta HCC algorithm. *Canadian Journal of Gastroenterology = Journal Canadien De Gastroenterologie*, 24(11):643–650, November 2010.

[34] David S. Carrell, Scott Halgrim, Diem-Thy Tran, Diana S. M. Buist, Jessica Chubak, Wendy W. Chapman, and Guergana Savova. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. *American Journal of Epidemiology*, 179(6):749–758, March 2014.

[35] Angel X. Chang and Christopher Manning. : A library for recognizing and normalizing time expressions. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, 2012.

[36] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011.

[37] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310, October 2001.

[38] Wendy W. Chapman, Guergana K. Savova, Jiaping Zheng, Melissa Tharp, and Rebecca Crowley. Anaphoric reference in clinical reports: characteristics of an annotated corpus. *Journal of Biomedical Informatics*, 45(3):507–521, June 2012.

[39] Lionel T. E. Cheng, Jiaping Zheng, Guergana K. Savova, and Bradley J. Erickson. Discerning tumor status from unstructured MRI reports–completeness of information in existing reports and utility of automated natural language processing. *Journal of Digital Imaging*, 23(2):119–132, April 2010.

[40] Lee M. Christensen, Peter J. Haug, and Marcelo Fiszman. MPLUS: A probabilistic medical language understanding system. In *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain - Volume 3*, BioMed '02, pages 29–36, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[41] Anni Coden, Guergana Savova, Igor Sominsky, Michael Tanenblatt, James Masanz, Karin Schuler, James Cooper, Wei Guan, and Piet C. de Groen. Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model. *Journal of Biomedical Informatics*, 42(5):937–949, October 2009.

[42] Rebecca S. Crowley, Melissa Castine, Kevin Mitchell, Girish Chavan, Tara McSherry, and Michael Feldman. caTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. *Journal of the American Medical Informatics Association: JAMIA*, 17(3):253–264, June 2010.

[43] Licong Cui, Satya S. Sahoo, Samden D. Lhatoo, Gaurav Garg, Prashant Rai, Alireza Bozorgi, and Guo-Qiang Zhang. Complex epilepsy phenotype extraction from narrative

clinical discharge summaries. *Journal of Biomedical Informatics*, 51:272–279, October 2014.

[44] Hamish Cunningham, Diana Maynard, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Kalina Bontcheva, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. *Text Processing with GATE (Version 6)*. 2011.

[45] Kim N. Danforth, Megan I. Early, Sharon Ngan, Anne E. Kosco, Chengyi Zheng, and Michael K. Gould. Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing. *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer*, 7(8):1257–1262, August 2012.

[46] Leonard W. D'Avolio, Thien M. Nguyen, Wildon R. Farwell, Yongming Chen, Felicia Fitzmeyer, Owen M. Harris, and Louis D. Fiore. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). *Journal of the American Medical Informatics Association*, 17(4):375–382, July 2010.

[47] Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):557–562, 2011.

[48] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, pages 47–56, New York, NY, USA, 2013. ACM.

[49] Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772, October 2009.

[50] Renumathy Dhanasekaran, Alpna Limaye, and Roniel Cabrera. Hepatocellular carcinoma: current trends in worldwide epidemiology, risk factors, diagnosis, and therapeutics. *Hepatic Medicine : Evidence and Research*, 4:19–37, May 2012.

[51] Dmitriy Dligach, Steven Bethard, Lee Becker, Timothy Miller, and Guergana K. Savova. Discovering body site and severity modifiers in clinical texts. *Journal of the American Medical Informatics Association: JAMIA*, 21(3):448–454, June 2014.

[52] Son Doan, Mike Conway, Tu Minh Phuong, and Lucila Ohno-Machado. Natural language processing in biomedicine: a unified system architecture overview. *Methods in Molecular Biology (Clifton, N.J.)*, 1168:275–294, 2014.

[53] George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC*, volume 2, page 1, 2004.

[54] Molla Sloane Donaldson. An overview of to err is human: Re-emphasizing the message of patient safety. In Ronda G. Hughes, editor, *Patient Safety and Quality: An Evidence-Based Handbook for Nurses*, Advances in Patient Safety. Agency for Healthcare Research and Quality (US), Rockville (MD), 2008.

[55] Rezarta Islamaj Doan and Zhiyong Lu. An improved corpus of disease mentions in PubMed citations. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, BioNLP '12, pages 91–99, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[56] Tracy Edinger, Aaron M. Cohen, Steven Bedrick, Kyle Ambert, and William Hersh. Barriers to retrieving patient information from electronic health record data: Failure analysis from the TREC medical records track. *AMIA Annual Symposium Proceedings*, 2012:180–188, November 2012.

[57] Noémie Elhadad, Sameer Pradhan, WW Chapman, Suresh Manandhar, and GK Savova. Semeval-2015 task 14: Analysis of clinical text. In *Proc of Workshop on Semantic Evaluation. Association for Computational Linguistics*, pages 303–10, 2015.

[58] Mehdi Embarek and Olivier Ferret. Learning patterns for building resources about semantic relations in the medical domain. *LREC 2008*, 2008.

[59] W. K. Evans, J. Crook, D. Read, J. Morriss, and D. M. Logan. Capturing tumour stage in a cancer information database. *Cancer prevention and control: CPC = Prevention and controle en cancerologie: PCC*, 2(6):304–309, December 1998.

[60] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008.

[61] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[62] A. V. C. Frana, J. Elias Junior, B. L. G. Lima, A. L. C. Martinelli, and F. J. Carrilho. Diagnosis, staging and treatment of hepatocellular carcinoma. *Brazilian Journal of Medical and Biological Research = Revista Brasileira De Pesquisas Mdicas E Biolgicas / Sociedade Brasileira De Biofsica [et Al.]*, 37(11):1689–1705, November 2004.

[63] Jeff Friedlin and Clement J. McDonald. A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. *AMIA Annual Symposium Proceedings*, pages 269–273, 2006.

[64] Jeff Friedlin and Clement J. McDonald. Using a natural language processing system to extract and code family history data from admission reports. *AMIA Annual Symposium Proceedings*, page 925, 2006.

[65] Jeff Friedlin, Marc Overhage, Mohammed A. Al-Haddad, Joshua A. Waters, J. Juan R. Aguilar-Saavedra, Joe Kesterson, and Max Schmidt. Comparing methods for identifying pancreatic cancer patients using electronic data sources. *AMIA Annual Symposium Proceedings*, 2010:237–241, 2010.

[66] C. Friedman. A broad-coverage natural language processing system. *Proceedings of the AMIA Symposium*, pages 270–274, 2000.

[67] Hongyuan Gao, Erin J Aiello Bowles, David Carrell, and Diana SM Buist. Using natural language processing to extract mammographic findings. *Journal of biomedical informatics*, 54:77–84, 2015.

[68] Glenn T. Gobbel, Ruth Reeves, Shrimalini Jayaramaraja, Dario Giuse, Theodore Speroff, Steven H. Brown, Peter L. Elkin, and Michael E. Matheny. Development and evaluation of RapTAT: a machine learning system for concept mapping of phrases from medical narratives. *Journal of Biomedical Informatics*, 48:54–65, April 2014.

[69] Omri Gottesman, Helena Kuivaniemi, Gerard Tromp, W Andrew Faucett, Rongling Li, Teri A Manolio, Saskia C Sanderson, Joseph Kannry, Randi Zinberg, Melissa A Basford, et al. The electronic medical records and genomics (emerge) network: past, present, and future. *Genetics in Medicine*, 15(10):761–771, 2013.

[70] A. Grieco, M. Pompili, G. Caminiti, L. Miele, M. Covino, B. Alfei, G. L. Rapaccini, and G. Gasbarrini. Prognostic factors for survival in patients with early-intermediate hepatocellular carcinoma undergoing non-surgical therapy: comparison of okuda, CLIP, and BCLC staging systems in a single italian centre. *Gut*, 54(3):411–418, March 2005.

[71] Ralph Grishman and Beth Sundheim. Message understanding conference-6: A brief history. In *COLING*, volume 96, pages 466–471, 1996.

[72] Udo Hahn, Martin Romacker, and Stefan Schulz. MEDSYNDIKATE–a natural language system for the extraction of medical information from findings reports. *International Journal of Medical Informatics*, 67(1-3):63–74, December 2002.

[73] 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. Asian consensus workshop report: expert consensus guideline for the management of intermediate and advanced hepatocellular carcinoma in asia. *Oncology*, 81 Suppl 1:158–164, 2011.

[74] Saeed Hassanpour and Curtis P. Langlotz. Information extraction from multi-institutional radiology reports. *Artificial Intelligence in Medicine*, October 2015.

[75] P. J. Haug, S. Koehler, L. M. Lau, P. Wang, R. Rocha, and S. M. Huff. Experience with a mixed semantic/syntactic parser. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 284–288, 1995.

[76] Brian Hazlehurst, H. Robert Frost, Dean F. Sittig, and Victor J. Stevens. MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. *Journal of the American Medical Informatics Association: JAMIA*, 12(5):517–529, October 2005.

[77] George Hripcsak and Adam S. Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298, May 2005.

[78] Phil Hughes. Python and tkinter programming. *Linux J.*, 2000(77es), September 2000.

[79] Fidel-David Huitzil-Melendez, Marinela Capanu, Eileen M. O'Reilly, Austin Duffy, Bolorsukh Gansukh, Leonard L. Saltz, and Ghassan K. Abou-Alfa. Advanced hepatocellular carcinoma: which staging systems best predict prognosis? *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 28(17):2889–2895, June 2010.

[80] Timothy D. Imler, Justin Morea, Charles Kahi, and Thomas F. Imperiale. Natural language processing accurately categorizes findings from colonoscopy and pathology reports. *Clinical Gastroenterology and Hepatology: The Official Clinical Practice Journal of the American Gastroenterological Association*, 11(6):689–694, June 2013.

[81] Institute of Medicine (US) Forum on Drug Discovery, Development, and Translation. *Transforming Clinical Research in the United States: Challenges and Opportunities: Workshop Summary*. The National Academies Collection: Reports funded by National Institutes of Health. National Academies Press (US), Washington (DC), 2010.

[82] N. L. Jain and C. Friedman. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. *AMIA Annual Symposium Proceedings*, pages 829–833, 1997.

[83] Mukund Jha and Nomie Elhadad. Cancer stage prediction based on patient online discourse. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, BioNLP '10, pages 64–71, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[84] Min Jiang, Yukun Chen, Mei Liu, S Trent Rosenbloom, Subramani Mani, Joshua C Denny, and Hua Xu. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):601–606, 2011.

[85] Prateek Jindal and Dan Roth. Extraction of events and temporal expressions from clinical narratives. *Journal of Biomedical Informatics*, 46, Supplement:S13–S19, December 2013.

[86] Clement Jonquet, Nigam H. Shah, and Mark A. Musen. The open biomedical annotator. *Summit Trans Bioinformatics*, pages 56–60, 2009.

[87] V. Jouhet, G. Defossez, A. Burgun, P. le Beux, P. Levillain, P. Ingrand, and V. Claveau. Automated classification of free-text pathology reports for registration of incident cases of cancer. *Methods of Information in Medicine*, 51(3):242–251, 2012.

[88] Ramakanth Kavuluru, Isaac Hands, Eric B. Durbin, and Lisa Witt. Automatic extraction of ICD-o-3 primary sites from cancer pathology reports. *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science*, 2013:112–116, 2013.

[89] Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Tobias Schreck, Gondy Leroy, Danielle L Mowery, Sumithra Velupillai, Wendy W Chapman, David Martinez, Guido Zuccon, et al. *Overview of the share/clef ehealth evaluation lab 2014*. Springer, 2014.

[90] Jin-Dong Kim, Ngan Nguyen, Yue Wang, Jun'ichi Tsujii, Toshihisa Takagi, and Akinori Yonezawa. The Genia Event and Protein Coreference tasks of the BioNLP Shared Task 2011. *BMC Bioinformatics*, 13(11):1–12, 2012.

[91] Jacqueline C. Kirby, Peter Speltz, Luke V. Rasmussen, Melissa Basford, Omri Gottesman, Peggy L. Peissig, Jennifer A. Pacheco, Gerard Tromp, Jyotishman Pathak, David S. Carrell, Stephen B. Ellis, Todd Lingren, Will K. Thompson, Guergana Savova, Jonathan Haines, Dan M. Roden, Paul A. Harris, and Joshua C. Denny. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *Journal of the American Medical Informatics Association: JAMIA*, March 2016.

[92] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[93] Kenton Lee, Yoav Artzi, Jesse Dodge, and Luke Zettlemoyer. Context-dependent semantic parsing for time expressions. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2014.

[94] Jianbo Lei, Buzhou Tang, Xueqin Lu, Kaihua Gao, Min Jiang, and Hua Xu. A comprehensive study of named entity recognition in chinese clinical text. *Journal of the American Medical Informatics Association*, 21(5):808–814, September 2014.

[95] Katherine P. Liao, Tianxi Cai, Vivian Gainer, Sergey Goryachev, Qing Zeng-Treitler, Soumya Raychaudhuri, Pete Szolovits, Susanne Churchill, Shawn Murphy, Isaac Kohane, Elizabeth W. Karlson, and Robert M. Plenge. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis care & research*, 62(8):1120–1127, August 2010.

[96] Chen Lin, Elizabeth W Karlson, Dmitriy Dligach, Monica P Ramirez, Timothy A Miller, Huan Mo, Natalie S Braggs, Andrew Cagan, Vivian Gainer, Joshua C Denny, et al. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. *Journal of the American Medical Informatics Association*, pages amiajnl–2014, 2014.

[97] Kaihong Liu, Kevin J. Mitchell, Wendy W. Chapman, and Rebecca S. Crowley. Automating tissue bank annotation from pathology reports - comparison to a gold standard expert annotation set. *AMIA Annual Symposium Proceedings*, pages 460–464, 2005.

[98] Mei Liu, Anushi Shah, Min Jiang, Neeraja B. Peterson, Qi Dai, Melinda C. Aldrich, Qingxia Chen, Erica A. Bowton, Hongfang Liu, Joshua C. Denny, and Hua Xu. A Study of Transportability of an Existing Smoking Status Detection Module across Institutions. *AMIA Annual Symposium Proceedings*, 2012:577–586, November 2012.

[99] W. L. Liu, S. Kasl, J. T. Flannery, A. Lindo, and R. Dubrow. The accuracy of prostate cancer staging in a population-based tumor registry and its impact on the black-white stage difference (connecticut, united states). *Cancer causes & control: CCC*, 6(5):425–430, September 1995.

[100] Edward Loper and Steven Bird. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[101] Xiaoqiang Luo. On Coreference Resolution Performance Metrics. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 25–32, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[102] Clement Ma, Shazeen Bandukwala, Debika Burman, John Bryson, Dori Seccareccia, Subrata Banerjee, Jeff Myers, Gary Rodin, Deborah Dudgeon, and Camilla Zimmermann. Interconversion of three measures of performance status: an empirical analysis. *European Journal of Cancer (Oxford, England: 1990)*, 46(18):3175–3183, December 2010.

[103] Diana Lynn MacLean and Jeffrey Heer. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. *Journal of the American Medical Informatics Association : JAMIA*, 20(6):1120–1127, November 2013.

[104] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[105] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The stanford CoreNLP natural language processing toolkit. i. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[106] Laure Martin, Battistelli Delphine, and Charnois Thierry. Symptom recognition issue. *Proceedings of the 2014 Workshop on Biomedical Natural Language Processing*, pages 107–111, June 2014.

[107] David Martinez, Lawrence Cavedon, and Graham Pitson. Stability of text mining techniques for identifying cancer staging. Canberra, Australia, 2013.

[108] David Martinez and Yue Li. Information extraction from pathology reports in a hospital setting. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1877–1882, New York, NY, USA, 2011. ACM.

[109] Michael E. Matheny, Fern FitzHenry, Theodore Speroff, Jennifer K. Green, Michelle L. Griffith, Eduard E. Vasilevskis, Elliot M. Fielstein, Peter L. Elkin, and Steven H. Brown. Detection of infectious symptoms from VA emergency department and primary

care clinical documentation. *International Journal of Medical Informatics*, 81(3):143–156, March 2012.

[110] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[111] Andrew Kachites McCallum. MALLET: A machine learning for language toolkit. `http://mallet.cs.umass.edu`, 2002.

[112] Iain A. McCowan, Darren C. Moore, Anthony N. Nguyen, Rayleen V. Bowman, Belinda E. Clarke, Edwina E. Duhig, and Mary-Jane Fry. Collection of cancer stage data by classifying free-text medical reports. *Journal of the American Medical Informatics Association: JAMIA*, 14(6):736–745, December 2007.

[113] Katherine A. McGlynn and W. Thomas London. The global epidemiology of hepatocellular carcinoma: present and future. *Clinics in Liver Disease*, 15(2):223–243, vii–x, May 2011.

[114] Eugenia R. McPeek Hinz, Lisa Bastarache, and Joshua C. Denny. A natural language processing algorithm to define a venous thromboembolism phenotype. *AMIA Annual Symposium Proceedings*, 2013:975–983, 2013.

[115] Saeed Mehrabi, C. Max Schmidt, Joshua A. Waters, Chris Beesley, Anand Krishnan, Joe Kesterson, Paul Dexter, Mohammed A. Al-Haddad, William M. Tierney, and Mathew Palakal. An efficient pancreatic cyst identification methodology using natural language processing. *Studies in Health Technology and Informatics*, 192:822–826, 2013.

[116] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of Medical Informatics*, pages 128–144, 2008.

[117] P. Mitra, S. Mitra, and S. K. Pal. Staging of cervical cancer with soft computing. *IEEE transactions on bio-medical engineering*, 47(7):934–940, July 2000.

[118] Sungrim Moon, Bridget McInnes, and Genevieve B. Melton. Challenges and practical approaches with word sense disambiguation of acronyms and abbreviations in the clinical domain. *Healthcare Informatics Research*, 21(1):35–42, January 2015.

[119] Harvey J. Murff, Fern FitzHenry, Michael E. Matheny, Nancy Gentry, Kristen L. Kotter, Kimberly Crimin, Robert S. Dittus, Amy K. Rosen, Peter L. Elkin, Steven H. Brown, and Theodore Speroff. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA*, 306(8):848–855, August 2011.

[120] Shawn N. Murphy, Griffin Weber, Michael Mendis, Vivian Gainer, Henry C. Chueh, Susanne Churchill, and Isaac Kohane. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, 17(2):124–130, March 2010.

