HHS Public Access

Author manuscript

Proc IEEE Int Conf Healthc Inform Imaging Syst Biol. Author manuscript; available in PMC 2016 August 26.

Published in final edited form as:

Proc IEEE Int Conf Healthc Inform Imaging Syst Biol. 2011 July ; 2011: 221–228. doi:10.1109/HISB.2011.3.

A Case-based Retrieval System using Natural Language Processing and Population-based Visualization

William Hsu¹, Ricky K. Taira¹, Fernando Viñuela², and Alex A.T. Bui¹

William Hsu: willhsu@mii.ucla.edu

¹Medical Imaging Informatics Group, Department of Radiological Sciences, University of California, Los Angeles, CA

²Division of Interventional Neuroradiology, Department of Radiological Sciences, University of California, Los Angeles, CA

Abstract

Electronic medical records capture large quantities of patient data generated as a result of routine care. Secondary use of this data for clinical research could provide new insights into the evolution of diseases and help assess the effectiveness of available interventions. Unfortunately, the unstructured nature of clinical data hinders a user's ability to understand this data: tools are needed to structure, model, and visualize the data to elucidate patterns in a patient population. We present a case-based retrieval framework that incorporates an extraction tool to identify concepts from clinical reports, a disease model to capture necessary context for interpreting extracted concepts, and a model-driven visualization to facilitate querying and interpretation of the results. We describe how the model is used to group, filter, and retrieve similar cases. We present an application of the framework that aids users in exploring a population of intracranial aneurysm patients.

Keywords

Case-based retrieval; natural language processing; knowledge representation; information visualization

I. Introduction

Given the data-intensive nature of today's clinical environment, large amounts of data such as clinical narratives, images, ordered tests, and hospital administrative data are being generated routinely. The medical community has become increasingly cognizant of the data's value for secondary applications such as measuring the quality of healthcare delivery, performing public health surveillance, and developing patient health records [1]. In addition to important socio-organizational issues related to data use and patient privacy, several technical challenges exist in working with observational clinical data: how can one integrate

heterogeneous data from different sources and institutions, standardize the meaning and representation of collected data elements, and permit efficient exploration, cohort identification, and comparative analysis [2]? We present one approach for addressing these challenges: a retrieval framework that structures, organizes, and visualizes patient data in a large clinical repository. The goal is to facilitate exploration of a large dataset consisting of multiple patients to support tasks such as understanding intrinsic patterns within a population (*e.g.*, proportion of patients who have undergone a specific procedure). As part of the framework, we have implemented several informatics tools to leverage clinical data for research: 1) extracting information from unstructured clinical narratives; 2) mapping extracted data elements to a standardized representation; and 3) facilitating the visual exploration of a large dataset using a direct manipulation paradigm. The system incorporates an open-source natural language processing system called clinical Text Analysis and Knowledge Extraction System (cTAKES) developed at the Mayo Clinic to identify relevant biomedical concepts and attributes from clinical reports [3]. Here, we use a disease model to provide a unifying representation to map extracted information, explicitly modeling how different concepts are related. The model is organized hierarchically, representing variables, states, and their relationships for a given disease. The contextual information provided by the model can be used to inform the natural language processing system and compute a similarity measure for comparing patient cases. We have created a mapping tool that links extracted concepts to the disease model. A population-based visualization has been developed to summarize the entire patient cohort in a single display. A user specifies query attributes through a form-based interface; results are reflected in the display by visually highlighting the corresponding proportion of patients who match those attributes. We discuss how our approach is applied towards exploring a dataset of 495 intracranial aneurysm patients. The goal is to understand the evolution and treatment of this debilitating disease. We highlight how our tools can be used to answer questions related to the population (*e.g.*, which patients have undergone surgical clipping but not endovascular embolization), stratify the population into semantically related groups (*e.g.*, anatomical outcome, Fisher grade), and identify other patient cases that have similar findings to a query case.

II. Related Work

A. Structuring Clinical Narratives

Clinical narratives contain useful descriptive information related to the patient, history, disease manifestation, prescribed interventions, and outcomes. However, much of this information is often in a form that is not amenable to analysis because of its free-text, unstructured nature. A trade-off exists between using structured reporting, which encourages standardized data collection and reuse, and unstructured reporting, which allows for expressivity and flexibility [4]. The advantage of developing natural language processing (NLP) systems is that it does not change the *modus operandi* for clinical reporting or restrict expression: a program interprets the physician's free-text descriptions and automatically formulates the information into knowledge structures (*e.g.*, logic-based frames). The recent releases of several open-source text processing frameworks such as General Architecture for Text Engineering (GATE) and Unstructured Information Management applications (UIMA)

have fostered development of new tools in medical NLP. Applications of these systems have been demonstrated across a gamut of targeted information extraction tasks (*e.g.*, smoking status, drug interactions, specific imaging findings) from sources such as radiology reports, discharge summaries, clinical trial participant information, and published literature [5]. For example, one system utilizes cTAKES to discover peripheral arterial disease cases from radiology reports; it achieves a recall of 93% when evaluated against an expert-derived gold standard. These recent works have demonstrated the promise of medical NLP in extracting and structuring information from clinical narratives.

