A Tool for Improving the Longitudinal Imaging Characterization for Neuro-Oncology Cases

Ricky K. Taira, Ph.D[1], Alex Bui, Ph.D[1], William Hsu, BS[1], Vijayaraghavan Bashyam, MS[1], Emily Watt, MLIS[1], Lewellyn Andrada, BA[1], Suzie El-Saden, MD[1], Timothy Cloughesy, MD[2], Hooshang Kangarloo, MD[1]

[1]Medical Imaging Informatics, UCLA, Los Angeles, CA
[2]Neuro-Oncology Clinic, UCLA Medical Center, Los Angeles, CA

Abstract

We describe the development of a prototype tool for the construction of longitudinal cases studies that can be used for teaching files, construction of clinical databases, and for patient education. The test domain is neuro-oncology. The features of the tool include: 1) natural language processing tools to assist structuring report information; 2) integration of imaging data; 3) integration of drug information; 4) target data model that includes the dimensions of space, time, existence, and causality; 5) user interface that provides three levels of information including overview, filtered summarization, and details on demand. The results of this preliminary work include a full prototype for neuro-oncology patients that allow users an efficient means for scanning a patient's imaging and support data.

Introduction

In this work, we investigate the requirements and prototype a toolkit for improving the documentation and modeling of longitudinal case studies for the domain of neuro-oncology. The resulting case studies can then be used to compile a clinical database. We define clinical databases as collections of patient cases intended for clinical reviews (e.g., tumor board), research (non-prospective), physician training, and patient education. The data collection and management protocols for these cases are not held to the same standards and formalisms associated with prospective clinical trials. For example, there is usually no explicit hypothesis associated with each patient encounter; there are no formal data collection forms; there are no formal documentation methods.

The purpose of this research is to investigate the requirements and development of prototype tools to facilitate the construction of case studies. In particular, we investigate tools for automatically structuring clinical documents, tools for editing/validating structured representations, and tools for filtering/visualizing patient case models. The current products of this research are a prototype system with design explanations and architecture for supporting future requirements needs.

We do not address the issue of confidentiality in this paper. We refer to some of our prior publications on this issue[1, 2].

Clinical Environment and Needs Assessment

Physicians involved in busy clinical services currently have few options concerning methods for creating quality research patient databases. Perhaps the most common means is the simple spreadsheet with patients listed along one dimension, and the patient variables along the other. This approach often lacks any advanced informatics tools and/or organizing principles. For example, time-oriented and causal associations are not well represented in this form. Multi-media information is often excluded. In practice, these databases that are often created at world-class facilities remain private. Their valuable observations are not sharable with other researchers. The data format and representation formalism exist in an *ad hoc* form and are not well modeled or easily visualized.

In this paper, we provide an initial exploration toward development of a general tool for formally capturing the longitudinal observations, interventions, and causal theories associated with patients with neuro-oncology disorders. The justification for this application domain includes the following:

Prevalence: In the US, over 40,000 patients are diagnosed annually with primary brain tumors[3]. Although there have been several decades of research, life expectancy remains 10-18 months.

Imaging: The NCI Cancer Imaging program has emphasized imaging's role in cancer detection, characterization, and monitoring. Thus far, imaging data for cancer patients have been poorly connected to clinical data and outcomes. Conventional MR imaging with gadolinium-based contrast agents is by far the most widely used modality for this purpose. Several studies have been conducted to investigate