[121] Anthony Nguyen, Darren Moore, Iain McCowan, and Mary-Jane Courage. Multi-class classification of cancer stages from free-text histology reports using support vector machines. *Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2007:5140–5143, 2007.

[122] Anthony N. Nguyen, Michael J. Lawley, David P. Hansen, Rayleen V. Bowman, Belinda E. Clarke, Edwina E. Duhig, and Shoni Colquist. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *Journal of the American Medical Informatics Association: JAMIA*, 17(4):440–445, August 2010.

[123] Tiago Nunes, David Campos, Srgio Matos, and Jos Lus Oliveira. BeCAS: biomedical concept recognition services and visualization. *Bioinformatics (Oxford, England)*, 29(15):1915–1916, August 2013.

[124] U.S. National Institutes of Health. National Cancer Institute: NCIthesaurus. `http://ncit.nci.nih.gov/`.

[125] United States Department of Veterans Affairs. VA child-turcotte-pugh (CTP) calculator. `http://www.hepatitis.va.gov/provider/tools/child-pugh-calculator.asp`.

[126] Naoaki Okazaki. CRFsuite: a fast implementation of Conditional Random Fields (CRFs). 2007.

[127] Ying Ou and Jon Patrick. Automatic population of structured reports from narrative pathology reports. In *Proceedings of the Seventh Australasian Workshop on Health Informatics and Knowledge Management - Volume 153*, HIKM '14, pages 41–50, Darlinghurst, Australia, Australia, 2014. Australian Computer Society, Inc.

[128] Jon Patrick and Min Li. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association: JAMIA*, 17(5):524–527, October 2010.

[129] Michael J. Paul and Mark Dredze. *You Are What You Tweet: Analyzing Twitter for Public Health*.

[130] Peggy L Peissig, Luke V Rasmussen, Richard L Berg, James G Linneman, Catherine A McCarty, Carol Waudby, Lin Chen, Joshua C Denny, Russell A Wilke, Jyotishman Pathak, David Carrell, Abel N Kho, and Justin B Starren. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. *Journal of the American Medical Informatics Association : JAMIA*, 19(2):225–234, 2012.

[131] John P. Pestian, Christopher Brew, Pawe\l Matykiewicz, D. J. Hovermale, Neil Johnson, K. Bretonnel Cohen, and W\lodzis\law Duch. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, BioNLP '07, pages 97–104, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[132] John P. Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K. Bretonnel Cohen, John Hurdle, and Christopher Brew. Sentiment Analysis of Suicide Notes: A Shared Task. *Biomedical Informatics Insights*, 5(Suppl 1):3–16, January 2012.

[133] Anne-Dominique Pham, Aurlie Nvol, Thomas Lavergne, Daisuke Yasunaga, Olivier Clment, Guy Meyer, Rmy Morello, and Anita Burgun. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. *BMC Bioinformatics*, 15(1):266, August 2014.

[134] P. Phinjaroenphan and S. Bevinakoppa. Automated prognostic tool for cervical cancer patient database. In *Proceedings of International Conference on Intelligent Sensing and Information Processing, 2004*, pages 63–66, 2004.

[135] Xiao-Ou Ping, Yi-Ju Tseng, Yufang Chung, Ya-Lin Wu, Ching-Wei Hsu, Pei-Ming Yang, Guan-Tarn Huang, Feipei Lai, and Ja-Der Liang. Information extraction for tracking liver cancer patients' statuses: from mixture of clinical narrative report types. *Telemedicine Journal and E-Health: The Official Journal of the American Telemedicine Association*, 19(9):704–710, September 2013.

[136] Fernando Pons, Maria Varela, and Josep M. Llovet. Staging systems in hepatocellular carcinoma. *HPB : The Official Journal of the International Hepato Pancreato Biliary Association*, 7(1):35–41, 2005.

[137] Sameer Pradhan, Nomie Elhadad, Wendy W Chapman, Suresh Manandhar, and Guergana Savova. SemEval-2014 task 7: Analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 54–62, Dublin, Ireland, August 2014.

[138] Sameer Pradhan, Nomie Elhadad, Brett R. South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman, and Guergana Savova. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association*, 22(1):143–154, January 2015.

[139] Dominik Pus, Nicolas Newcomb, and Silvia Hofer. Appraisal of the Karnofsky Performance Status and proposal of a simple algorithmic system for its evaluation. *BMC medical informatics and decision making*, 13:72, 2013.

[140] A. L. Rector, J. E. Rogers, P. E. Zanstra, and E. van der Haring. OpenGALEN: Open source medical terminology and tools. *AMIA Annual Symposium Proceedings*, 2003:982, 2003.

[141] Angus Roberts, Robert Gaizauskas, Mark Hepple, Neil Davis, George Demetriou, Yikun Guo, Jay (Subbarao) Kola, Ian Roberts, Andrea Setzer, Archana Tapuria, and Bill Wheeldin. The CLEF corpus: Semantic annotation of clinical text. *AMIA Annual Symposium Proceedings*, 2007:625–629, 2007.

[142] Kirk Roberts, Bryan Rink, Sanda M. Harabagiu, Richard H. Scheuermann, Seth Toomay, Travis Browning, Teresa Bosler, and Ronald Peshock. A machine learning approach for identifying anatomical locations of actionable findings in radiology reports. *Proceedings of the AMIA Symposium*, 2012:779–788, 2012.

[143] Cornelius Rosse and Jos L. V. Mejino Jr. The Foundational Model of Anatomy Ontology. In Albert Burger BSc MSc, Duncan Davidson BSc, and Richard Baldock BSc, editors, *Anatomy Ontologies for Bioinformatics*, number 6 in Computational Biology, pages 59–117. Springer London, 2008. DOI: 10.1007/978-1-84628-885-2_4.

[144] Patrick Ruch, Julien Gobeill, Christian Lovis, and Antoine Geissbhler. Automatic medical encoding with SNOMED categories. *BMC Medical Informatics and Decision Making*, 8(Suppl 1):S6, October 2008.

[145] Yvonne Sada, Jason Hou, Peter Richardson, Hashem El-Serag, and Jessica Davila. Validation of case finding algorithms for hepatocellular cancer from administrative data and electronic health records using natural language processing. *Medical Care*, August 2013.

[146] N. Sager, M. Lyman, C. Bucknall, N. Nhan, and L. J. Tick. Natural language processing and the representation of clinical data. *Journal of the American Medical Informatics Association: JAMIA*, 1(2):142–160, April 1994.

[147] Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association: JAMIA*, 17(5):507–513, October 2010.

[148] L. J. Schouten, J. A. Langendijk, J. J. Jager, and P. A. van den Brandt. Validity of the stage of lung cancer in records of the maastricht cancer registry, the netherlands. *Lung Cancer (Amsterdam, Netherlands)*, 17(1):115–122, May 1997.

[149] Karin Kipper Schuler. VerbNet: A broad-coverage, comprehensive verb lexicon. *Dissertations available from ProQuest*, pages 1–146, January 2005.

[150] Merlijn Sevenster, Jeffrey Bozeman, Andrea Cowhy, and William Trost. A natural language processing pipeline for pairing measurements uniquely across free-text CT reports. *Journal of Biomedical Informatics*, 53:36–48, February 2015.

[151] Tracy Sexton, George Rodrigues, Ed Brecevic, Laura Boyce, Denise Parrack, Michael Lock, and David D'Souza. Controversies in prostate cancer staging implementation at a tertiary cancer center. *The Canadian Journal of Urology*, 13(6):3327–3334, December 2006.

[152] Nigam H. Shah, Nipun Bhatia, Clement Jonquet, Daniel Rubin, Annie P. Chiang, and Mark A. Musen. Comparison of concept recognizers for building the open biomedical annotator. *BMC Bioinformatics*, 10(Suppl 9):S14, September 2009.

[153] Mohamed I. F. Shariff, I. Jane Cox, Asmaa I. Gomaa, Shahid A. Khan, Wladyslaw Gedroyc, and Simon D. Taylor-Robinson. Hepatocellular carcinoma: current trends in worldwide epidemiology, risk factors, diagnosis and therapeutics. *Expert Review of Gastroenterology & Hepatology*, 3(4):353–367, August 2009.

[154] Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter J. Embi, Noemie Elhadad, Stephen B. Johnson, and Albert M. Lai. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association: JAMIA*, 21(2):221–230, April 2014.

[155] Yongyut Sirivatanauksorn and Chutwichai Tovikkai. Comparison of staging systems of hepatocellular carcinoma. *HPB Surgery*, 2011:e818217, June 2011.

[156] American Cancer Society. Cancer facts & figures 2015. Technical report, Atlanta, 2014.

[157] Sunghwan Sohn, Cheryl Clark, Scott R. Halgrim, Sean P. Murphy, Christopher G. Chute, and Hongfang Liu. : an open source medication extraction and normalization tool for clinical text. *Journal of the American Medical Informatics Association: JAMIA*, 21(5):858–865, October 2014.

[158] Roderick Y. Son, Ricky K. Taira, and Hooshang Kangarloo. Inter-document coreference resolution of abnormal findings in radiology documents. *Studies in Health Technology and Informatics*, 107(Pt 2):1388–1392, 2004.

[159] Irena Spasi, Jacqueline Livsey, John A. Keane, and Goran Nenadi. Text mining of cancer-related information: review of current status and future directions. *International Journal of Medical Informatics*, 83(9):605–623, September 2014.

[160] Irena Spasi, Bo Zhao, Christopher B. Jones, and Kate Button. KneeTex: an ontologydriven system for information extraction from MRI reports. *Journal of Biomedical Semantics*, 6, September 2015.

[161] P. Spyns. Natural language processing in medicine: an overview. *Methods of Information in Medicine*, 35(4-5):285–301, December 1996.

[162] Sylvie Stacy, Omar Hyder, David Cosgrove, Joseph M. Herman, Ihab Kamel, Jean-Francois H. Geschwind, Ahmet Gurakar, Robert Anders, Andrew Cameron, and Timothy M. Pawlik. Patterns of consultation and treatment of patients with hepatocellular carcinoma presenting to a large academic medical center in the US. *Journal of Gastrointestinal Surgery: Official Journal of the Society for Surgery of the Alimentary Tract*, 17(9):1600–1608, September 2013.

[163] Pontus Stenetorp, Sampo Pyysalo, Goran Topi, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. BRAT: A web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[164] Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. Conundrums in Noun Phrase Coreference Resolution: Making Sense of the State-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 656–664, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[165] Justin A. Strauss, Chun R. Chao, Marilyn L. Kwan, Syed A. Ahmed, Joanne E. Schottinger, and Virginia P. Quinn. Identifying primary and recurrent cancers using a

SAS-based natural language processing algorithm. *Journal of the American Medical Informatics Association: JAMIA*, 20(2):349–355, April 2013.

[166] Jannik Strtgen and Michael Gertz. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 321–324, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[167] Amber Stubbs, Christopher Kotfila, Hua Xu, and zlem Uzuner. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. *Journal of Biomedical Informatics*, 58, Supplement:S67–S77, December 2015.

[168] Amber Stubbs and Ozlem Uzuner. Annotating risk factors for heart disease in clinical narratives for diabetic patients. *Journal of Biomedical Informatics*.

[169] William Styler, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet de Groen, Brad Erickson, Timothy Miller, Lin Chen, Guergana Savova, and James Pustejovsky. Temporal annotations in the clinical domain. In *Transactions of the Association for Computational Linguistics.*, 2014.

[170] Somasundaram Subramaniam, Robin K. Kelley, and Alan P. Venook. A review of hepatocellular carcinoma (HCC) staging systems. *Chinese Clinical Oncology*, 2(4), August 2013.

[171] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association: JAMIA*, 20(5):806–813, October 2013.

[172] Hanna Suominen, Sanna Salanter, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J. F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot, David Martinez, and Guido Zuccon. Overview of the ShARe/CLEF eHealth evaluation lab 2013. In Pamela Forner, Henning Mller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, number 8138 in Lecture Notes in Computer Science, pages 212–231. Springer Berlin Heidelberg, 2013.

[173] R. K. Taira, S. G. Soderland, and R. M. Jakobovits. Automatic structuring of radiology free-text reports. *Radiographics: A Review Publication of the Radiological Society of North America, Inc*, 21(1):237–245, February 2001.

[174] Michael Tepper, Daniel Capurro, Fei Xia, Lucy Vanderwende, and Meliha Yestisgen-Yildiz. Statistical Section Segmentation in Free-Text Clinical Records. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Istanbul, May 2012.

[175] Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association : JAMIA*, 19(5):786–791, 2012.

[176] Ozlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):552–556, 2011.

[177] J. van der Lei. Closing the loop between clinical practice, research, and education: the potential of electronic patient records. *Methods of Information in Medicine*, 41(1):51–54, 2002.

[178] Nynke van Dijk, Lotty Hooft, and Margreet Wieringa-de Waard. What are the barriers to residents' practicing evidence-based medicine? a systematic review. *Academic Medicine: Journal of the Association of American Medical Colleges*, 85(7):1163–1170, July 2010.

[179] Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*, pages 45–52, 1995.

[180] Chunye Wang and Ramakrishna Akella. A Hybrid Approach to Extracting Disorder Mentions from Clinical Notes. 2015.

[181] Hui Wang, Weide Zhang, Qiang Zeng, Zuofeng Li, Kaiyan Feng, and Lei Liu. Extracting important information from chinese operation notes with natural language processing methods. *Journal of Biomedical Informatics*, 48:130–136, April 2014.

[182] Jeremy L. Warner, Peter Anick, Pengyu Hong, and Nianwen Xue. Natural language processing and the oncologic history: is there a match? *Journal of Oncology Practice / American Society of Clinical Oncology*, 7(4):e15–19, July 2011.

[183] Richard A. Wilson, Wendy W. Chapman, Shawn J. Defries, Michael J. Becich, and Brian E. Chapman. Automated ancillary cancer history classification for mesothelioma patients from free-text clinical reports. *Journal of Pathology Informatics*, 1:24, 2010.

[184] Andrew Worster and Ted Haines. Advanced statistics: Understanding medical record review (MRR) studies. *Academic Emergency Medicine*, 11(2):187–192, 2004.

[185] Stephen Wu, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, and Cheryl Clark. Negation's not solved: generalizability versus optimizability in clinical natural language processing. *PloS One*, 9(11):e112774, 2014.

[186] Yonghui Wu, Jianbo Lei, Wei-Qi Wei, Buzhou Tang, Joshua C. Denny, S. Trent Rosenbloom, Randolph A. Miller, Dario A. Giuse, Kai Zheng, and Hua Xu. Analyzing differences between chinese and english clinical text: a cross-institution comparison of discharge summaries in two languages. *Studies in Health Technology and Informatics*, 192:662–666, 2013.

[187] Hua Xu, Zhenming Fu, Anushi Shah, Yukun Chen, Neeraja B. Peterson, Qingxia Chen, Subramani Mani, Mia A. Levy, Qi Dai, and Josh C. Denny. Extracting and Integrating Data from Entire Electronic Health Records for Detecting Colorectal Cancer Cases. *AMIA Annual Symposium Proceedings*, 2011:1564–1572, 2011.

[188] Hua Xu, Shane P. Stenner, Son Doan, Kevin B. Johnson, Lemuel R. Waitman, and Joshua C. Denny. MedEx: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association: JAMIA*, 17(1):19–24, February 2010.

[189] Yan Xu, Kai Hong, Junichi Tsujii, and Eric I.-Chao Chang. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *Journal of the American Medical Informatics Association: JAMIA*, 19(5):824–832, October 2012.

[190] Yan Xu, Ji Hua, Zhaoheng Ni, Qinlang Chen, Yubo Fan, Sophia Ananiadou, Eric I.-Chao Chang, and Junichi Tsujii. Anatomical entity recognition with a hierarchical framework augmented by external resources. *PloS One*, 9(10):e108396, 2014.

[191] Ju Dong Yang and Lewis R. Roberts. Epidemiology and management of hepatocellular carcinoma. *Infectious Disease Clinics of North America*, 24(4):899–919, viii, December 2010.

[192] Ju Dong Yang and Lewis R. Roberts. Hepatocellular carcinoma: A global view. *Nature Reviews. Gastroenterology & Hepatology*, 7(8):448–458, August 2010.

[193] Jonathan C. Yau, Arlene Chan, Tamina Eapen, Keith Oirourke, and Libni Eapen. Accuracy of the oncology patients information system in a regional cancer centre. *Oncology Reports*, 9(1):167–169, February 2002.

[194] Wen-wai Yim, Tyler Denman, Sharon Kwan, and Meliha Yetisgen. Tumor information extraction in radiology reports for hepatocellular carcinoma patients. In *Proceedings of AMIA 2016 Joint Summits on Translational Science.*, San Francisco, USA, March 2016.

[195] Wen-wai Yim, Sharon Kwan, and Meliha Yetisgen. Classifying tumor event attributes in radiology reports. *Journal of Biomedical Informatics (UNDER REVIEW)*, 2016.

[196] Wen-wai Yim, Sharon Kwan, and Meliha Yetisgen. Tumor reference resolution and characteristic extraction in radiology reports for liver cancer stage prediction. *Journal of Biomedical Informatics (UNDER REVIEW)*, 2016.

[197] Wen-Wai Yim, Meliha Yetisgen, William P. Harris, and Sharon W. Kwan. Natural Language Processing in Oncology: A Review. *JAMA oncology*, April 2016.

[198] Qing T. Zeng, Sergey Goryachev, Scott Weiss, Margarita Sordo, Shawn N. Murphy, and Ross Lazarus. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC medical informatics and decision making*, 6:30, 2006.

[199] Shaodian Zhang and Nomie Elhadad. Unsupervised biomedical named entity recognition: experiments with clinical and biological texts. *Journal of Biomedical Informatics*, 46(6):1088–1098, December 2013.

[200] Li Zhou, Joseph M Plasek, Lisa M Mahoney, Neelima Karipineni, Frank Chang, Xuemin Yan, Fenny Chang, Dana Dimaggio, Debora S. Goldman, and Roberto A. Rocha. Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to Process Medication Information in Outpatient Clinical Notes. *AMIA Annual Symposium Proceedings*, 2011:1639–1648, 2011.

[201] Camilla Zimmermann, Debika Burman, Shazeen Bandukwala, Dori Seccareccia, Ebru Kaya, John Bryson, Gary Rodin, and Christopher Lo. Nurse and physician inter-rater agreement of three performance status measures in palliative care outpatients. *Supportive Care in Cancer: Official Journal of the Multinational Association of Supportive Care in Cancer*, 18(5):609–616, May 2010.