B. Representing Extracted Concepts

Another longstanding informatics challenge is how to best represent extracted concepts from medical reports. In specific cases, NLP results can be codified to declarative representations (*e.g.*, Unified Medical Language System, UMLS) if the concept is simple and self-contained. In many practical situations, however, the context of the finding is critical and must also be captured to completely understand the observation itself. Early works in representing observational data have emphasized the importance of differentiating objective facts from observed, perceived facts and capturing the requisite context to accurately interpret the information later [6]. More recently, researchers have described systems that map extracted concepts to domain-specific models, which capture all of the relevant problems, findings, and attributes for a specific disease. One example of such a system is the Medical Text Analysis System/Pathology (MedTAS/P) [7], which structures pathology reports into a codified cancer disease model. They map identified terms to the International Classification of Diseases-Oncology (ICD-O v3), achieving an F-score of 0.96, 0.97, and 0.95 for anatomical site, histology, and tumor grade value, respectively. [8] presents a phenomenon-centric data model in which medical problems are made up of lower-level phenomena: each medical problem is associated with a set of (ab)normal findings, states, behaviors, and theories that provide explanations for the problem under investigation. In [9], the authors demonstrate how patient data stored in a relational database is mapped to a custom ontology that is used to facilitate information retrieval. They developed a prostate cancer ontology for this purpose, utilizing a combination of NCI Thesaurus (NCIt) and Foundational Model of Anatomy (FMA) to represent common concepts found in the domain.

C. Visualizing Population of Patients

Visualizations that present data representing a large population face unique challenges in comparison to those that solely examine individual records. Population-based visualizations need to support: 1) a greater quantity and variety of data; 2) different views of the data with varying levels of abstraction (*e.g.*, from overview to details); and 3) visual information seeking through interactive filters and querying mechanisms that help expound patterns in the data. Treemaps have been shown to be effective for visualizing a large number of hierarchically organized data using nested rectangles. DBMap [10] utilizes a treemap to organize data stored in a neuroscience database to support epilepsy research; the treemap is dynamically generated based on a user-defined hierarchy. Another approach, stadium diagrams, uses an array of person icons that are organized in a grid. Icons can be spatially

grouped and colored based on some attribute of interest (*e.g.*, health state); [11] uses this representation to visualize changes in health status for patients in a population based on actions given during the course of a management strategy. [12] examines how temporal trends across a population can be elucidated. The tool provides filters that allow a user to specify a sentinel event, realigning multiple timelines (representing different patients) around the selected time point, permitting the user to find patterns of interest (*e.g.*, adverse reaction). A supported query would be: *which patients have abnormal laboratory test values within a week of receiving drug x?* VISITORS [13] is an example of an application that incorporates timelines and parallel coordinates to highlight relationships between concepts (*e.g.*, lab values) for multiple patients simultaneously and facilitate identification of temporal relationships. Their system is designed to answer queries such as *what associations exist between hemoglobin, platelet count and total bilirubin for patients seen in 2005?*

In the following sections, we present the components of our case-based retrieval framework. We demonstrate an application of the framework to explore a large dataset of intracranial aneurysm patients. We conclude by discussing current limitations and future directions of our work.

III. Retrieval Framework

Our case-based retrieval framework incorporates the following components: 1) a comprehensive disease model that encodes domain knowledge drawn from sources such as biomedical ontologies, published literature, and clinical datasets to capture the context associated with observations related to patient state; 2) extraction and mapping tools that identify relevant biomedical concepts and phrases from clinical narratives and map them to a standardized disease model; and 3) an interactive visualization that facilitates exploration, filtering, and case-based reasoning on a population of patients. The flow of information between the framework's components is illustrated in Figure 1; a description of each component is provided in the following sections.