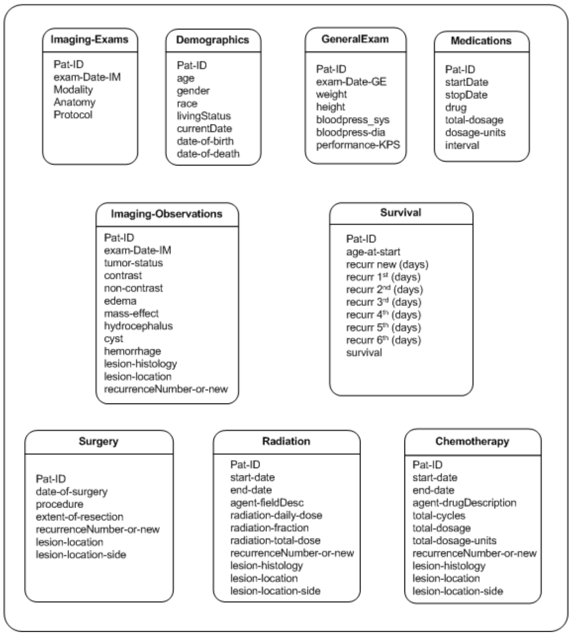

Figure 1. Partial schema of existing neuro-oncology patient database (genetic data not shown).

how tumoral properties and behaviors manifest themselves on clinical MR image studies[4, 5, 6, 7].

Clinical Trials: Methods for structuring and accurately documenting imaging and clinical observations from routine studies has been mentioned as a major goal by the CaBIG consortium due to the fact that only 2% of cancer patients are on formal clinical trials. Methods for documenting and characterizing non-trials cases for compilation of clinical databases can accelerate basic and evaluation research efforts as well as contribute to education efforts to students and patients[8, 9].

Description of Existing Clinical Database

For the past ten years, clinicians at the UCLA Neuro-oncology program have been carefully constructing a structured clinical neuro-oncology database for patients with primary brain tumors, recording key variables on histopathology, demographics, living status, interventions (*e.g.,* surgical resection; chemo/radiotherapy, etc.), time to progression, time to survival, medications, and key radiographic findings regarding tumor appearance (Figure 1). Data for over 2,018 confirmed brain tumor patients have been collected to date. Data for these cases have been manually abstracted from the patient record by a qualified researcher, or recorded manually by physicians managing the patient case into a set of spreadsheets. Gene expression data is now also being collected. From this database, 110 patients with histologically proven glioblastoma and 41 patients with anaplastic astrocytoma were used in a study investigating the correlation of MRI findings to survival in patients with high-grade gliomas[4].

A large part of this research is motivated by improving the characterization of radiographic imaging evidence. This important functionality has been lacking in clinical research databases. A number of recent works have shown the importance of imaging in characterizing brain tumor and remains the most common method of monitoring patient response to therapy. It provides both: 1) gross details regarding tumor extent and the effects of tumor and treatment on normal tissues as well as 2) characterization of tumor micro-environment (e.g., local O2 levels, acidity, tumor cell density, acidity, etc.)[5, 10, 11, 12, 13]. A tool to integrate imaging data with clinical information is needed.

System Overview

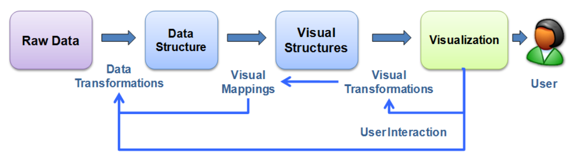

Figure 2. System overview

Figure 2 shows the high level overview of the system. The input is the collection of documents for a given patient. A natural language processing module[14] analyzes the free text reports and outputs a formal representation of targeted information content. Currently, the documents processed by NLP system include radiology, pathology, and discharge summaries. The resulting formal information model includes characterizations of time, space, existence and causality for the given patient. The formal information model can then be mapped to a variety of visual structures such as trees, graphs, tables, image stacks, and timelines. These visual structures can then be interpreted by various workstation rendering programs to visualize the data and interact with the user.

System Design

The summarization and organizational framework for medical records has its roots in the work of Larry Weed and his notion of the Problem-oriented Medical Record. Many practical issues regarding its clinical realization linger including: 1) Data Entry - how does information in the case study get entered? 2) Information Content - What information should be structured and to what granularity? 3)

Representation - How does the representation get standardized? 4) User Interface - What user interface and data visualization paradigm should be deployed to maximize the cognitive understanding and reasoning capabilities of a physician/researcher for a particular task while minimizing user fatigue?