[202] Qinghua Zou, Wesley W. Chu, Craig Morioka, Gregory H. Leazer, and Hooshang Kangarloo. : A Method of Extracting Key Concepts from Clinical Texts for Indexing. *AMIA Annual Symposium Proceedings*, 2003:763–767, 2003.

[203] Sandra Zwolsman, Ellen te Pas, Lotty Hooft, Margreet Wieringa-de Waard, and Nynke van Dijk. Barriers to GPs' use of evidence-based medicine: a systematic review. *British Journal of General Practice*, 62(600):511–521, July 2012.

[204] Sandra Zwolsman, Ellen te Pas, Lotty Hooft, Margreet Wieringa-de Waard, and Nynke van Dijk. Barriers to GPs' use of evidence-based medicine: a systematic review. *The British Journal of General Practice: The Journal of the Royal College of General Practitioners*, 62(600):e511–521, July 2012.

Appendix A

TEXT ANNOTATION GUIDELINES

GENERAL COMMENTS

In general, annotate text span that gives the relevant information. If it is unclear, then mark the entire sentence
If not text is marked for a specific parameter, the default will be the first value on these lists
If you encounter a report or note that is not relevant and should be excluded, mark the first word with Parameter "Exclude file"="Yes"
Examples: Addenda without new useful information, Abbreviated notes with no additional useful information, radiology reports from after treatment
If you encounter a patient that is not relevant and should be excluded, mark the first word of the first file with Parameter "Exclude patient"="Yes"
Examples: Patients with other types of cancer, patients without cancer

STEP 1

Start with Radiology reports, if available
If multiple studies available, annotate all diagnostic studies available as long as it is pre-treatment
If a study is not diagnostic or post-treatment, then tag the first word in the report with parameter "Exclude file"="Yes"
Start with Impression section. If information not available in Impression, use Findings section.

Annotate for:	Parameter	Values	Notes
	Tumor number	Single	Mark in Impression section, if data available there
		2-3	Otherwise mark in Findings section
		>3	Mark enough text to allow identification of tumor(s) location in the liver (e.g. vs a lung metastasis)
	Tumor size	<3 cm	Mark in Impression section, if data available there
		3-5 cm	Else mark in Findings section
		>5 cm	Mark enough text to allow identification of tumor(s) location in the liver (e.g. vs a lung metastasis)
	Tumor morphology	Uninodular and extension <50% of liver	Highlight any text that gives relevant clues-e.g. "single lesion" "multiple tumors"
		Multinodular and extension <50% of liver	If more than one tumor, should be multinodular
		Massive or extension >= 50% of liver	
	Macrovascular invasion	No	Mark in Impression section, if data available there
		Yes-minor branch	Otherwise mark in Findings section
		Yes-major branch	Major branch defined as anything larger than left or right PV or left, middle, or right HV
	Extrahepatic invasion	No	Defined as direct invasion of an adjacent organ other than gallbladder or perforation of visceral peritoneum
		Yes	
	Metastasis	No	Often will be same text as for extrahepatic invasion
		Yes-regional lymph nodes	
		Yes-distal	
	Ascites	None	Mark relevant text and score literally; if clinic note give additional information, mark separately and score accordingly
		Mild-Suppressed on medications	
		Moderate-Severe/Refractory	
	Portal hypertension	No	Mark any text that may suggest its presence, such as ascites, varices, splenomegaly
		Yes	

Figure A.1: Text annotation guidelines (part1).

STEP 2

Annotate clinic notes

If parameters above not found in radiology report, then annotate the above parameters using information from the clinic notes, if can be found
One exception is ascites- where if additional information affects your final score, that should be annotated also
Start in HPI. If information not available there, then annotate whereever it is found elsewhere in the clinic note
If same information is repeated multiple times in clinc note, mark the first instance (exception being ECOG, where one descriptor and score should both be marked).
If multiple pieces of information all give clue to value of a parameter (e.g. presence of portal hypertension), mark them all.

Parameter	Values	Notes
Child-Pugh Class	A	This may not be directly available from clinic notes. If not, it will be calculated from labs and other parameters
	B	If clinic note is equivocal (e.g. "CP score is between B or C", do not mark
	C	
ECOG performance status	0	If a single person's clinic note gives a range, use the lower value
	1	If clinic notes don't give specific score, mark text that gives clues to score
	2	If clinic note gives descriptives AND specific score, mark both
	>=3	If two notes conflict on score, annotate all scores and give appropriate respective score, but when assigning final score during staging by consensus, this will be deemed unscorable
Hepatic encephalopathy	None	
	Mild/Grade 1-2/Suppressed	If on medications for HE and this affects your score, annotate medication (name only, do not include dose)
	Severe/Grade 3-4/Refractory	
Ascites	None	If clinic note gives additional information that changes score from that using radiology report, then annotate and score based on additional information
	Mild-Suppressed on medications	If patient is post-TIPS and well controlled, score as mild.
	Moderate-Severe/Refractory	If patient has hydrothorax, score as mild or moderate (depending on symptoms)

Figure A.2: Text annotation guidelines (part2).

Appendix B

STAGE ANNOTATION LOOKUP TABLES

AJCC STAGE					
Tumor number	Tumor size	Macrovascular invasion	Extrahepatic invasion	Metastasis	AJCC Stage
ANY	ANY	ANY	ANY	Yes-distal	IV-b
ANY	ANY	ANY	ANY	Yes-regional lymph nodes	IV-a
ANY	ANY	ANY	Yes	No	III-c
ANY	ANY	Yes-major branch	No	No	III-b
2-3	>5	No	No	No	III-a
2-3	>5	Yes-minor branch	No	No	III-a
>3	>5	No	No	No	III-a
>3	>5	Yes-minor branch	No	No	III-a
>3	3-5	No	No	No	II
ANY	3-5	Yes-minor branch	No	No	II
ANY	<3	Yes-minor branch	No	No	II
2-3	3-5	No	No	No	II
2-3	<3	No	No	No	II
2-3	3-5	No	No	No	II
>3	<3	No	No	No	II
Single	ANY	No	No	No	I
Single	>5	Yes-minor branch	No	No	II

Figure B.1: AJCC lookup table

BCLC STAGE								
ECO	Child Pu	Tumor numb	Tumor s	Macrovascular invasi	Metasta	Portal hypertensi	Bilirub	BCLC Sta
>=3	ANY	ANY	ANY	ANY	ANY	ANY	ANY	D
ANY	C	ANY	ANY	ANY	ANY	ANY	ANY	D
2	A	ANY	ANY	ANY	ANY	ANY	ANY	C
2	B	ANY	ANY	ANY	ANY	ANY	ANY	C
1	A	ANY	ANY	ANY	ANY	ANY	ANY	C
1	B	ANY	ANY	ANY	ANY	ANY	ANY	C
0	A	ANY	ANY	ANY	Yes-regional lymph nodes	ANY	ANY	C
0	B	ANY	ANY	ANY	Yes-regional lymph nodes	ANY	ANY	C
0	A	ANY	ANY	ANY	Yes-distal	ANY	ANY	C
0	B	ANY	ANY	ANY	Yes-distal	ANY	ANY	C
0	A	ANY	ANY	Yes	ANY	ANY	ANY	C
0	B	ANY	ANY	Yes	ANY	ANY	ANY	C
0	A	>3	ANY	No	No	ANY	ANY	B
0	B	>3	ANY	No	No	ANY	ANY	B
0	A	Single	>5cm	No	No	ANY	ANY	B
0	B	Single	>5cm	No	No	ANY	ANY	B
0	A	2-3	3-5cm	No	No	ANY	ANY	B
0	B	2-3	3-5cm	No	No	ANY	ANY	B
0	A	2-3	<3cm	No	No	ANY	ANY	A4
0	B	2-3	<3cm	No	No	ANY	ANY	A4
0	A	Single	<3cm	No	No	Yes	Abnormal	A3
0	A	Single	3-5cm	No	No	Yes	Abnormal	A3
0	B	Single	<3cm	No	No	Yes	Abnormal	A3
0	B	Single	3-5cm	No	No	Yes	Abnormal	A3
0	A	Single	<3cm	No	No	Yes	Normal	A2
0	A	Single	3-5cm	No	No	Yes	Normal	A2
0	B	Single	<3cm	No	No	Yes	Normal	A2
0	B	Single	3-5cm	No	No	Yes	Normal	A2
0	A	Single	<3cm	No	No	No	Normal	A1
0	A	Single	3-5cm	No	No	No	Normal	A1
0	A	Single	<3cm	No	No	No	Abnormal	A1
0	A	Single	3-5cm	No	No	No	Abnormal	A1
0	B	Single	<3cm	No	No	No	Normal	A1
0	B	Single	3-5cm	No	No	No	Normal	A1
0	B	Single	<3cm	No	No	No	Abnormal	A1
0	B	Single	3-5cm	No	No	No	Abnormal	A1
0	B	2-3	>5cm	No	No	ANY	ANY	B
0	A	2-3	>5cm	No	No	ANY	ANY	B

Figure B.2: BCLC lookup table (Bilirubin values >1.3 mg/dL are abnormal)

CLIP SCORE				
Child Pugh	Tumor morphology	AFP	Macrovascular invasion	CLIP SCORE
A	Uninodular	<400	No	0
A	Multinodular	<400	No	1
A	Massive	<400	No	2
A	Uninodular	>400	No	1
A	Multinodular	>400	No	2
A	Massive	>400	No	3
A	Uninodular	>400	Yes	2
A	Multinodular	>400	Yes	3
A	Massive	>400	Yes	4
A	Uninodular	<400	Yes	1
A	Multinodular	<400	Yes	2
A	Massive	<400	Yes	3
B	Uninodular	<400	No	1
B	Multinodular	<400	No	2
B	Massive	<400	No	3
B	Uninodular	>400	No	2
B	Multinodular	>400	No	3
B	Massive	>400	No	4
B	Uninodular	>400	Yes	3
B	Multinodular	>400	Yes	4
B	Massive	>400	Yes	5
B	Uninodular	<400	Yes	2
B	Multinodular	<400	Yes	3
B	Massive	<400	Yes	4
C	Uninodular	<400	No	2
C	Multinodular	<400	No	3
C	Massive	<400	No	4
C	Uninodular	>400	No	3
C	Multinodular	>400	No	4
C	Massive	>400	No	5
C	Uninodular	>400	Yes	4
C	Multinodular	>400	Yes	5
C	Massive	>400	Yes	6
C	Uninodular	<400	Yes	3
C	Multinodular	<400	Yes	4
C	Massive	<400	Yes	5

Figure B.3: CLIP lookup table

Appendix C

PRIMARY LIVER CLINIC NOTE SECTION ONTOLOGY

A clinical note has the following general sections with the various subsections.

GENERAL INFORMATION

Clinic note type

Date of Service

ID/Chief Complaint

HISTORY

History of present illness

Past Medical history

Past Surgical history

Social history

Family history

Allergies

Medications

Review of systems

Performance status

OBJECTIVE DATA

Vital signs

Physical examination

Laboratory results

Imaging studies

ASSESSMENT & PLAN

Assessment

Plan

OTHER

Attending statement

Timed billing statement

Journal references

Appendix D

TUMOR TEMPLATE ANNOTATIONS GUIDELINES

1. Introduction

In this document we describe the annotation process and lay out the annotation guidelines for marking pieces of information that will be used for tumor information extraction. Strictly speaking the goal of our annotation is to locate all potential tumor mentions, which is different than identifying text that is judged to be a tumor by experts.

In this guide, we will start with a general overview about the problem space and introduce the primary annotation elements, before establishing some general annotation decisions. Afterwards, we will provide solutions for possible ambiguous cases.

2. Tumor references in medical diagnostic reports

When there is abnormal growth of tissue, a mass may form, which constitutes a tumor, also known as a neoplasm. Tumors may be malignant, when they are still growing and invading surrounding tissue, or benign, when growing has stopped. Malignant tumors are commonly called cancer.

Because tumors occur inside the body, to gather information about potential cancer, diagnostic radiology tools are used to identify tumors. However, the resolutions of these tools are limited and it may not always be clear whether structures in an image are truly tumors or may constitute something else.

For example, the following passages refer to artifacts that are indeed tumors

> *1.9 x 1.8 cm hyperenhancing **mass** on the arterial phase with enhancing pseudocapsule, corresponding washout on portal venous phase as well as T2 hyperintensity and restricted diffusion, **characteristic of HCC**.*

> *2. Segment 5 **lesion**, 1.1 cm , image 12: 50.*
> *The lesion is hypervascular on the arterial phase with possible delayed enhancement.*
> ***Diagnostic considerations include HCC versus cholangiocarcinoma**.*

whereas the next two are artifacts from other causes.

> *1.1 cm **focus** of nonenhancing liver parenchyma with signal dropout on opposed phase images most **suggestive of focal fatty infiltration**.*

> *There are multiple scattered hepatic **hypodensities** that exhibit no enhancement and **likely represent cysts**.*

In some cases, it is not clear what an artifact is.

> *In segment 4a, there is a stable **hypovascular lesion** which is **indeterminate** and could represent a regenerative nodule. Would recommend MRI with Eovist specifically to further evaluate this lesion.*

In addition to this technology challenge, medical reports are not immune to the challenges of general NLP obstacles. In particular, anaphora, when one sentence refers to context from another sentence, is a persistent problem. Below we initially receive information about two lesions, and only in the Impression section do we know that the first lesion may be an HCC. However, the sentence in the Impression section does not contain the measurements of the tumor, so it must be inferred by connecting it with sentence 16.

> *15: Focal Lesions:*
> *16: **Lesion 1**: Segment 3, 2.7 x 2.2 cm, (series 6, image 6 and 3/28).*
> *17: Lesion 2: Segment 8, greater than 1 cm, indeterminate.*

18: Impression:
*19: **One focal lesion** in segment 3, suspicious for HCC.*

References may also involve split antecedents, i.e. multiple "first mentions," later referred to collectively.[199]

27: Focal lesions:
28: Total number: 4:
29: Lesion 1: segment 8, 2.2 x 2.0cm , image 4/41, hypervascular with washout on veinous phase - HCC
*30: **Lesion 2**: segment 5, 0.7cm , image 4/49, hypervascular with no washout.*
*31: **Lesion 3**: segment 4A, 0.5cm , image 4/24,hypervascular with no washout.*
*32: **Lesion 4**: Segment 5, 0.6 cm, image 4/56, hypervascular with no washout.*
……
37: Impression:
38: 1 focal lesion in segment 8, highly suggestive of HCC.
39: Technically limited study for the characterization of HCC due to the absence of the delayed phase.
*40: **3 indeterminate focal lesions** in segment 4 a and 5.*

Moreover, this problem is exasperated by temporal references.

*The previously visualized **mass** involving segment 5 and segment 6 has increased in size (cranial caudal measuring 11 mm, **previously 8.5 mm**) and now extends to involve segment 4.*

And at times there are no local tumor references.

Wedge-shaped distribution pattern is suggestive of vascular invasion, however, direct visualization of vascular tumor thrombus is not present.
The other hypervascular foci in the liver as follows:
Segment 4A/2, 0.8 cm, no washout.
Segment 8, 0.6 cm, image 34/6, no washout.
Segment 6, medially 1.5 x 1.3 cm with washout, image 45/6, HCC.
Segment 6, adjacent to the gallbladder, 0.9 cm, image 62/6 with no clear washout.
The hypervascular lesions with no clear washout are indeterminate, but HCC is not excluded.
Gallbladder.

Typical tumor extraction problems involve gathering information such as the tumor size, number, and location. Recognizing the reality of medical technology and reporting language, we approach annotation by taking a greedy mindset. We annotate for all possible tumor mentions, e.g. "lesions," and possible tumor measurments, e.g. "0.6 cm." And additionally annotate for the tumor characteristics and the reason a particular mention should be considered a tumor.

3. Entity Annotations

Anat: Anatomy, values: (Liver,NonLiver)
Location that a potential tumor has been found in

- Test question: Is this an anatomical part?
  - If YES, then mark as anatomy
  - If NO, then do not mark
- Capture largest span, e.g. "left lobe" instead of just "lobe"
- Test question for including adjectives: Does this adjective describe an anatomic unit of the organ?
  - If YES, then include in anatomy
  - If NO, then only highlight the noun
- When possible mark separate entities, e.g. "segment 6 and segment 7"
  - This includes cases like "left lobe of the liver"
- If multiple mentions appear in a conjunction, highlight entire region

1. 10.2 cm hypervascular mass in hepatic segments 4b and 3 with imaging characteristics suggestive of cholangiocarcinoma.

200

- You must assign one of 2 possible values:
  - **Liver**: anatomy that refers to a liver
  - **NonLiver:** anatomy DOES NOT refer to the liver

Meas: measurement
Size or measurements indicated in the text
Mark sizes even if they are from a past diagnostic test

- Test question: Is this a measurement?
  - If YES, then mark as a measurement
  - If NO, do not mark
- Can often be found with measurement units, e.g. "cm" or "mm"
- Do NOT mark qualitative sizes, e.g. "large" or "small"

Lesion 1: segment 8, 2.2 x 2.0cm , image 4/41, hypervascular with washout on veinous phase - HCC

Negation
Whether or not information related to a tumor reference or tumor measurement is negated

No other suspicious liver lesions are identified.

TumCount: Tumor count
Number of potential tumors referenced

- This entity may not be used for potential tumor references alluding to a single tumor
- Include numeric or text numbers, e.g. "two" or "2"
- Mark qualitative sizes, e.g. "single," "few," "multiple"

1. Multiple right hepatic lobe arterial enhancing masses which are either unchanged or slightly smaller on the follow up study which could be secondary to treatment response.

TumRef: Tumor reference
Potential tumor mention

- Test question: Can this entity possibly refer to a tumor?
  - If YES, then mark as a tumor reference
  - If NO, do not mark
- Test question for adjective inclusion: Does the adjective describe radiology diagnostic information?
  - If YES, then mark as part of the tumor reference
    - Ex: "focal lesion" and "hypervascular mass" => these adjectives come from describing radiology artifacts
  - If NO, then only mark noun
    - Ex: "visualized mass" => "visualized" is not radiological adjective
    - Ex. "wedge-shape", "satellite", "subcapsular", "vascular" => describes findings
- Often, "mass", "lesion", "foci", "hyperintensity"
- Do NOT annotate demonstratives, e.g. "this" or "those"

5.7 x 7.4 cm hypervascular segment 6 lesion with washout, characteristics of HCC.