A. Disease Model

Given the highly fragmented and heterogeneous nature of observational clinical data, disease models play an integral role in bridging gaps between variables and capturing necessary context to interpret the data. The purpose of the model is to: 1) enumerate all relevant variables and states that are associated with a particular disease and its management; 2) explicate the relationships among variables, integrating descriptions of the disease state across different scales (*e.g.*, genetic, cellular, organ); 3) provide a consistent representation of variables, reusing knowledge from other ontologies whenever possible; and 4) drive applications that utilize information in the model to structure observational data and enable comparisons between patient cases. In generating the disease model, three sources of information are referenced: existing biomedical ontologies, scientific literature identifying pertinent variables specific to the disease, and a combination of expert knowledge and available clinical data. Existing knowledge sources such as the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) and NCI provide the initial set of terms. From these sources, a subset of variables is selected based on their relevance to the disease.

Additional variables are enumerated based on manual examination of published literature (*e.g.*, systematic reviews), expert opinion (*e.g.*, experiential knowledge), and clinical data (*e.g.*, variables collected as part of a disease registry). Each variable in the disease model is recorded along with associated metadata such as synonymous terms, external unique identifiers, semantic types, and states:

- Synonyms help identify variants of a term (including abbreviations) that may be used to represent a concept in the clinical narrative. For example, high blood pressure can also be expressed in multiple ways: hypertension, increased blood pressure, or HTN among others. Any instance of these terms found in a report should be mapped to the variable "hypertensive disease."
- External unique identifiers connect a variable to other ontologies (*i.e.*, through a UMLS concept unique identifier), whenever possible.
- The semantic type permits concepts to be grouped using higher levels of abstraction. For example, the variable "seizures" is assigned the semantic type "sign or symptom." This information allows concepts to be retrieved using higher-level classes (*e.g.*, identify all signs and symptoms related to lung cancer).
- The variable state indicates a specific value that a variable could be assigned. For example, the variable "tumor size" could be represented (arbitrarily) as four states: small (<4 mm), medium (4–8 mm), large (8–12 mm), and giant (>12 mm). Variable states can be ordinal, nominal, or interval.

B. Extraction and Mapping

The first step is to extract relevant findings and attributes related to the disease of interest from unstructured clinical reports. The framework incorporates an open-source system called cTAKES to perform information extraction given its extensibility, modularity, and growing popularity within the informatics community. cTAKES performs several tasks including sentence boundary detection, tokenization, word normalization, parts-of-speech tagging, shallow parsing, and named entity recognition. The system also identifies negations (*e.g.*, no evidence of) and status (*e.g.*, current, family history). The system is customizable, allowing users to include additional lookup dictionaries. To incorporate the UMLS, the Metathesaurus is loaded into a relational database, and the dictionary lookup configuration is altered to query the database. Additional domain-specific terms found in the disease model can be re-encoded as a delimited text file and incorporated into cTAKES by providing a revised lookup method that searches the custom dictionary. Once processed, named entities and attributes identified in each report is encoded using eXtensible Markup Language (XML).

Based on the output of cTAKES, we then created a mapping tool to link extracted concepts with variables and states represented in the disease model. The purpose of the mapping tool is to utilize information encoded in the disease model to contextualize extracted concepts (*e.g.*, definitions, restrictions) and to facilitate similarity matching in case-based reasoning. Not all terms identified by cTAKES are mapped; matching terms to concepts in a disease

model ensures that matched terms are relevant to the domain of interest, filtering out unrelated information. In many cases, this mapping is straightforward; it simply matches the concept unique identifiers provided by UMLS. However, special mapping is needed to address situations in which: 1) an extracted concept does not map to any existing variables in the model; or 2) an attribute value does not directly correspond to a variable state. For example, tumor size is often expressed in radiology reports as one or more continuous numbers representing measurements taken along different axes (*e.g.*, anterior-posterior length). While cTAKES identifies these values as "Measurement" objects, it does not automatically associate this value to the tumor. The mapping tool makes an assumption that entities and attributes identified in the same sentence are related. In addition, numerical values in models are often represented as discrete states. The mapping tool provides classifiers that categorize a continuous value into one of the predefined states. For example, given the sentence, "the tumor was 3 mm in diameter," the classifier would identify 3 mm and map this value to the state "small (<4 mm)". We utilize a process similar to the mapping table developed for classifying whether a medical problem has worsened, improved, or remained unchanged in comparison to a previous observation as described in [14].

C. Querying and Visualization

The value of structuring observational clinical data is achieved when the user is able to generate new knowledge from the available data; this goal can be attained by providing users with a flexible interface that supports a rich set of queries. In this framework, we describe three ways that users can interact with the underlying data: filtering patients based on attributes, stratifying patients into semantic groups, and retrieving similar patient cases using case-based reasoning.