The functionality of the targeted application was first designed onto paper by an informatician and neuro-radiologist. The informatician first analyzed all imaging, pathology, and discharge records for a complex neuro-oncology patient case. The goal was to specify the types of information to be extracted and the presentation method specifically for longitudinal monitoring of tumor burden, for observing the effect of anatomical abnormality on physiologic function, and to infer causal effects of observed data patterns. An organizational framework that includes the following relational aspects fundamental to any disease phenomenon was proposed: 1) Space-Continuum – The anatomic focus and extent of disease is central to understanding of many pathologic processes; 2) Time Continuum – disease phenomena exists within a space-time continuum. We need basic references within this continuum to better understand the extent and dynamics of a disease; 3) Existence status – what things exist within this continuum, how strong is our belief in this existence proposition, and how does this existence change over time (new, old, resolved, recurrent); 4) Causality – what is the genesis of how things came to exist either due to the natural course of a disease and/or its response to a medical intervention. A paper-based summarization of the single patient case underwent an iterative design cycle of design, construction, and evaluation by various clinicians involved in managing patients with brain tumors.

Structuring Text Reports

Sentence-Level Processing – We utilized an existing NLP toolkit to analyze every sentence of a given patients radiology, pathology, and discharge summary reports[14]. The processing is performed within a development environment with all patient identifier information removed. For each sentence, the NLP system performs the extraction and classification of the following types of information: 1) mention of a finding/disease process; 2) the location description of the finding; 3) existence description of the finding; 4) temporal modifiers of the finding; 5) associations with other findings (e.g., causal, differential interpretation, and co-occuring). The identified findings as well as their corresponding anatomical locations were coded to UMLS.

Co-Reference Resolution – Descriptions of findings described over serial studies need to be linked in order to capture the behavior (*i.e.,* state changes) of a finding that is being followed. Co-reference resolution is the problem of linking all descriptions for a single finding at all structural levels including: within a sentence; between sentences in the same report; and across multiple reports. The features for the co-reference classifier include:

Linguistic cues. Words or phrases such as *a*, *the*, *this*, *another*, *a second*, or *a new*, are indicators of whether the current reference is a new finding or a finding mentioned previously in the text.

Semantic class agreement. This identifies candidate pairs as either lexically identical, conceptually identical, generalization / specialization, group / member, or none.

Syntactic agreement. The co-referent candidate pairs should agree in number. By way of illustration, the pairing {solitary mass, lesions} is likely not an equivalence pairing nor is the pairing {masses, it}.

Frame description compatibility. The compatibility of attributes contained in the NLP sentence-level frame descriptions are assessed. For example, a mass that is 3 cm in the right anterior temporal lobe and a 2 cm mass located in the right frontal lobe are not consistent descriptions.

Existence and behavior agreement. A rule base is used to identify existence descriptions that are incompatible. For instance, a tumor cannot have the following existence pairings: {newly seen, increasing in size}, {resolved, increasing in size}, {again seen, new}, {recurrent, no longer seen}. Likewise, the changes in an entity must be consistent (*e.g.,* a tumor cannot both increase and decrease in size at the same time).

Expected discourse sequence. Features that provide topical context such as whether the candidate words exist in the same sentence, the same paragraph or the same section were used as well as if one candidate was in the findings and the other in the conclusion.

A maximum entropy classifier was used to build a statistical model to combine the features into a single joint probability model. The classifier was trained on 50 patient cases that had an average of 5 imaging studies each.