TumEvid: Tumorhood evidence, values: (isCancer, isBenign/notTumor, inDet)

- Test question: Does this information tell me whether or not something is a tumor/cancer?
  - If YES, then mark as a tumorhood evidence
  - If NO, do not mark
- Should get minimum text-span but still gives a complete idea
  - Usually starts with a "hedging" term such as "suspicious," "likely," "possibly"
- You must also assign one of 3 possible values:
  - **isCancer:** is a malignant tumor, e.g. "characteristic of HCC"

5.7 x 7.4 cm hypervascular segment 6 lesion with washout, characteristics of HCC.

- Imaging indicators of **isCancer**:
  - enhancement AND washout
  - LIRADS 4-5

  - **isBenign/notTumor:** when the tumor reference or measurement refers to something that is not a benign tumor and will not be in the future

1.1 cm focus of nonenhancing liver parenchyma with signal dropout on opposed phase images most suggestive of focal fatty infiltration.

- Imaging indicators of **isBenign/notTumor**:
  - LIRADS 1-2

  - **inDet:** when imagining cannot determine the whether cancer or not or if tumor reference or measurement refers to something that is neither cancer or benign, e.g. "indeterminate" or "dysplastic nodule"

There is a hypovascular lesion seen in segment 4a which measures approximately 2.2 x 3.0 cm and is indeterminate.

- Imaging indicators of **inDet**:
  - LIRADS 3

In text examples of positive tumorhood evidence:

Radiologist makes diagnosis
is definitively a hepatocellular carcinoma
highly suspicious for hepatocellular carcinoma
typical for HCC
classic hypervascular hepatocellular carcinoma
consistent with HCC
consistent with a fat containing HCC
represents biopsy-proven HCC
typical for largely necrotic HCC
characteristic for HCC
LIRADS/LI-RADS/Li-rads/LR 5 (5+any letter would count)
LIRADS/LI-RADS/Li-rads/LR 4 (4+any letter would count)

Imaging descriptors (all essentially examples of the same concept of arterial enhancement and washout)
arterially enhancing hepatic lesions with delayed phase washout

heterogeneous arterial enhancement on the arterial phase and washout on the portal venous and delayed phases
lesions exhibiting arterial hyperenhancement and delayed washout
hypervascular subcapsular lesion demonstrates washout on the more delayed images
minimally enhancing but washing out in the venous and delayed
heterogeneously hypervascular lesion demonstrating washout
Numerous hypovascular lesions scattered throughout the liver with delayed hypodensity
hypervascular lesion that demonstrates hypointensity on hepato specific sequences
enhancement and hypointense on 20min delayed hepatospecific phase
Infiltrative, hyper enhancing mass on arterial phase which washes out on delayed phase
arterially enhancing mass demonstrating delayed hypointensity

Mixed cases
hypervascular mass with no washout suspicious for HCC (In this case, "suspicious for HCC" trumps "no washout")
interval increase in size and central necrosis is suspicious for HCC (again, "suspicious for HCC" trumps an otherwise less specific imaging descriptor)
The first part "interval increase in size and central necrosis" is less specific. Other types of cancers (colon metastasis, lymphoma, etc) could also grow and develop necrosis. The "is suspicious for HCC" says that the radiologist that in this context it is consistent with HCC.

4. Relation Annotations

hasCount: Number of potential tumors being referenced
Relates the tumor reference to the tumor count associated with it.

3 indeterminate focal lesions in segment 4 a and 5 ***hasCount(focal lesions, 3)***

isNegated: Indicates a negated concept
Relates the negation to the concept

No other suspicious liver lesions identified. ***isNegated(lesions, No)***

locatedIn: Location Indicator
Relates the tumor reference to the liver anatomy where it is found.

1. 10.2 cm hypervascular mass in hepatic segments 4b and 3 with imaging characteristics suggestive of cholangiocarcinoma. ***locatedIn(hypervascular mass, segment 4b and 3)***

hasMeasurement: Present tense indicator of size of potential tumor
Relates the tumor reference to a measurement.

1. 10.2 cm hypervascular mass in hepatic segments 4b and 3 with imaging characteristics suggestive of cholangiocarcinoma ***hasMeasurement(hypervascular mass, 10.2 cm)***

hadMeasurement: Past tense indicator of size of potential tumor
Relates the tumor reference to a tumor measurement that was true in the past but is not currently.

The previously visualized mass involving segment 5 and segment 6 has increased in size (cranial caudal measuring 11 mm, previously 8.5 mm) and now extends to involve segment 4.
hadMeasurement(mass, 8.5 mm)

hasTumEvid: Evidence of tumorhood
Relates the tumore reference to a positive, negative, or uncertain declaration of tumorhood evidence.

1.1 cm focus of nonenhancing liver parenchyma with signal dropout on opposed phase images most suggestive of focal fatty infiltration. ***hasTumEvid (focus, suggestive of focal fatty infiltration)***

referTo: Indicator of a anatomy entity measurement
Relates a measurement to an anatomy entity

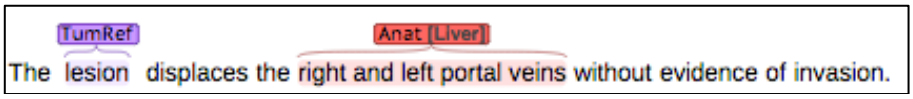
Spleen: Splenomegaly , measuring 17 cm in long axis. ***referTo(17 cm, Spleen)***

5. Annotation Decisions

Major

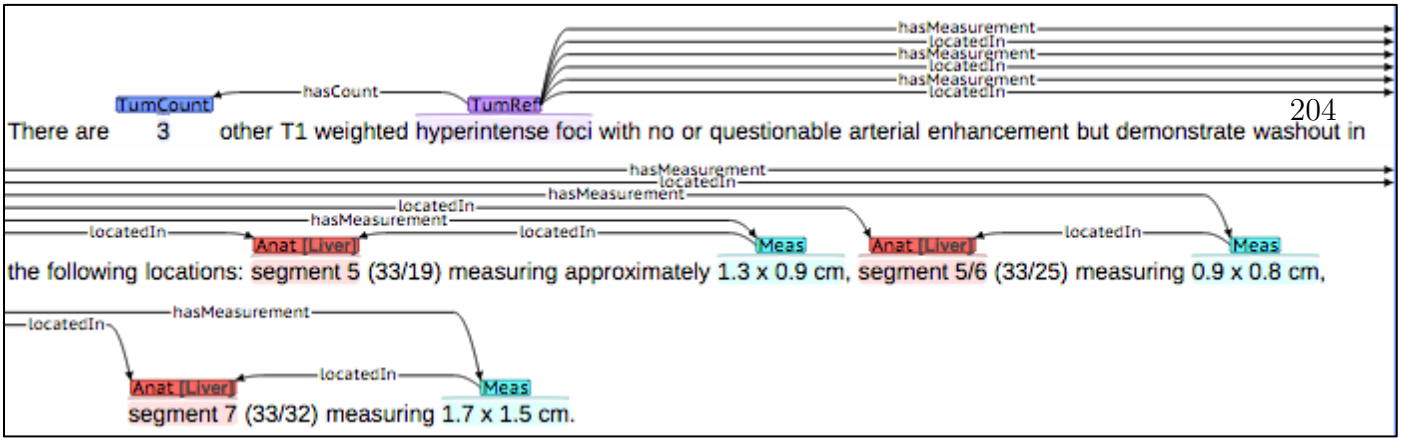
- Annotation will only occur for **Findings** and **Impression** sections of **radiology reports**.
- The reports will be pre-annotated for report sections.
- Annotate all possible tumor reference and measurements in all sentences. Other entities should only be marked if in relation to the tumor reference or measurement.
    - Reasoning:
        - A statement characterizing a tumor may not always have a tumor reference, (may be referred to as "the largest" "this" or just a measurement).
        - Marking every entity consistently in other context is too exhaustive and does not contribute significantly better information.
            - For example, "cysts" may be described by themselves without a mention of a lesion and without any measurement information. So annotating those instances would contribute little value.
- Additionally annotate anatomy entities if they appear in the same line as a tumor reference or measurement, even if they are not to be connected to the tumor reference or measurement in any way. Also annotate if they appear in the same line as a anatomy that is attached to a tumor reference or measurement.

The lesion displaces the right and left portal veins without evidence of invasion **(lesion is not locatedIn portal veins, but "right and left portal veins should still be annotated")**



- Always attach entities to either a tumor reference (preferred) or to a tumor measurement.
- Try to treat each tumor reference or measurement as a separate event, i.e. don't attach coreferential tumor references / measurement description items together.
- Additionally attach anatomy entities to measurements even if a tumor reference appears, if tumor count is greater than 1.

There are 3 other T1 weighted hyperintense foci ... in the following locations: segment 5 measuring 1.3 x 0.9 cm, segment 5/6 measuring 0.9 x 0.8 cm, segment 7 measuring 1.7 x 1.5 cm



Minor

- Do not annotate ending punctuations
- When possible annotate separate entities, e.g. "segment 5 and segment 6" but highlight together if not possible to separate.

A 3.6 cm mass in segment 6/7 with imaging characteristics favoring focal nodular hyperplasia

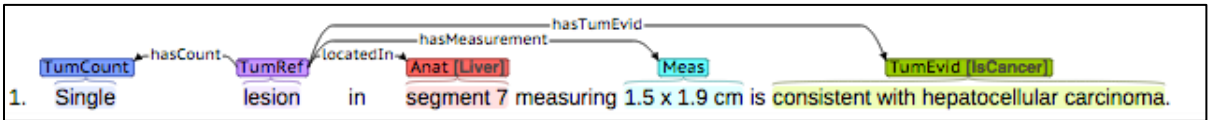
- In general annotate largest noun-phrase considered a single medical entity, e.g. "hypervascular mass" OK, but "visualized mass" only annotate the "mass"
- Include limits indicators for measurements, e.g. "greater than 1 cm" (but "approximately" not considered a limit indicator)
- Do not annotate section header body parts as location, except when they are on the same line as one line section.
- Noun phrase tumor reference: Do not highlight if over 2 words are not considered "radiological"

  ex. "heterogenously enhancing partially circumscribed mass", just mark "mass"

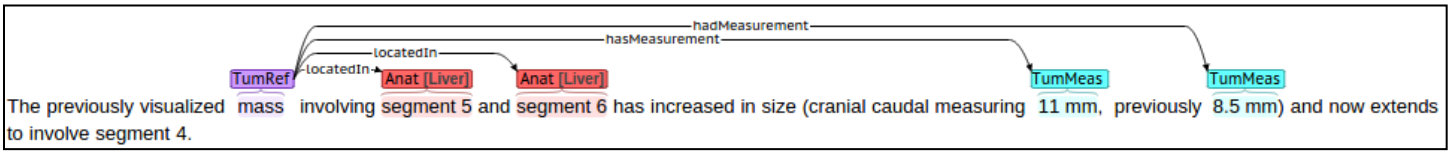
- For tumorhood evidence isCancer, for "enhancement and washout" get the minimum span containing this information. Only mark this if it is on the same line as the tumor reference or measurement.
- For tumorhood evidence, do not annotate ancillary features.
- For anatomy, for structures that are on the border of the liver and the outside liver, include as a liver anatomy.
  - porta hepatis, fissure, periportal – include as liver
  - left and right portal vein – include as liver
  - main portal vein – include as NonLiver
  - left, right, middle hepatic vein

6. Full mark-up examples

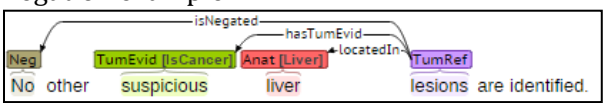
Simple example with a canonical tumor reference attached with a anatomy, count, measurement, and tumorhood evidence.



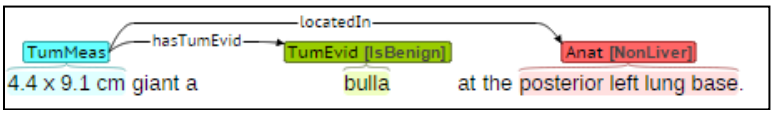
Mass with a current and previous measurement. Tumor measurements may be related to the reference as "hasMeasurement" for present tense and "hadMeasurement" for past tense.



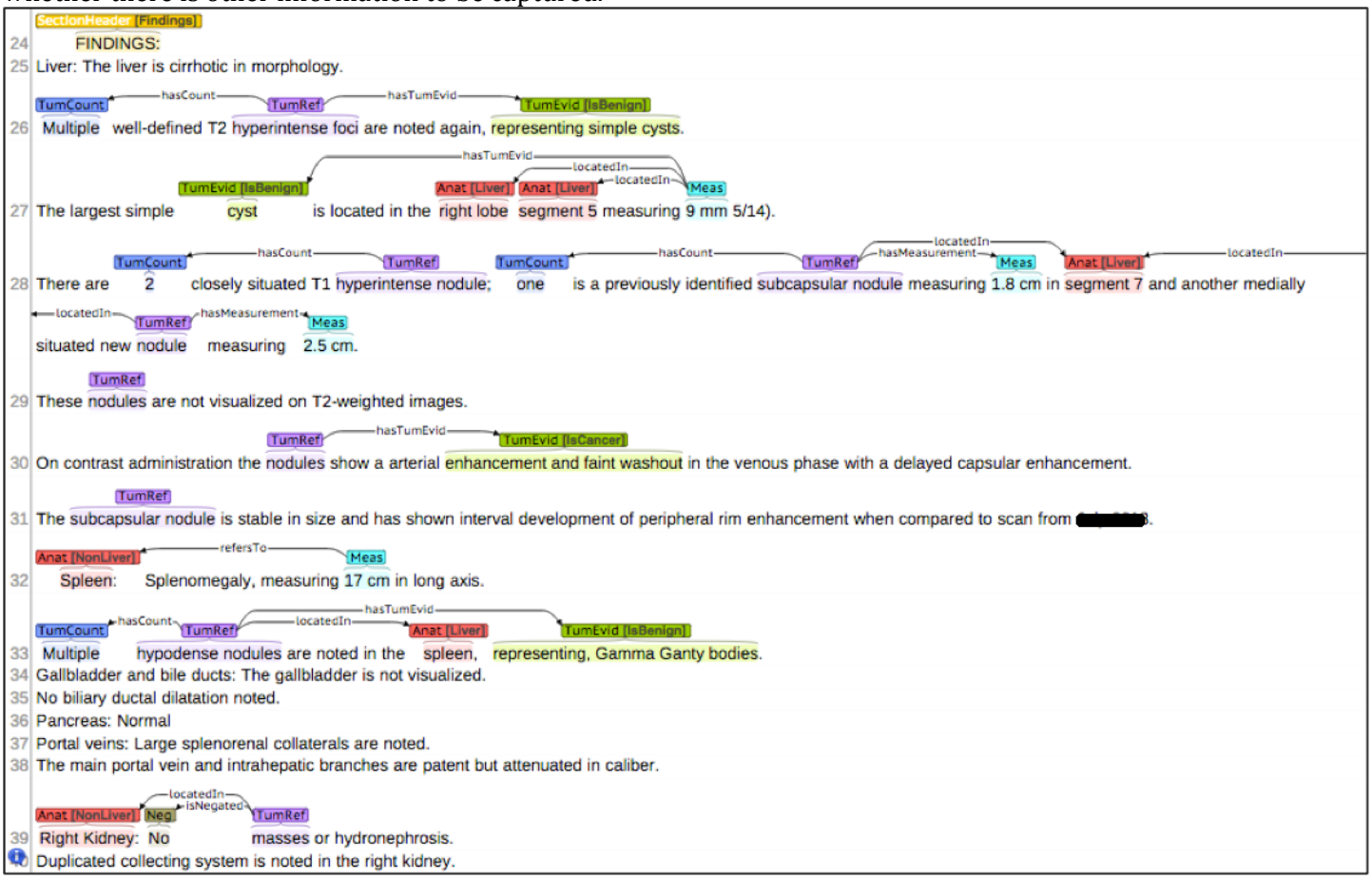
Negation example



Annotated measurement that is not a tumor



Annotations in brat. The "SectionHeader [Findings]" is part of the pre-annotated section headers. All possible tumor references and tumor measurements are annotated regardless of whether they are actual tumors or whether there is other information to be captured.



7. Annotation Questions

What should be connected in case of no tumor reference?

Segment 6, adjacent to the gallbladder, 0.9 cm, image 62/6 with no clear washout.

=> Attach everything to the tumor measurement, "0.9cm"

How do I annotate a lesion that is given a "Proper name"?

Lesion 1: segment 8, 2.2 x 2.0cm , image 4/41, hypervascular with washout on veinous phase - HCC

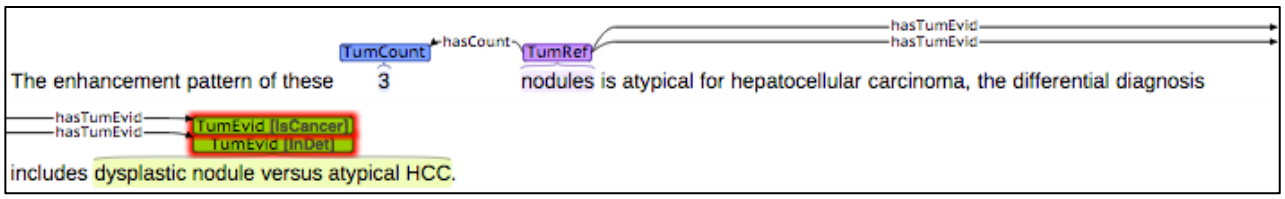
=> Include the entire "name", "Lesion 1"

Should I annotate all anatomy mentions?

Gallbladder: Normal

=> NO, only annotate anatomy mentions if they appear in the same line as a tumor reference or measurement, or if it is on the same line as a anatomy mention that is attached to a tumor reference or measurement.

What happens when there are multiple tumorhood evidence suggestions?



=> Annotate multiple and attach to the appropriate reference

Should I annotate a sentence even if it gives no other information?

No other hepatic lesions.

Mild interval increase in size of this lesion compared to the ultrasound from dd/dd/dddd is likely at least partially secondary to differences in technique

=> YES, only for tumor reference and measurements.

Should I annotate a measurement even though it is a measurement for an organ (ex. spleen) or something else (ex. Lymph node)?