Filtering—The most fundamental type of interaction supported is specifying one or more attributes that are used to filter the entire population for matching cases. Cases that match are highlighted in red. The user specifies a constraint by inputting the value for each attribute using a form-based interface consisting of text fields, check boxes, and drop down lists. Attributes represented in the interface are drawn from variables and states represented in the disease model. A user can select one or more states for each variable; the result will be the union of cases that match. Users can also sequentially add multiple constraints to formulate complex queries.

Grouping—When performing comparative studies that assess the effectiveness of two different treatments, it is often necessary to identify homogeneous groups of patients for comparison to control for confounding and selection bias. We can use the process of grouping to provide insights into whether the available population is capable of supporting the types of comparisons desired by the user. First, the user selects one or more variables by which the population is stratified. For example, a population of brain tumor patients can be divided by survival time (*i.e.*, < 1 year, 1 – 2 year, 2+ years) and by chemotherapy type (*e.g.*, temozolomide, carmustine, lomustine). Then, based on initial inspection, the user can visually estimate the size of each group based on the relative size of each partition. Finally, the user can specify additional constraints through the query interface, highlighting patient

cases that match in each subgroup. The resulting display allows users to determine whether subgroups are relatively homogeneous based on the pattern of highlighted cases (*i.e.*, subgroups with the majority of patients highlighted are more homogeneous).

Reasoning—Given a large database of structured patient cases, one application is to identify and rank cases in the population based on inputted characteristics using a similarity measure. A variety of applications has been described in the past, showcasing how case-based reasoning utilizes previous cases as a precedent for interpreting unseen cases [15, 16]. In medicine, case-based reasoning has been shown to inform clinical decision support by identifying past patient cases with known treatments and outcomes that are similar to a new patient case [17, 18]. In this work, we explore how the disease model, which is represented as a graph, can be used to facilitate the comparison of patient cases. Our approach is based on previous work that explored four graph-based metrics for computing inter-patient distance using SNOMED CT [19]. Patient cases are represented as a vector of concepts that are extracted from the medical record. This information is mapped to variables and states in the disease model using the mapping tool. The model's graph structure is then used to compute similarity: essentially, the distance metric is the average minimum number of links between a variable in the query patient to a variable in an indexed case. The user initiates the retrieval process by selecting a patient as the query case. The system will automatically compute the distance metric for the query case along with all other cases indexed in the database. Results are shown as a tabular listing of other cases, ranked by increasing distance. Users can then select an entry in the listing to retrieve detailed information about the patient such as clinical reports and images.

User interface—The interface is divided into three panes, as depicted in Figure 2: 1) the population pane is comprised of an array of person icons arranged in a rectangular grid, representing the entire population; 2) the query input and processing panes contain a set of form fields for inputting values for attributes as well as controls for adjusting and executing the query; and 3) a results pane displays relevant statistics and the results of case-based reasoning.

IV. Case Study: ICA Database

During the development of this system, we worked with an interventional neuroradiologist to apply the case-based retrieval framework towards understanding a population of patients at risk of intracranial aneurysms (ICAs). ICAs are complex lesions that are only partially understood. While their true incidence rate is unknown, an estimated 1–6% of the population is affected [20]. Within the affected population, an estimated 20–50% of ICAs will rupture within an individual's lifetime [21]. Variables such as smoking, aneurysm size, and age are commonly used to determine the risk of rupture, but little is known about the true etiology of ICAs and optimal treatment is still largely debated. Presently, three common types of treatments exist: observation, surgical clipping, or endovascular embolization. Despite these treatment options, the overall mortality rate for a ruptured aneurysm remains approximately 50% [22]. Undoubtedly, early diagnostic and better therapeutic strategies are needed to prevent the poor outcomes observed today. To develop the case study, the neuroradiologist

identified specific types of questions that he wished to ask about the patient population to better understand the treatment options and long-term outcome of ICA patients seen at UCLA.

A. ICA Disease Model

The disease model is based on concepts originally enumerated by the @neurist project, which is a consortium of European institutions [23] who collaborate to develop tools and standards for characterizing cerebral aneurysms. The original ontology contains approximately 2,800 concepts enumerated by manually surveying a selection of data terminology and knowledge sources such as the UMLS, FMA, and Gene Ontology for molecular concepts. The model represents common clinical findings, software tools to enable data fusion and querying across multi-institutional databases, and generation of systematic reviews that promote better practices for the management of ICA patients. We created our model using the @neurist ontology as a foundation. We represented each concept in the ontology as a variable in our hierarchical model. The model encodes "is-a" relationships in the model as directed arrows. Each variable is associated with two or more states. Figure 3 depicts how the variable "hemorrhage" is represented in our model. For example, states associated for "subarachnoid hemorrhage" include "acute", "non acute", and "none". In addition, we supplemented the model with additional variables identified through review of systematic reviews (*i.e.*, Cochrane collaboration), reporting guidelines such as [24], and consultations with our clinical collaborators.