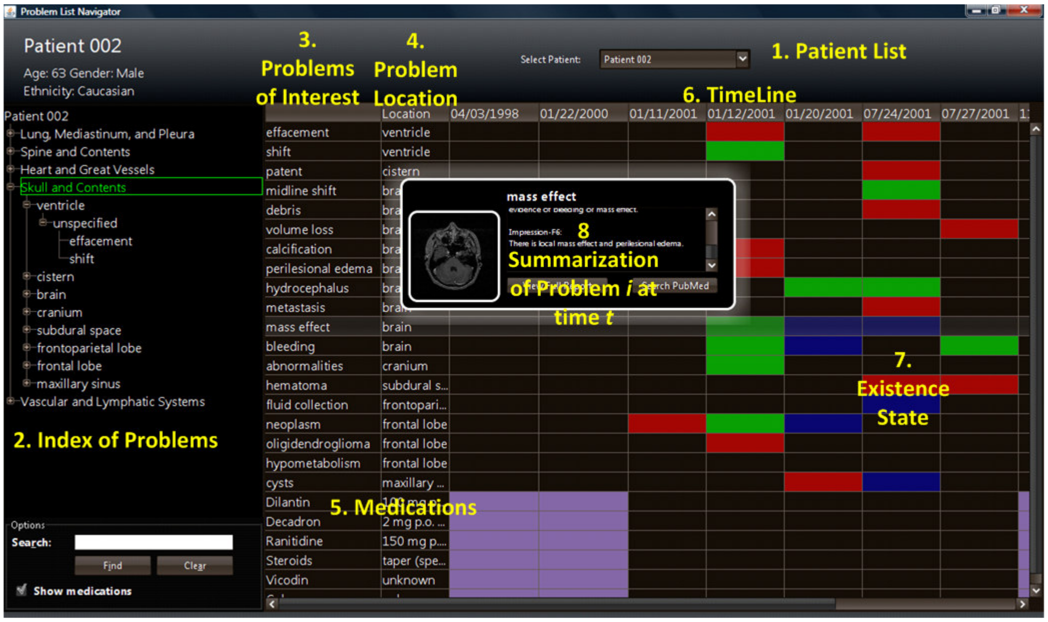

Figure 3. Screenshot of prototype application.

Prototype Application Design

The prototype application includes three levels of information summarization to facilitate both breadth and depth of detail: overview (patient level), zoom and filter (targeted summarizations), and details on demand (source reports and image data)[15]. Figure 3 shows a screenshot of the prototype application. (Note: numbers in figure correspond to descriptions below):

1. Patient list - Can be used to browse patients which are part of the research database.
2. For a given patient, there is an index of all problems over time. The problems are categorized according to the diagnostic index of the American College of Radiology.
3. The user can pick a category get a listing of all problems over time under the select category.
4. For each problem, the most specific location that encompasses the disease extent is listed.
5. A list of medications is also presented on the timeline. Cells show the dosage information.
6. Problem descriptions are summarized along a custom timeline for the patient.
7. Cells in the timeline are color coded to show existence information. For example, red means the finding is confirmed by some study; green means that there is no evidence of the finding; blue means that the problem/finding still is a differential diagnosis.
8. As the user hovers over a cell in the timeline, a pop-up box is displayed in real time showing the report sentences describing a particular finding. Additionally, if an imaging study is available, an icon represented by the central slice of study is shown. This gives the user a quick means of scanning the highly relevant information for a given problem over time. The pop-up box includes buttons to allow the user to access the full report corresponding to the time cell as well as access to the full imaging study.

Results

Performance results of the sentence level NLP were previously evaluated and summarized here for reference[14]. The NLP parser evaluation drew from a test sample of three hundred random sentences, with an average of 17.12 word tokens; the average number of unknown tokens per sentence was 0.086; the average number of conjunctions per sentence was 0.601; and the average number or prepositions per sentence was 1.858. The performance statistics were: recall 84.99%; precision 89.95%.

For the task of just identifying mentions of findings and problems, we obtained a precision of 87% and a recall of 96% for a test set which contained 592 true positive instances.

The co-reference resolution algorithm was tested using a 10-fold cross validation design utilizing the fifty patient cases which had an average of 5 imaging studies each. The results were a precision of 72.0% and a recall of 63.1%.

Discussion

Imaging data is the most in-vivo method for evaluating the effect of therapies and monitoring the progression of primary brain tumor disease. Thus far, imaging has not been prevalent within research databases in documenting the status of patients undergoing a particular therapeutic plan. In this paper, we develop a prototype station that utilizes NLP methods to assist in identifying all findings and diseases for a given patient, the corresponding location, its existence description, and causal inferences. This structuring process allows the front end application to tie together text and imaging data related to a finding of interest over time. The user can access the report descriptions and imaging data in real time by moving a mouse pointer over the desired finding-time point cell in the timeline.