Spleen: Mild splenomegaly, AP diameter 14.8 cm

17 x 10-mm right paratracheal lymph node (d/dd), previously measured 13 x 9 mm

Right kidney: 1.4-cm cortical hypodense cyst in the upper zone of the right kidney [...]

=> YES, mark the measurement. Also mark the anatomy entity attach it as to the measurement with a "referTo" relation.

Should I annotate tumor-like references?

Multiple small cysts

=> NO, a "cyst" already determines that it is not a tumor, therefore it cannot be a possible tumor reference.

Should anatomic adjectives be marked as a anatomy location?

Hepatic segment 5 lesion (1401/111) measures 2.1 x 2.4 x 1.4 cm, previously 1.8 x 1.9 x 1.4 cm

6. right upper pole solid renal mass concerning for renal cell carcinoma

=> YES, mark anatomic adjectives as location.

What if the note gives a tumorhood evidence, but also seems to be unsure?

8.7 cm focal lesion in hepatic segment VII suspicious for HCC, Multiphase CT is recommended for better characterization.

=> Continue with marking the "suspicious for HCC".

Do I annotate "Findings" as possible tumor reference?

Findings are typical of HCC.

=> No, but HCC may be annotated as a tumorhood evidence

What if the tumor only takes up part of a segment?

Lesion on the border between segment 4A and 8 measuring 2.9 x 3.5 cm

=> Don't annotate qualifiers for segments

Appendix E

TUMOR REFERENCE RESOLUTION ANNOTATIONS GUIDELINES

Annotation Guidelines – Tumor Reference Resolution

1. Introduction

In this document we describe the annotation process and lay out the annotation guidelines for marking reference mentions of tumor artifacts in radiology report. Here, we assume that the corpus has already been marked with tumor events.

2. Reference resolution

Reference resolution involves identifying information in text that refer to the same entity. For example,

"John has a three ducks. They were just born yesterday. He is very proud of them."

John is referred to in the last sentence as *He*, while the *ducks* are alternatively addressed with *they* and *them*.

The general English domain focuses on named entities, where constraining features are based on subject-verb agreement, pronoun (he vs. she vs. it) usage, etc. For example, in the following example, *Victoria Chen* is referred to with *Chief Financial Officer*, *her*, *the 37-year-old*, and *company's president*. And *Megabucks Banking Corp* is referred to later as *the Denver-based financial-services company*. Target named entity types are **person**, **location**, and **organization**.

Victoria Chen, Chief Financial Officer of Megabucks Banking Corp since 2004, saw her pay jump 20%, to $1.3 million as the 37-year-old also because the Denver-based financial-services company's president.

Jurasky and Martin, Speech and Language 2nd edition

In the biomedical literature domain, target named entity types have been on **genes** and **proteins**. Below "p65" and "this transcription factor."

To investigate the molecular basis for the critical regulatory interaction between NF-kappa B and I Kappa B/MAD-3 a series of human NF-kappa B p65 mutants was identified that functionally segregated DNA binding, I kappa B-mediated inhibition, and I kappa B-induced nuclear exclusion of this transcription factor.

BioNLP 2011 Shared Task Protein/Gene Coreference Task

In the medical reference resolution domain, there has been much emphasis on identifying coreferences for **persons**, **symptoms**, and **tests**. This was the subject of the 2011 i2b2/VA cincinatti challenge. While overall scores have been high, these depend on the evaluation metric and the particular categories of the challenge. Below "Patient" and "She" refer to the same entity, but the two "Pathology" are not the same.

Patient underwent a total abdominal hysterectomy in 02/90 for a 4x3.6x2 cm cervical mass felt to be a fibroid at Vanor. Pathology revealed poorly differentiated squamous cell carcinoma of the cervix […] with extensive lymphatic invasion. She underwent exploratory laparotomy […]. Pathology was negative for tumor and showed peritubal and periovarian adhesions.

Jonnalagadda, et al. Coreference analysis in clinical notes: a multi-pass sieve with alternate anaphora resolution modules.

3. Event reference resolution

Event reference resolution is related entity reference resolution, which identifies events that are the same. An event is a predefined template with certain attributes. For example, in the general English domain, an event has "agent" "patient" and "location."

Example from (Liu, Supervised Within-Document Event Coreference using Information Propagation, LREC 2014)

Indian naval forces **came to the rescue (E1)** of a merchant vessel under **attack (E2)** by pirates in the Gulf of Ade on Saturday, **capturing (E3)** 23 of the raiders, India **said (E4)**.

Event 1: ***came to the rescue***
Agent: Indian naval forces
Patient: merchant vessel
Location: Gulf of Ade
Time: Saturday

Event 2: ***attack***
Agent: pirates
Patient: merchant vessel
Time: Saturday

Event 3: ***capturing***
Agent: pirates
Patient: 23 of the raiders

Event 4: ***said***
Agent: India

The Indian navy **captured (E5)** 23 piracy suspects who **tried (E6)** to **take over (E7)** a merchant vessel in the Gulf of Aden, between the Horn of Africa and the Arabian Peninsula, Indian officials **said (E8)**.

Event 5: ***captured***
Agent: Indian navy
Patient: 23 piracy suspects
Location: Gulf of Aden

Event 6: ***tried***
Agent: 23 piracy suspects
Location: Gulf of Aden

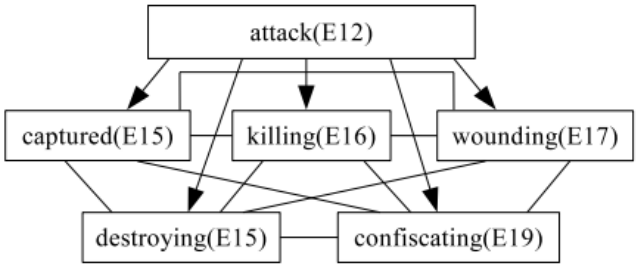
Event 7: ***take over***
Agent: 23 piracy suspects
Patient: merchant vessel
Location: Gulf of Aden

Event 8: ***said*** 211
Agent: Indian officials

In this case, events 3 and 5, events 2 and 7, and events 4 and 8 co-refer.

In the following example, we note **subevent** cases in which there are parent-child and sister relations in addition to a straightforward **coreference**.

Ismail said the fighting, which lasted several days, intensified when forces loyal to Egal's Ha-bar Awal sub-clan of the Issak **attacked**(E12) a militia stronghold of his main opposition rival, . . . Egal militia, claiming to be the national defence force, said they had **captured**(E15) two opposition posts, **killing**(E16) and **wounding**(E17) many of the fighters, **destroying**(E18) three technicals (armed pick-up trucks) and **confiscating**(E19) artillery guns and assorted ammunition



Examples from (Araki, Detecting Subevent Structure for Event Coreference Resolution, LREC 2014)

4. Tumor reference resolution in medical diagnostic reports

Reference resolution of tumors in radiology reports are in many ways similar to the traditional entity/event reference resolution task. For example emphasis on the WH- preceding determiner, e.g. "the" and "these", within-sentence repeated mentions, and intra-document mentions. Some event cues are based on attributes: anatomic location, measurements, and cancer description.

There are numerous hypervascular liver **lesions (E1)** involving all segments of the liver several of which demonstrate washout on delayed scans in keeping with multifocal hepatocellular carcinoma. The largest in the *left lobe* and is in inferior lateral segment 3 and measures **6.7 x 6.8 cm (E2)** in transverse diameter. … There are numerous additional confluent **lesions (E3)** in the *left lobe of the liver*. The largest lesion in the right lobe of the liver is in posterior superior segment 7 and measures **2.7 and 3.8 cm (E4)** in transverse diameter

Here, events 2 and 3 refer to the same lesions.

4. Our annotation

212

Our annotation is a direct extension of our previous sparse tumor annotation, which is described in our guidelines and in this paper [1]. Our goal is to annotate not only for equivalence event references but also for parent-child non-transitive, non-symmetric referring expressions, between tumor events. Because our annotation extends our previous scheme, we highlight some key differences with prior work.

Key Differences:

1. **Sparse annotation:** Not all lines are marked, and not all elements in the sentences are marked.
    a. "E.g. "The largest in the left lobe and is in inferior lateral segment 3 and measures 6.7 x 6.8 cm in transverse diameter." – "the largest" is not annotated
2. **Use of measurements as referring expressions of a tumor entity:** This was a consequence of the telegraphic nature of the medical text and as well as our sparse annotation (the goal to avoid excess annotation for all sentence elements, e.g. determiners).
    a. "Segment VII: 2.6 x 2.4 cm (37/4)." – "2.6 x 2.4 cm" is marked as the *referring expression* to a tumor

Our annotation is performed over a clinical radiology text corpus, primarily focusing on hepatocellular carcinoma patients. Thus the language and style of the reports yields important observations.

Domain particularities:

1. **Approximate measurement equivalences** – measurement sizes may not be exact, sometimes being referred to as "approximately X" where X may be the same number without a significant figure or rounded up or down.
2. **Largest measurement as referring expressions** – tumor measurements differ depending on the scan angle and axis. Typically, after mentioning all the dimensions, the maximum dimension may be used as a representative of the full measurement.

2.1 x 2.1 hyper enhancing mass in the arterial phase in segment 4a (5/16, 8/12) with washout and single pseudocapsule consistent with HCC.
Hyperenhancing lesion in segment 2 (5/12) measures 0.9 x 0.7 cm, without any suspicious washout on the portal venous and delayed images.
…
Impression:
1. 2.1 cm segment 4a mass consistent with HCC.
2. 0.9 cm hyper enhancement in segment 2 without any suspicious washout, is indeterminate

3. **Summarization by measurements** – More than one tumor may be described at once summarized by a bounded number range, e.g. "3 lesions all under 2 cm."
4. **Summarization by anatomic region** –More than one tumor may be described at once, summarized by a region, e.g. "3 lesions in the left lateral section" or "numerous tumors all over the right lobe."

Coreference of 2 summary statements with non-exact number agreement

Focal lesions:
Total number Two large and multiple small in right lobe
Lesion 1: Large lesion involving segment 8,7,6,5, At least 13.5cm

Lesion 2: segment 5, 7cm
Lesion 3: Multiple satellite lesions for example, segment 4 measures 2 cm. 213
….
Impression:
Multiple focal lesions in right lobe the largest measuring 13.5 cm.
Findings are highly suspicious of fibrolameliar hepatocellar carcinoma.
Smaller lesions are suspicious of satellite lesions.

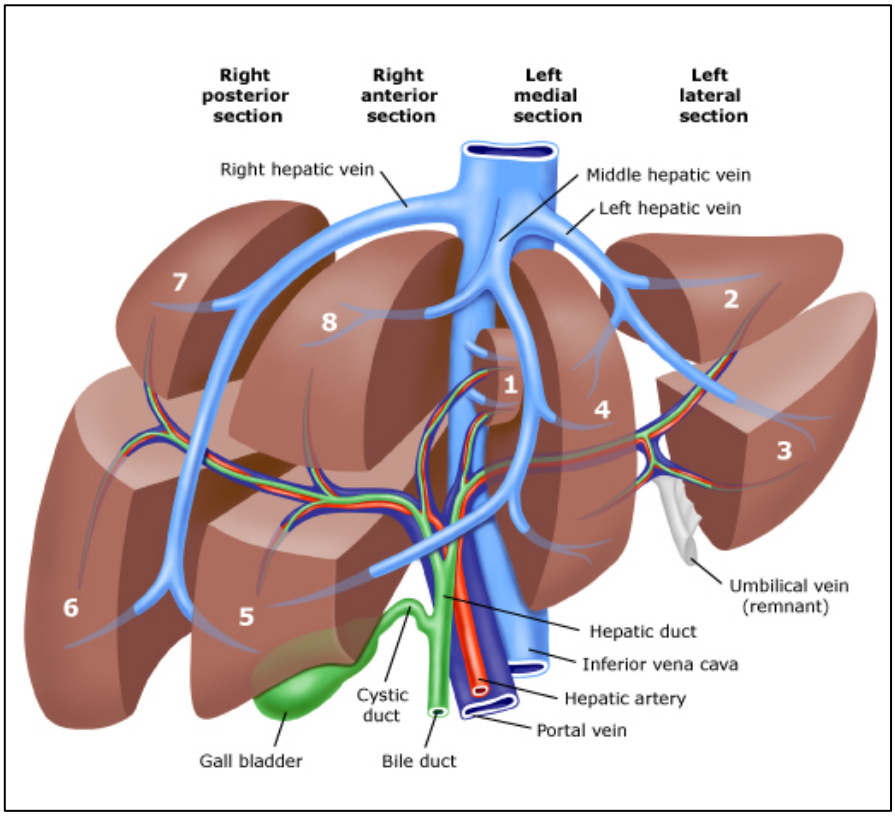


Image from www.aboutcancer.com

For more information about liver anatomy, there is an online version of Gray's anatomy (http://www.bartleby.com/107/250.html) and a site that describes how radiologists read images (http://www.radiologyassistant.nl/en/p4375bb8dc241d/anatomy-of-the-liver-segments.html).

5. Reference Annotations

This section describes the reference resolution annotations. Annotations should be between **event heads**. Tumor event heads can either be tumor references or measurements, typically the highest in the tumor template graph annotation. For tumor events with more than one measurement, those measurements are also considered tumor event heads.

Coref: Coreference
This type of relation annotate for equivalence references for two event heads. The relationship should be both transitive and symmetric.

Particularization: Reference of general referring expression to specific a specific one

This annotates for references from an event that is a superset to another event. This relation is non-symmetric, but it is transitive. 214

Multiple tumors in same place

In response to the question regarding the size of the masses seen in the liver: The mass in segment 6 measures 2.6 x 2.3 x 2.8 cm, measured in the hepatocyte phase. The mass in segment 8 measures 1.6 x 1.6 x 1.6 cm, measured in hepatocyte phase.

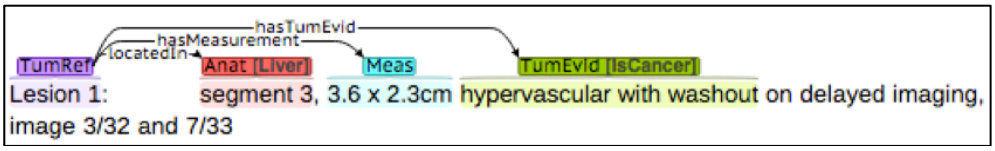
Particularization: {masses -> mass},{masses->mass}

Focal lesions: Total number: 9: Largest Lesion : segment 6, 5.5cm , image 601/61, at least 4 more lesions are probably HCC. Largest lesion has some fat attenuation in it. …. Impression: Agree Single phase scan limits the diagnostic specificity Nine hypervascular lesions in the liver. At least 5 lesions are probably HCC. The largest located in segment 6 measuring 5.5cm.

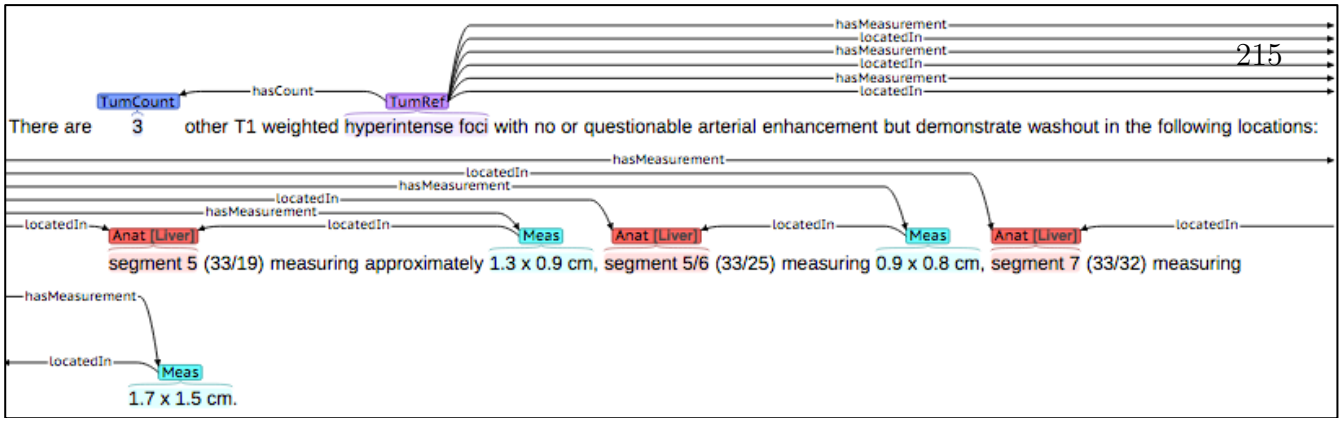
Coreference: {Lesion, lesion, 5.5cm}, {Focal lesions, lesions}
Particularization: {Focal lesions -> Lesion}{Focal lesions -> lesions} {lesions -> lesions} {lesions->5.5cm}

Connection with previous annotations

Previous annotation hasMeasurement, for a single measurement can be seen as a special case of COREF



Previous annotation hasMeasurement for more than one measurement can be seen as a special case of *particularization*



4. Annotation Decisions

1. Only mark references based on previous mention annotation. Do not correct past annotations errors.

2. Mark all head "tumor reference" and "measurement" with reference annotations.

3. **Mark non-head "measurements"** if there are more than one "hasMeasurement" for that tumor reference, and if the measurement is mentioned by itself at a later time.

4. Put the minimum amount of reference annotations to fully constrain the problem. (So if *A corefers with B* and *B corefers with C*, it is not necessary to also put *A corefers with C*. Similarly, if *A is a particularization to B*, and *B particularization to C*, there is no need to put *A particularization to C*).

5. Annotate all references, including the benign entities from other anatomic locations.

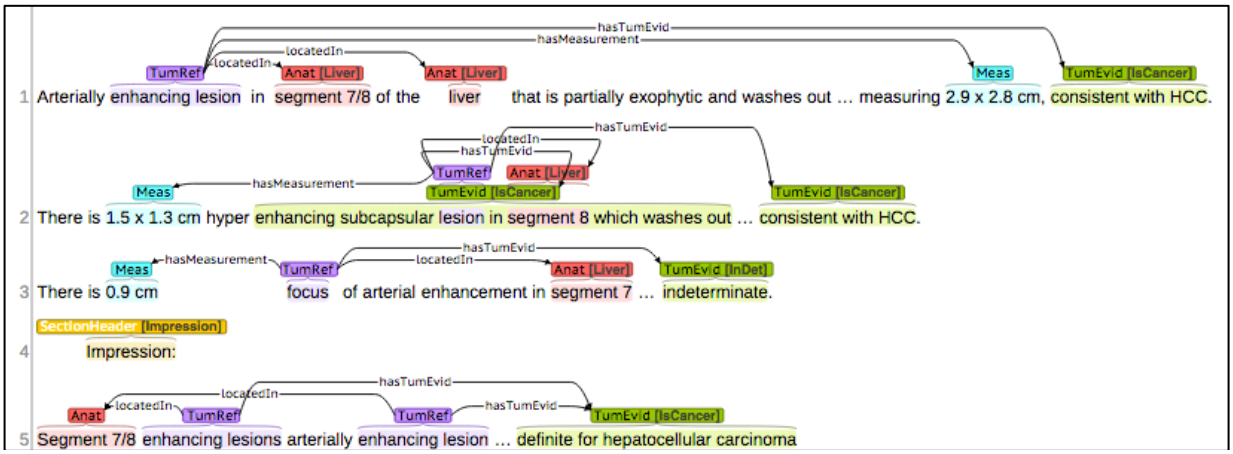
Findings:
Scans demonstrate at least 12 scattered pulmonary nodules in the 3-8 mm size range some of which are slightly larger than on the prior exam of 2 weeks ago
….
Assessment:
….
4. Multiple pulmonary nodules

Mark *nodules* from findings and assessment to COREF.

6. Mark-up examples

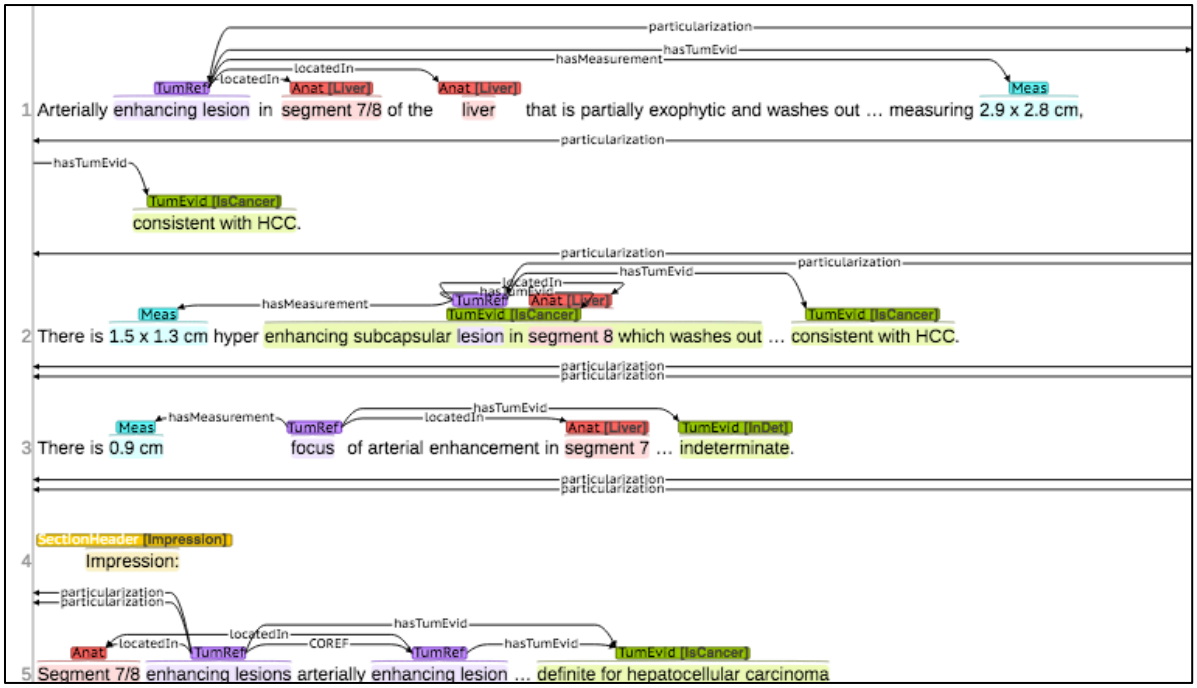
Example 1 Different tumor in same area, later referred to with a combined anatomy location

Arterially enhancing lesion in segment 7/8 of the liver that is partially exophytic and washes out … measuring 2.9 x 2.8 cm, consistent with HCC. There is 1.5 x 1.3 cm hyper enhancing subcapsular lesion in segment 8 which washes out … consistent with HCC. There is 0.9 cm focus of arterial enhancement in segment 7 … indeterminate. Impression: Segment 7/8 enhancing lesions arterially enhancing lesion … definite for hepatocellular carcinoma



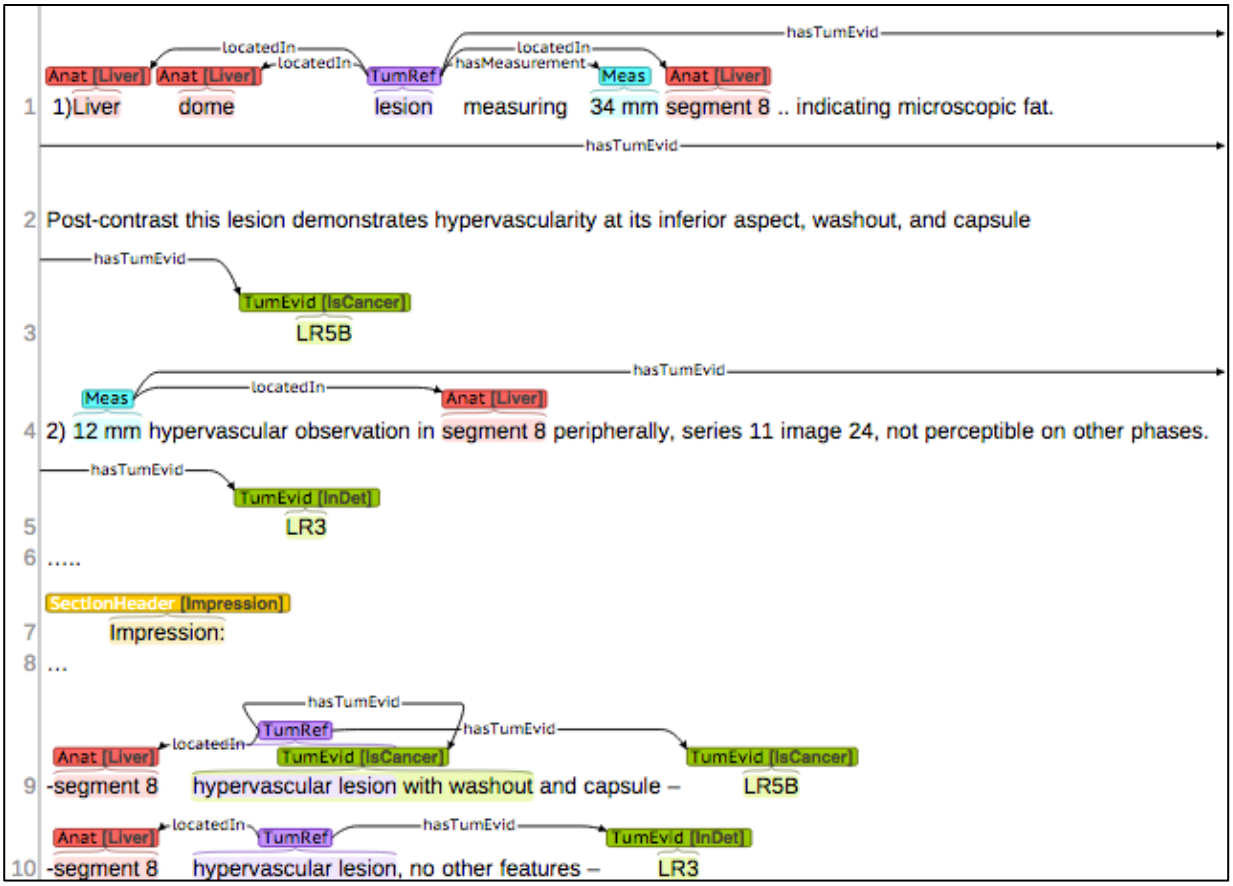
After adding reference resolution annotations:

Last (Line 5) *enhancing lesions* connected in a *particularization* relation to (Line 1) *enhancing lesion* and to (Line 2) *lesion*.



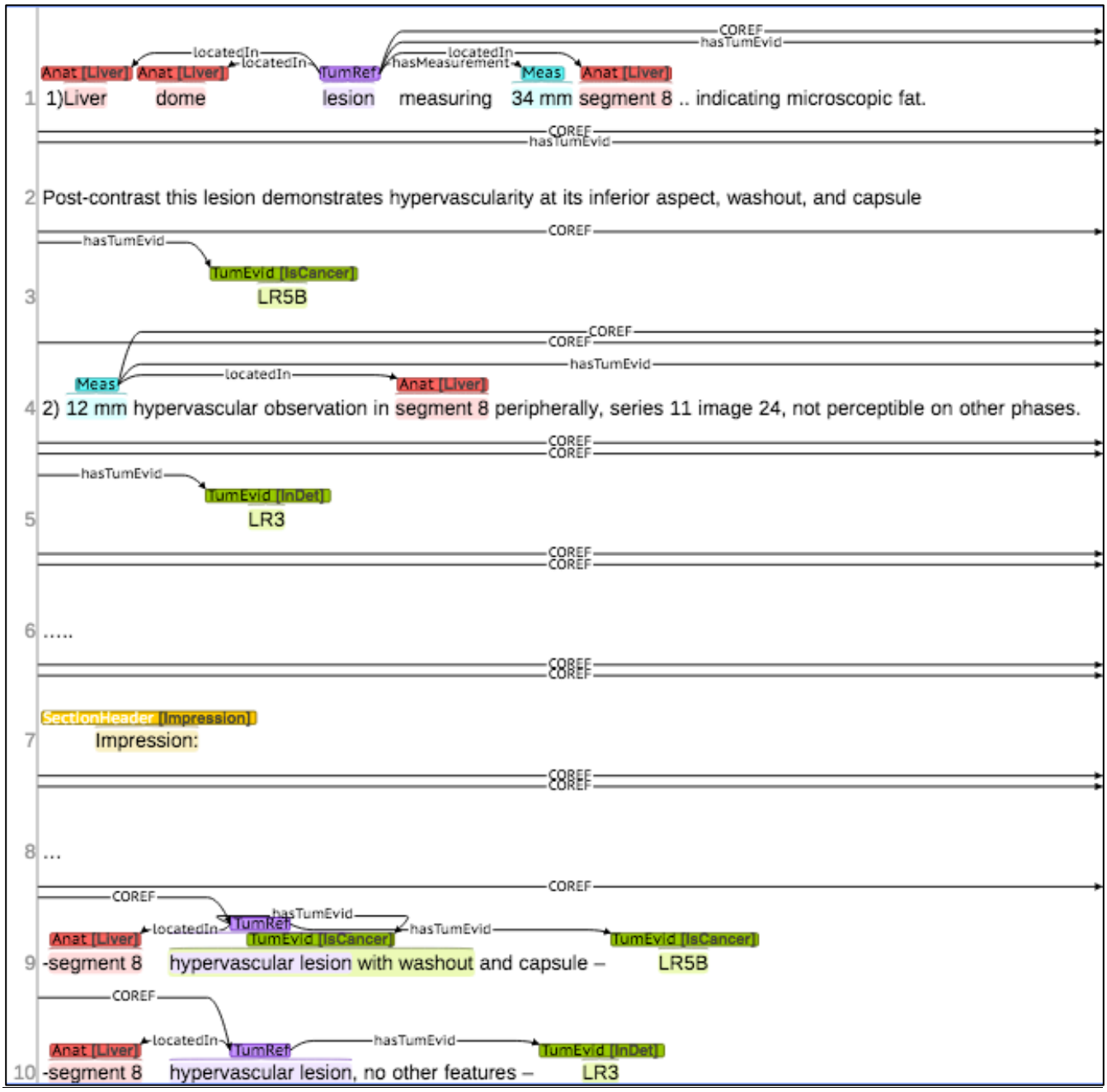
Example 2. **Needs cancer type to distinguish** 217

1)Liver dome lesion measuring 34 mm segment 8 .. indicating microscopic fat.
Post-contrast this lesion demonstrates hypervascularity at its inferior aspect, washout, and capsule
LR5B
2) 12 mm hypervascular observation in segment 8 peripherally, series 11 image 24, not perceptible on other phases.
LR3
…..
Impression:
…
-segment 8 hypervascular lesion with washout and capsule – LR5B
-segment 8 hypervascular lesion, no other features – LR3



After adding reference resolution annotations:

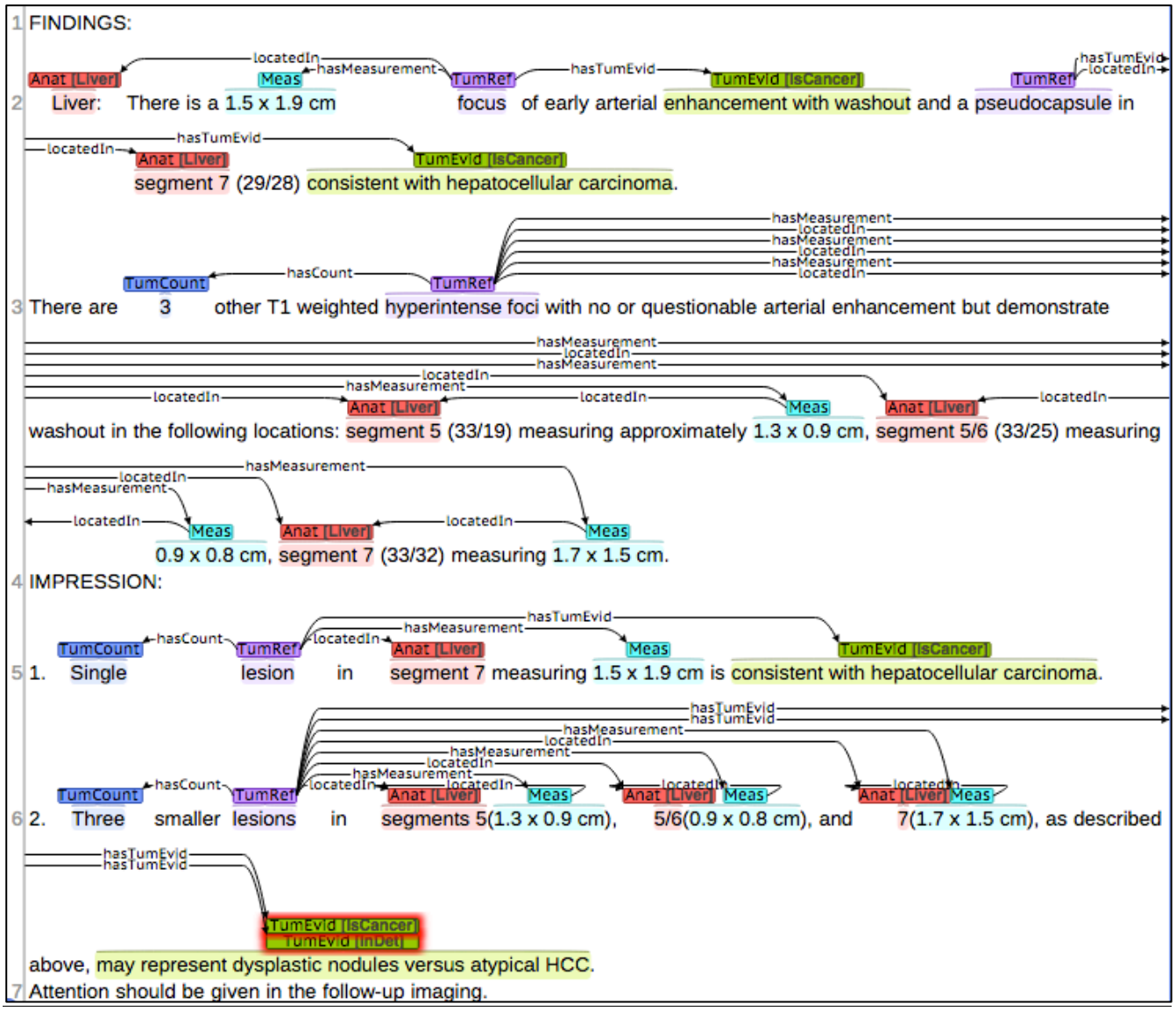
(Line 1) *lesion* connects with (Line 9) *hypervascular lesion* in a COREF relation
(Line 2) *12 mm* connects with (Line 10) *hypervascular lesion* in a COREF relation


1 1)Liver dome lesion measuring 34 mm segment 8 .. indicating microscopic fat.
Anat [Liver]
Anat [Liver]
locatedIn
locatedIn
TumRef
hasMeasurement
Meas
Anat [Liver]
COREF
hasTumEvid
2 Post-contrast this lesion demonstrates hypervascularity at its inferior aspect, washout, and capsule
hasTumEvid
TumEvid [IsCancer]
3 LR5B
Meas
locatedIn
Anat [Liver]
4 2) 12 mm hypervascular observation in segment 8 peripherally, series 11 image 24, not perceptible on other phases.
hasTumEvid
TumEvid [InDet]
5 LR3
6 .....
SectionHeader [Impression]
7 Impression:
8 ...
Anat [Liver]
locatedIn
TumRef
TumEvid [IsCancer]
hasTumEvid
TumEvid [IsCancer]
9 -segment 8 hypervascular lesion with washout and capsule – LR5B
Anat [Liver]
locatedIn
TumRef
hasTumEvid
TumEvid [InDet]
10 -segment 8 hypervascular lesion, no other features – LR3

Example 3. General statements with subevents

219

FINDINGS:
Liver: There is a 1.5 x 1.9 cm focus of early arterial enhancement with washout and a pseudocapsule in segment 7 (29/28) consistent with hepatocellular carcinoma.
There are 3 other T1 weighted hyperintense foci with no or questionable arterial enhancement but demonstrate washout in the following locations: segment 5 (33/19) measuring approximately 1.3 x 0.9 cm, segment 5/6 (33/25) measuring 0.9 x 0.8 cm, segment 7 (33/32) measuring 1.7 x 1.5 cm.
IMPRESSION:
1. Single lesion in segment 7 measuring 1.5 x 1.9 cm is consistent with hepatocellular carcinoma.
2. Three smaller lesions in segments 5(1.3 x 0.9 cm), 5/6(0.9 x 0.8 cm), and 7(1.7 x 1.5 cm), as described above, may represent dysplastic nodules versus atypical HCC.
Attention should be given in the follow-up imaging.