The long term goals of this project are to allow researchers to "decompose" the documentation for a patient case along the lines of individual findings, and thread instances of these findings over time. Currently, we summarize the state of a finding at any given moment in time by isolating the particular sentence within a report that mentions the finding. Thus, the assessment of the finding over time is the responsibility of a human, rather than machine. The future direction is to have a NLP output that models the information within these isolated sentences.

Longitudinal case studies that include imaging documentation are important for evaluation research in neuro-oncology: How did the tumor characteristics and patient symptoms change during the course of the patient's therapy. Clinical educators can benefit from longitudinal studies by allowing students to experience the complexity of a full patient report where a student needs to learn to distinguish between noise and/or independent problems relative to the current clinical questions.

Acknowledgements

The authors would like to thank DIA, GYU, MA, and MA for their valuable feedback to this work. This work was supported in part by NIH Grants P01-EB00216 and R01-EB002247.

References

1. Taira RK, Bui AAT, Kangarloo H. Identification of patient name references within medical documents using semantic selectional restrictions. Proc AMIA, 2002; pp. 757-761
2. Bui AAT, Morioka C, Dionisio JD, Johnson DB, Sinha U, Ardekani S, Taira RK, Aberle DR, El-Saden SE, Kangarloo H. OpenSourcePACS: An Extensible Infrastructure for Medical Image management. Inf Technol Biomed 11(1):94-109, 2007.
3. Jenkinson MD, Du Plessis DG, Smith TS, et al. Histological growth patterns and genotype in oligodendroglial tumors: Correlation with MRI features. Brain, 129:1884-1891, 2006.
4. Pope WB, Sayre J, Perlina A, Villablanca JP, Mischel PS, Cloughesy TF. "Magnetic resonance imaging correlates of survival in patients with high grade gliomas," Am J Neuroradiology, 26:2466-2474, 2005.
5. Li GZ, Yang J, Ye CZ, Geng DY. "Degree prediction of malignancy in brain glioma using support vector machines," Computers in Biology and Medicine, 36:313–325, 2006.
6. Aghi M, Gaviani P, Henson JW, Batchelor TT, et al. Magnetic resonance imaging characteristics predict epidermal growth factor receptor amplification status in glioblastoma. Clin Cancer Res, 11(24):8600-8605, 2005.
7. Higano S, Yun X, Kumabe T, Watanabe M, Mugikura S, Umetsu A, Sato A, Yamada T, Takahashi S. "Malignant astrocytic tumors: Clinical importance of apparent diffusion coefficient in prediction of grade and prognosis," Radiology 241(3):839-846, 2006.
8. Altman M. "The clinical data repository: a challenge to medical student education," J Am Med Inform Assoc., 14:697-699, 2007.
9. Black N. "Using clinical databases in practice," British Medical Journal 326:203, 2003.
10. Gillies RJ, Raghunand N, Karczmar GS, Bhujwalla ZM. "MRI of the tumor microenvironment," J of Magnetic Resonance Imaging 16(4)430-450, 2002.
11. Nakano T, Asano K, Miura H, et al. Meningiomas with brain edema. Radiologic characteristics on MRI and review of the literature. J Clin Imag, 26:243-249, 2002
12. Maldaun MV, Suki D, Lang FF, et al. Cystic glioblastoma multiforme: Survival outcomes in 22 cases. J Neurosurg, 100:61–67, 2004.
13. Megyesi JF, Kachur E, Lee DH, et al. Imaging correlates of molecular signatures in oligodendrogliomas. Clinc Canc Res, 10:4303-4306, 2004.
14. Taira RK, V Bashyam, Kangarloo H. A field theory approach to medical natural language processing. IEEE Trans Information Technology in Biomedicine, 11(4):364-375, 2007.
15. Shneiderman, B. Designing the User Interface: Strategies for Effective Human-Computer Interaction, 3rd ed. Reading, Mass.: Addison, 1998