After adding reference resolution annotations:

220

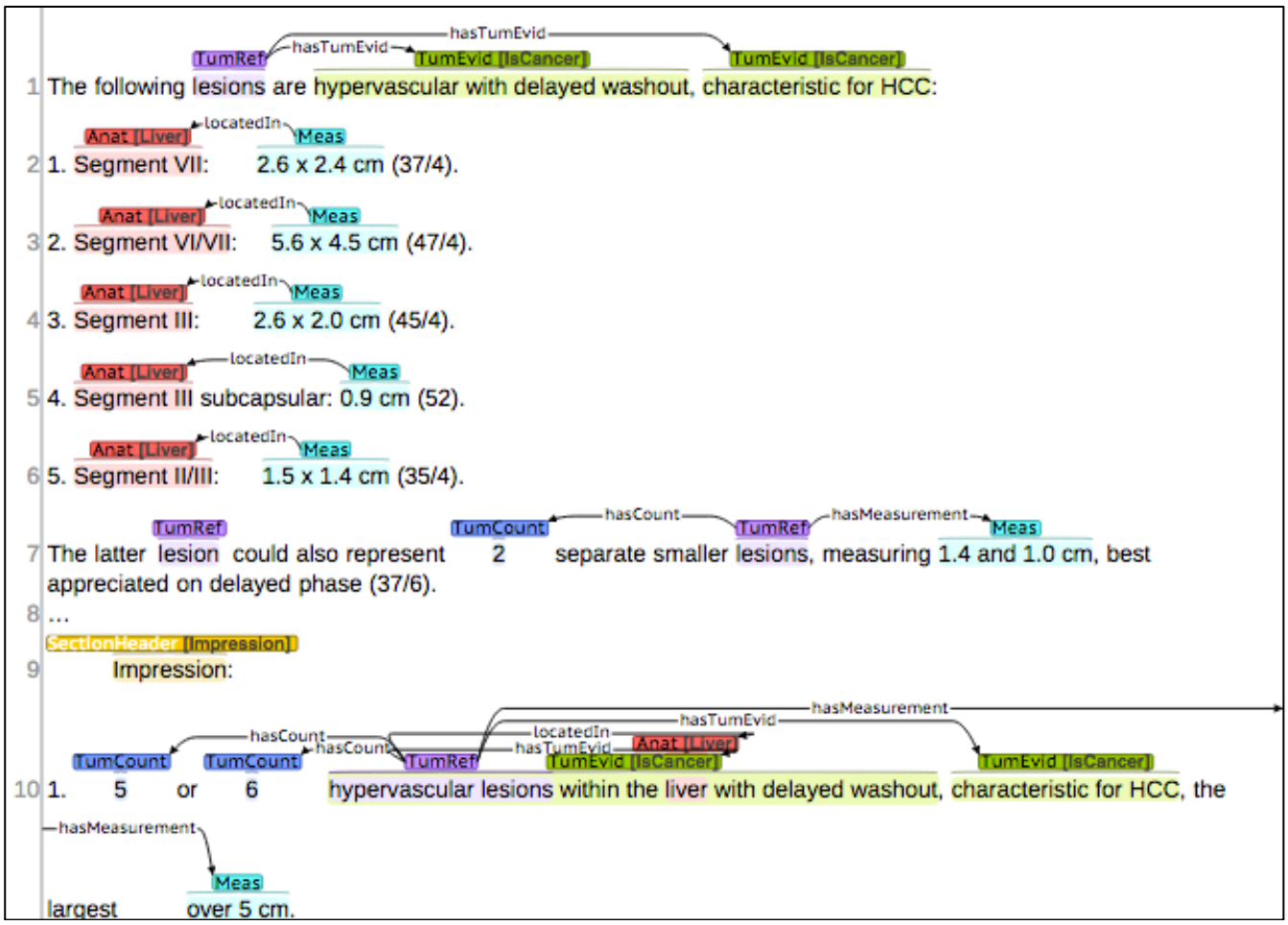
(Line 2) *focus* connects with (Line 5) *lesion* in a COREF relation
(Line 3) *hperintense foci* connects with (Line 6) *lesions* in a COREF relation
(Line 3) *1.4 x 0.9 cm* connects with (Line 6) *1.3 x 0.9 cm* in a COREF relation
(Line 3) *0.9 x 0.8 cm* connects with (Line 6) *0.9 x 0.8 cm* in a COREF relation
(Line 3) *1.7 x 1.5 cm* connects with (Line 6) *1.7 x 1.5 cm* in a COREF relation

Example 4. Ambiguity with number

221

The following lesions are hypervascular with delayed washout, characteristic for HCC:
1. Segment VII: 2.6 x 2.4 cm (37/4).
2. Segment VI/VII: 5.6 x 4.5 cm (47/4).
3. Segment III: 2.6 x 2.0 cm (45/4).
4. Segment III subcapsular: 0.9 cm (52).
5. Segment II/III: 1.5 x 1.4 cm (35/4).

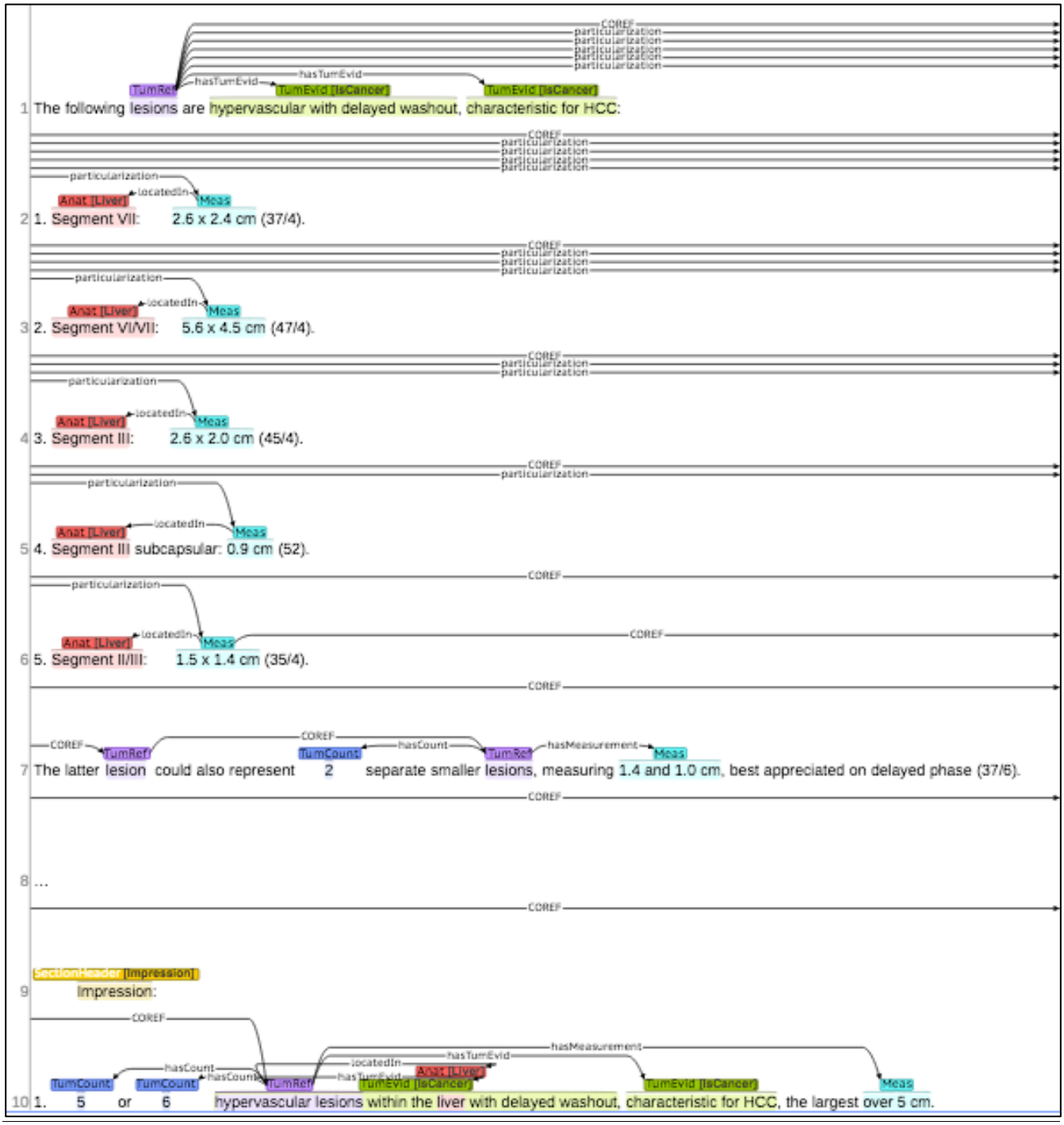
The latter lesion could also represent 2 separate smaller lesions, measuring 1.4 and 1.0 cm, best appreciated on delayed phase (37/6).
…
Impression:
1. 5 or 6 hypervascular lesions within the liver with delayed washout, characteristic for HCC, the largest over 5 cm.



After adding reference resolution annotations:

(Line 1) *lesions* connected to (Line 2-6) *2.6 x 2.4 cm*, *5.6 x 4.5 cm*, *2.6 x 2.0 cm, 0.9cm*, *1.5 x 1.4 cm* in a *particularization* relation.

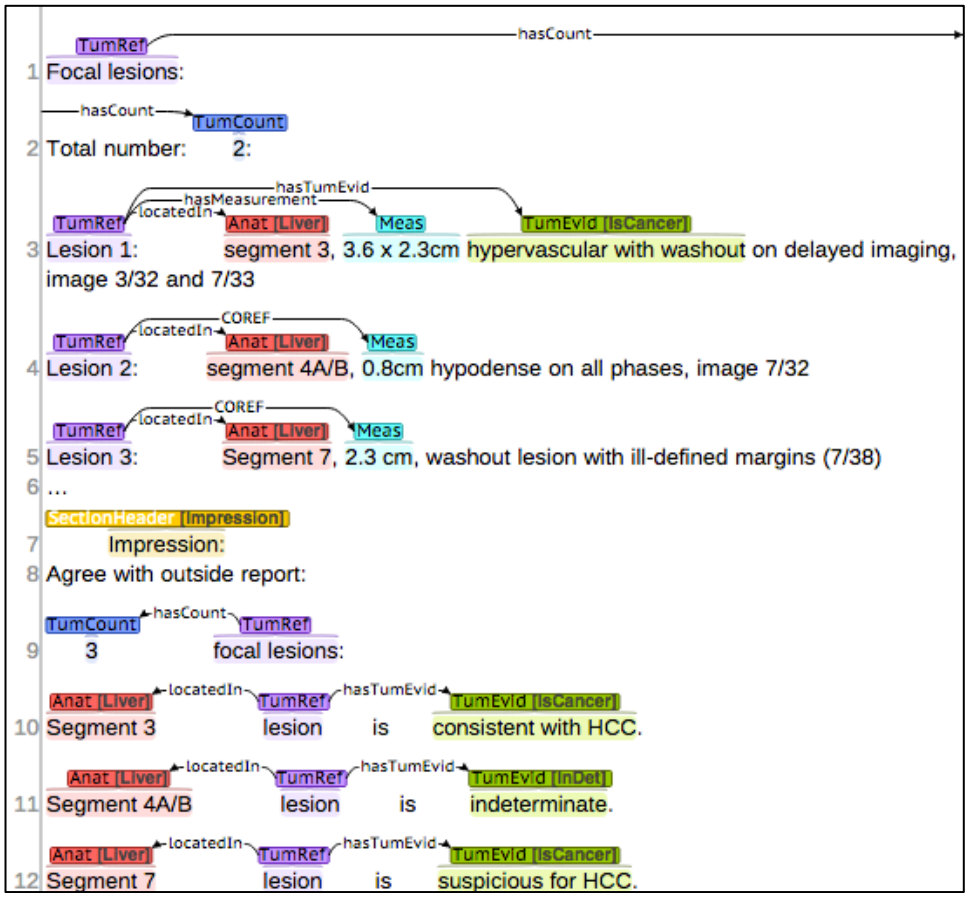
(Line 6) *1.5 x 1.4 cm* connects to (Line 7) *lesion* as a COREF relation
(Line 7) *lesion* connects with (Line 7) *lesions* as a COREF relation


1 The following lesions are hypervascular with delayed washout, characteristic for HCC:
2 1. Segment VII: 2.6 x 2.4 cm (37/4).
3 2. Segment VI/VII: 5.6 x 4.5 cm (47/4).
4 3. Segment III: 2.6 x 2.0 cm (45/4).
5 4. Segment III subcapsular: 0.9 cm (52).
6 5. Segment II/III: 1.5 x 1.4 cm (35/4).
7 The latter lesion could also represent 2 separate smaller lesions, measuring 1.4 and 1.0 cm, best appreciated on delayed phase (37/6).
8 ...
9 Impression:
10 1. 5 or 6 hypervascular lesions within the liver with delayed washout, characteristic for HCC, the largest over 5 cm.

Example 5. Incorrect numbers

223

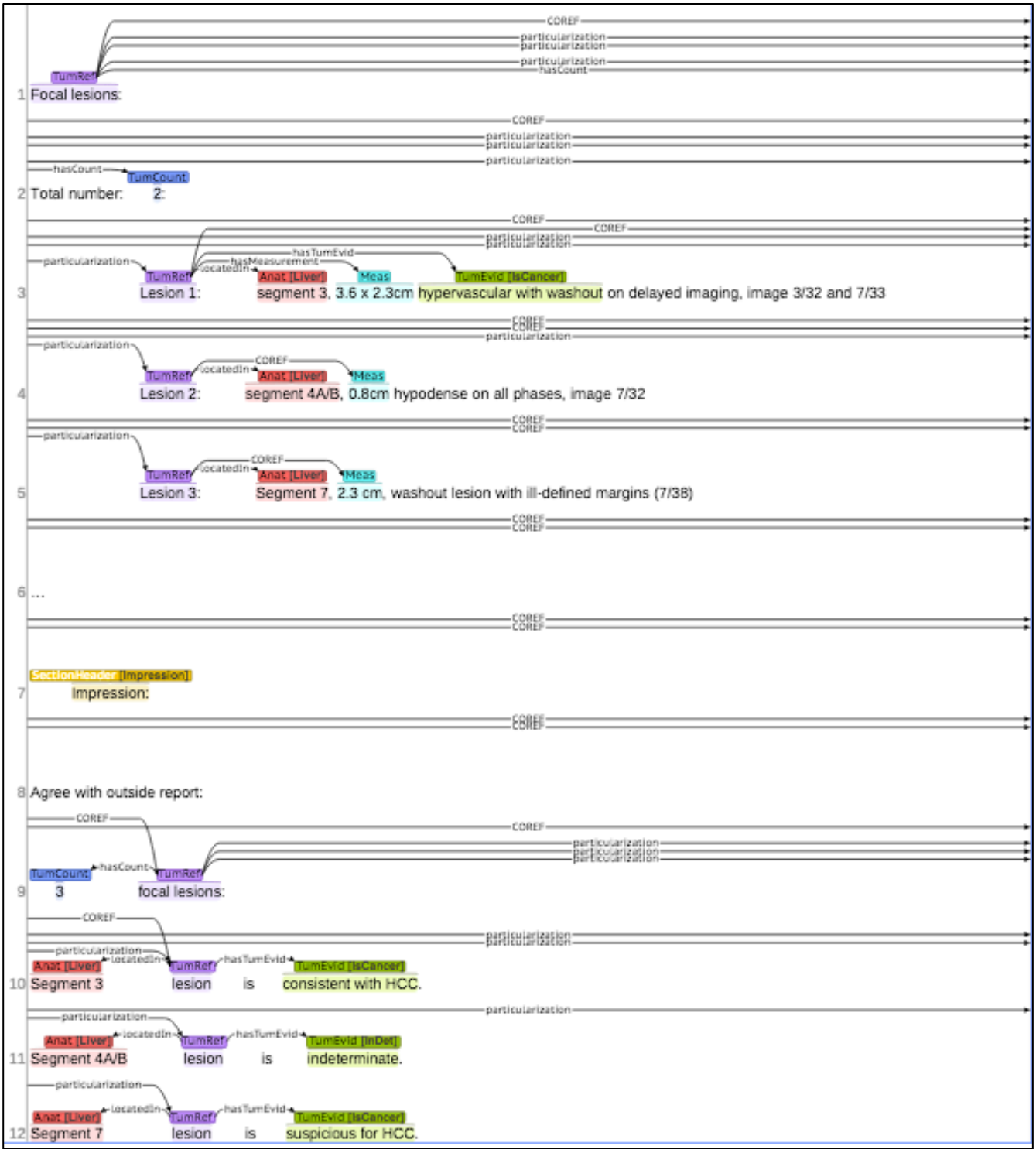
Focal lesions: Total number: 2: Lesion 1: segment 3, 3.6 x 2.3cm hypervascular with washout on delayed imaging, image 3/32 and 7/33 Lesion 2: segment 4A/B, 0.8cm hypodense on all phases, image 7/32 Lesion 3: Segment 7, 2.3 cm, washout lesion with ill-defined margins (7/38) … Impression: Agree with outside report: 3 focal lesions: Segment 3 lesion is consistent with HCC. Segment 4A/B lesion is indeterminate. Segment 7 lesion is suspicious for HCC.



After adding reference resolution annotations:

(Line 1) *Focal lesions* should connect with (Line 3) *Lesion 1*, *Lesion 2*, *Lesion 3* in a *particularization* relation.
(Line 9) *focal lesions* should connect with (Line 10) *lesion*, (Line 11) *lesion*, and (Line 12) *lesion* in a *particularization*.

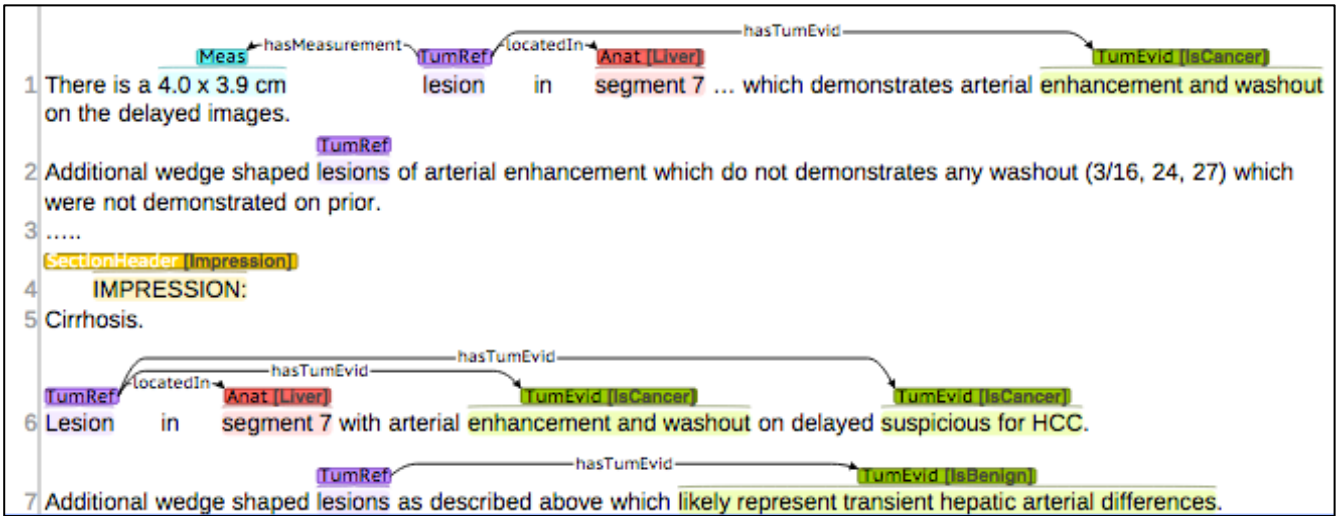
(Line 1) *Focal lesions* should connect with (Line 9) *focal lesions* in COREF relation 224
(Line 3) *Lesion 1* should connect with (Line 10) *lesion* in COREF relation
(Line 4) *Lesion 2* should connect with (Line 11) *lesion* in COREF relation
(Line 5) *Lesion 3* should connect with (Line 12) *lesion* in COREF relation



Example 6. Information in different places (size, and HCC diagnosis)

225

There is a 4.0 x 3.9 cm lesion in segment 7 … which demonstrates arterial enhancement and washout on the delayed images.
Additional wedge shaped lesions of arterial enhancement which do not demonstrates any washout (3/16, 24, 27) which were not demonstrated on prior.
…..
IMPRESSION:
Cirrhosis.
Lesion in segment 7 with arterial enhancement and washout on delayed suspicious for HCC.
Additional wedge shaped lesions as described above which likely represent transient hepatic arterial differences.



After adding reference resolution annotations:

(Line 1) *lesion* connects with (Line 6) *Lesion* in a COREF relation
(Line 2) *lesions* connects with (Line 7) *lesions* in a COREF relation


COREF
Meas
hasMeasurement
TumRef
locatedIn
Anat [Liver]
hasTumEvid
TumEvid [isCancer]
1 There is a 4.0 x 3.9 cm lesion in segment 7 ... which demonstrates arterial enhancement and washout on the
COREF
delayed images.
COREF
COREF
TumRef
2 Additional wedge shaped lesions of arterial enhancement which do not demonstrates any washout (3/16, 24, 27) which were not
COREF
demonstrated on prior.
COREF
3 .....
COREF
SectionHeader [Impression]
4 IMPRESSION:
COREF
5 Cirrhosis.
COREF
COREF
TumRef
locatedIn
Anat [Liver]
hasTumEvid
hasTumEvid
TumEvid [isCancer]
TumEvid [isCancer]
6 Lesion in segment 7 with arterial enhancement and washout on delayed suspicious for HCC.
COREF
TumRef
hasTumEvid
TumEvid [isBenign]
7 Additional wedge shaped lesions as described above which likely represent transient hepatic arterial differences.

7. Annotation Questions

1. What do I do for a general statement with one example?

Focal lesions: Total number: 1. Lesion 1: Segment V, 4.5 cm, series 5, image 53. Diagnosis: Suspicious for HCC, incompletely characterized, needs 3 phase CT or MRI

Mark relation between "Focal lesions" and "Lesion 1" as *particularization*.

2. How should I mark tumor thrombus?

1. Large infiltrative mass involving entire right lobe, infiltrating into the caudate lobe and segment 4 of the left lobe and invading into the right and main portal vein (tumor thrombus) consistent with infiltrative HCC.

Mark relation between "mass" and "tumor" as *particularization*.

3. What if there 2 general statements, but have no number match?

There are multiple arterial enhancing lesions which demonstrate washout consistent with HCC. … Impression: 1. Three arterial enhancing lesions within the liver involving segments 2, 3 and 8 which demonstrate washout concerning for HCC.

Decide if it is a COREF or a *particularization* based on context.

4. Should we connect negated instances?

No other foci of arterial enhancement or washout are demonstrated. …. 2. No other lesions suspicious for HCC.

Do not connect unless it is more specific
For example, connect these:

No nodules or masses are seen in the lungs.

5. Incorrect plural

Should the latter tumor mention ("1 focal lesions") be associated with first ("Focal lesions") or 2nd mention ("Lesion 1")?

Rule: If expression is in sentence format (in oppose to section format), and the number is "1" then it should be considered a specific example.

Focal lesions: 228
Total number: 1:
Lesion 1: Hypervascular with washout, segment 4A, 5.0 x 4.5 cm, image 7 series 9, subcapsular in location.
……
Impression:
Outside report not available at time of dictation:
1 focal lesions in semgnet 4A, typical for HCC.

8. References

[1] W. Yim, T. Denman, S. Kwan, M. Yetisgen. Tumor information extraction in radiology reports for hepatocellular carcinoma patients. To Appear in Proceedings of the American Medical Informatics Association Clinical Research Informatics Summit (AMIA CRI'16), San Francisco. March, 2016.

Appendix F

TUMOR CHARACTERISTICS ANNOTATION GUIDELINES

Annotation Guidelines – Tumor Characteristic Annotation for HCC staging

1. Introduction

In this document we describe the annotation guidelines for labeling set of documents with gold standard annotation of tumor characteristics relevant for hepatocellular carcinoma (HCC) staging.

2. Tumor characteristics for HCC staging

Tumor information is a typical parameter used for cancer staging. Our goal is to collect information for 3 liver cancer staging schemes: the AJCC (American Joint Committee on Cancer), the BCLC (Barcelona Clinic Liver Cancer) and the CLIP (Cancer of the Liver Italian Program). The relevant information for each staging scheme which affects stage are described in the following tables and figures:

AJCC

Stage	Description
I	There is a **single tumor** (any size) that has not grown into any blood vessels. The cancer has not spread to nearby lymph nodes or distant sites.
II	Either there is a **single tumor** (any size) that has grown into blood vessels, OR there are **several tumors, and all are 5 cm (2 inches)** or less across. The cancer has not spread to nearby lymph nodes or distant sites.
IIIA	There is **more than one tumor**, **and at least one is larger than 5 cm** (2 inches) across. The cancer has not spread to nearby lymph nodes or distant sites.
IIIB	**At least one tumor** is growing into a branch of a major vein of the liver (portal vein or hepatic vein). The cancer has not spread to nearby lymph nodes or distant sites.
IIIC	A tumor is growing into a nearby organ (other than the gallbladder), OR a tumor has grown into the outer covering of the liver. The cancer has not spread to nearby lymph nodes or distant sites.
IVA	**Tumors in the liver can be any size or number** and they may have grown into blood vessels or nearby organs. The cancer has spread to nearby lymph nodes. The cancer has not spread to distant sites.
IVB	The cancer has spread to other parts of the body. (Tumors can be any size or number, and nearby lymph nodes may or may not be involved.)

www.cancer.org/cancer/livercancer/detailedguide/liver-cancer-staging

Conclusion: Need to know (1) Single or multiple, and (2) If largest tumor size bigger 5cm

BCLC

Table 5 Barcelona Clinic Liver Cancer (BCLC) staging classification

Stage	PST	Tumor status		Liver function studies
		Tumor stage	Okuda stage	
Stage A: early HCC				
A1	0	Single	I	No portal hypertension and normal bilirubin
A2	0	Single	I	Portal hypertension and normal bilirubin
A3	0	Single	I	Portal hypertension and abnormal bilirubin
A4	0	3 tumors <3 cm	I-II	Child-Pugh A-B
Stage B: intermediate HCC	0	Large multinodular	I-II	Child-Pugh A-B
Stage C: advanced HCC	1-2*	Vascular invasion or extrahepatic spread	I-II	Child-Pugh A-B
Stage D: end-stage HCC	3-4†	Any	III	Child-Pugh C

PST, Performance Status Test; Stage A and B, All criteria should be fulfilled; *, Stage C, at least one criteria: PST1-2 or vascular invsion/extrahepatic spread; †, Stage D, at least one criteria: PST3-4 or Okuda Stage III/Child-Pugh C.

Conclusion: Need to know (1) Single or <=3 nodules, and (2) If largest tumor size >=3cm

CLIP 231

Table 7 Cancer of Liver Italian Program (CLIP) scoring system

Variables	Scores		
	0	1	2
Child-Pugh stage	A	B	C
Tumor morphology	uninodular and extension ≤50%	multinodular and extension ≤50%	massive or extension >50%
AFP (ng/dL)	<400	≥400	
Portal vein thrombosis	No	Yes	

Subramaniam. A review of hepatocellular carcinoma (HCC) staging systems. Chinese Clinical Oncology. Vol 2, No 4, Dec 2013

Conclusion: Need to know (1) Single or multiple, (2) If extension >50%

3. Annotation

While our annotation is motivated for HCC staging, we would like to capture the most general characteristics that allow inference of our information so that the task can be used towards other staging schemes. Below we identify more generalized factors to annotate.

Required Aggregated Information:

1. Given tumor number related to cancer, we can derive AJCC(1), BCLC(1), and CLIP(1).
2. Given the size of the largest tumor related to malignant cancer, we can derive AJCC(2) and BCLC(2).
3. Given if extension of all malignant cancer tumors takes up over 50% of the liver, CLIP (2).

Furthermore, while we are primarily interested in tumors associated with malignant cancer. Knowledge of other benign and indeterminate lesions would be useful to disambiguate for training examples.

Therefore, our target annotations are the following:

1. Tumor number [BENIGN, INDET, ISCANCER, UNK] – the number of tumors that are benign, indeterminate, malignant, or unknown.
2. Largest tumor size [ISCANCER] – the size of the largest tumor related to malignant cancer
3. Tumor extension [ISCANCER] – binary whether or not tumors related to a malignant cancer in the liver has extended beyond 50% of the liver.

Because whether or not malignant tumor extension has gone beyond 50% is difficult to assess with observing the image, we provide a definition described in the following table.

Tumor extension for malignant tumors are considered over 50% if **ANY** of the following conditions are met: -Tumor >= 10 cm ->4 segments involved -majority of right lobe involved -entire left lobe plus some right -some description to suggest much of liver involved e.g. “massive” “very extensive”

Finally, we recognize that radiology reports follow a structure in which the "Findings" section typically explain in detail each tumor, meanwhile "Impression" may offer additional or more summarized information. Therefore, mark gold standard labels at 3 levels:

1. Whole document – considers the entire document for the (1-3) target annotations described before
2. "Findings" section – considers only the "Findings" section for the (1) target annotations
3. "Impression" section – considers only the "Impression" section for the (1) target annotations

4. Annotation decisions

1. Corpus will be pre-annotated with tumor events from previous annotations.
2. Label all quantity ranges in inequality form with "," to separate multiple inequalities, for example if there are 2-3 lesions, then represent >1,<4.
3. If "more than one," "several," or "multiple" are indicated, only assume the number is >1.
4. Several inequalities can be added as in regular arithmetic.
5. If a tumor is both "Indeterminate" and "Benign," label it as Indeterminate.
6. Take the full dimensions for the size of the largest lesion related to malignant cancer. If more than one measurement exists, take the one most precise (in case one version is a rounded version).
7. If there are discrepancies between the tumor measurements (typo or other), take the one from Impressions section first, or the largest otherwise.

5. Annotation example

Findings:
Arterially enhancing **lesion** in segment 7/8 of the liver that is partially exophytic and washes out … measuring 2.9 x 2.8 cm, consistent with HCC.
There is 1.5 x 1.3 cm hyper enhancing subcapsular **lesion** in segment 8 which washes out … consistent with HCC.
There is 0.9 cm **focus** of arterial enhancement in segment 7 … indeterminate.
Impression:
Segment 7/8 enhancing **lesions** arterially enhancing lesion with … washout … definite for hepatocellular carcinoma

Annotations

Section	ISCANCER	INDET	BENIGN	UNK	Largest Malignant Lesion Size	>50%
Findings	2	1	0	0		
Impression	>1	0	0	0		
Whole	2	1	0	0	2.9 x 2.8 cm	NO

Tumors that are indeterminate are not mentioned in the "Impression" section, therefore the "INDET" is written as 0.

5. Annotation Questions

233

1. How do I handle multiple inequalities?

-> First determine if they are *separate inequalities* or *related inequalities*.

Separate inequalities example:

Liver: The liver is cirhootic in morphology **Multiple well-defined T2 hyperintense foci are noted again, representing simple cysts.** … Spleen: Splenomegaly, measuring 17 cm in long axis. **Multiple hypodense nodules are noted in the spleen, representing, Gamma bodies**

For benign:

(hyperintense foci) # of lesions > 1
\+ (hypodense nodules) # of lesions > 1

(# benign entities) # of lesions >2

Related inequalities example:

The lesions are too numerous to count. **The largest** is in the left lobe and is inferior lateral segment 3 and measures 6.7 x 6.8 cm in transverse diameter. It is hypervascular in the arterial phase with washout and a thin enhancing capsule in the venous phase.

"lesions are too numerous" already implies "# of lesions > 1".
Therefore any mention of up to 2 more specific examples does not change the inequality. If there are 3 specific examples, the inequality can then change to ">2".

(# malignant entities) # of lesions >1

2. Should I count tumors that were missed in previous annotations?

-> Yes, adjust for annotation errors from tumor event annotation.

3. Should I count tumors that were missed due to the annotation guidelines from previous annotation?

-> If the tumor is a malignant tumor (e.g. HCC) which was not marked due to annotation guidelines, then **YES** count it.

Large HCC in the proximal and central portion of segment VIII as described above.

The "HCC" may not have been identified as a tumor reference due to annotation rules. For counting number of tumors, DO count these.

-> If there is a mention of a benign or indeterminate tumor, without the measurement, then **NO,** do not count it.

234

Pancreas: Tiny 2 mm **cyst** in the pancreatic head (15/27), unchanged since pror CT from 2007 suggesting benign.

The “cysts” are not identified as a tumor reference or a tumorhood evidence due to annotation rules. For counting number of tumors (benign), DO NOT count these.

Appendix G

ORGAN ADJECTIVE WORDLIST

CUI	PreferredName	Adjectives
C0041600	bone structure of ulna	ulnar
C0009194	bone structure of coccyx	coccygeal
C0035561	bone structure of rib	costal, intercostal
C0018787	heart	cardiopulmonary, cardiorespiratory, cardiac
C0037303	bone structure of cranium	cranial, intracranial
C0001625	adrenal glands	adrenal
C0015392	eye	ophthalmic, binocular, optic, optical, ocular
C0006441	synovial bursa	bursal
C0030274	pancreas	pancreatic
C0038351	stomach	gastric gastroesophageal, pneumogastric, gastroduodenal, stomachal, stomachic
C0024947	maxilla	maxillary, maxillodental
C0041967	urethra	urethral
C0024109	lung	pulmonic, lung-like, pulmonary, pneumogastric, pneumonic, cardiopulmonary, intrapulmonary
C0549207	bone structure of spine	vertebral, intervertebral
C0026845	muscle	muscular, neuromuscular, myoid, intramuscular, musculoskeletal
C0223792	phalanx of hand	phalangeal
C0022646	kidney	nephritic, renal, adrenal
C0262950	skeletal bone	osseous, osteal, bony, ossiferous
C0004457	axis vertebra	axial, axile, biaxial, biaxal, biaxate
C0026367	molar tooth	molar
C0030558	parietal bone structure	parietal
C0039597	testis	testicular
C0033572	prostate	prostatic, prostate
C0039316	tarsal bones	tarsal
C0025526	metacarpal bone	metacarpal

Table G.1: Organ adjective wordlist by using WordNet pertainyms with MetaMap. (Part 1)

CUI	PreferredName	Adjectives
C0030580	parotid gland	parotid
C0040426	tooth structure	dental
C0006655	bone structure of calcaneum	calcaneal
C0448350	scalene muscle	scalene
C0008913	bone structure of clavicle	subclavian
C0030786	hip bone	pelvic
C0559499	biceps brachii muscle structure	bicipital
C1123023	skin	integumentary, percutaneous, hypodermic, transcutaneous, intradermal, endermatic, subcutaneous, endermic, integumental, mucocutaneous, skinny, intracutaneous, intradermic
C0032005	pituitary gland	pituitary, hypophysial, hypophyseal
C0040184	bone structure of tibia	tibial
C0014876	esophagus	esophageal, gastroesophageal
C0016976	gallbladder	biliary
C0034627	bone structure of radius	radial
C1744702	gluteal muscle	gluteal
C0040132	thyroid gland	thyroid antithyroid, thyroidal
C0036037	bone structure of sacrum	lumbosacral, sacral
C0016068	fibula	peroneal
C0042276	vagus nerve structure	vagal
C0036277	bone structure of scapula	scapular, scapulohumeral
C0042232	vagina	vaginal
C0022907	lacrimal gland structure	lacrimal, lachrymal
C0029939	ovary	ovarian
C0024687	mandible	mandibulate, mandibular, inframaxillary, maxillomandibular
C0038293	sternum	sternal
C0005682	urinary bladder	bladderlike, abdominovesical, bladdery
C0816871	skeleton	skeletal, musculoskeletal
C0020164	bone structure of humerus	scapulohumeral
C0013313	dura mater	dural, extradural, epidural, subdural
C0015811	femur	femoral
C0278403	subcutaneous tissue	hypodermal
C0020417	hyoid bone structure	hyoid
C0031050	pericardial sac structure	pericardial, pericardiac
C0011980	respiratory diaphragm	phrenic
C0042149	uterus	intrauterine, uterine
C0037993	spleen	lienal, splenetic, splenic
C0223741	trapezoid bone structure	trapezoidal
C0023884	liver	hepatic, hepatovenous
C0224434	structure of sartorius	muscle, sartorial
C0030647	patella	patellar

Table G.2: Organ adjective wordlist by using WordNet pertainyms with MetaMap. (Part 2)