B. Patient Data and Extraction

We obtained a set of 495 patients who had been seen at the UCLA Division of Interventional Neuroradiology with a confirmed diagnosis of intracranial aneurysms. While most of the patient cases in the set had previously been reviewed and manually abstracted by a physician, not all of the fields in Table I were extracted and populated into the structured database. Moreover, we wish to automate the process of adding new prospective cases into the database. Hence, we performed a preliminary assessment on the extraction tool to evaluate its ability to identify specific concepts and attributes in the ICA domain. We retrieved the full medical records for 20 of the 495 patients. We used DataServer [25], an application framework that enables real-time querying of clinical repositories, to access reports, labs, and images from UCLA Medical Center's health information systems. The output of DataServer is a single, uniform XML representation, abstracting underlying data sources and access mechanisms. Data acquisition and handling protocols were previously reviewed and approved by the institutional review board.

Out of the entire medical record, we examined only admission notes, radiology reports, surgical reports, discharge summaries, and consultation notes that had relation to ICA, resulting in a total of 567 documents examined from the 20 patient cases. We configured cTAKES to utilize the 2010AA UMLS release and a custom dictionary containing domain-specific variables from the ICA disease model. Concepts that had matching terms identified as negated by cTAKES were assigned states such as "none" or "not present" in the disease model. We assessed the ability of the extraction tool to identify relevant findings and

attributes. We targeted a subset of variables listed in Table I: anatomical location, thrombosis, aneurysm shape, aneurysm size, and calcification. To generate a gold standard, one medical resident was asked to manually review every report, identifying values for the selected set of variables. cTAKES then processed the same documents, generating a set of extracted concepts. In summary, our initial tests showed that the system achieved a recall of 77% for anatomical location, 100% for identifying thrombosis, 78% for categorizing aneurysm shape, 97% for extracting aneurysm size, and 100% for identifying presence of calcifications. A full listing of precision and recall for each category is reported in Table II. The poor result in precision and recall for anatomical location and shape is due to the variety of ways that location and shape descriptions are expressed in reports. While we achieved high recall for thrombosis, the classifier had low precision given that it could not differentiate between thrombosis events unrelated to ICA. We intend to improve the capabilities of our extraction tool using methods described in the discussion section.

C. Querying and Result Visualization

In exploring the ICA dataset, the neuroradiologist wished to perform two tasks: 1) pose queries that helped him explore the makeup of the patient population; and 2) identify subsets of patients that share common characteristics. In this section, we describe how the query interface and result visualization are used to answer two different queries. The first query involves filtering: *identify all patient cases that are female, older than 50 years old, presents with acute subarachnoid hemorrhage, and has undergone both surgery and endovascular embolization*. This query demonstrates how the system can be used to identify patients with a specific set of characteristics and compare them with respect to the entire population. The query is inputted by selecting the appropriate check boxes in the form-based interface. Selecting multiple checkboxes for a given attribute (e.g., choosing options for "51–70 years old" and "71 and older") returns the union of the patients matching those values. Results of the aforementioned query are depicted in Figure 2. Each new attribute (e.g., gender, age, clinical presentation) is treated as an additional constraint; the result of a query is the intersection of result sets across all specified attributes. The interface permits a user to sequentially input new constraints; the impact of the constraint on the overall query is emphasized by animating the transition between one result set and another. While a large proportion of the population becomes (un)highlighted given a constraint, it becomes visually apparent that the attribute has a large impact on the population. Selecting a person icon brings up a contextual menu that allows a user to view the raw patient data (e.g., reports, images) or use the selected patient as a query for case-based reasoning. Figure 2e illustrates how cases similar to the one selected (highlighted in purple) are ranked and presented to the user. A second type of query focuses on highlighting differences between cases stratified by related properties. For example, patient cases can be grouped based on outcome (e.g., Fisher grade, Glasgow coma score). In Figure 4, the population is divided into four groups based on the patients' Fisher grade. When additional constraints are specified using the query interface, matching cases within each group are highlighted in red, revealing spatial clusters that may elucidate common properties of each group. When the user inputs a query such as *identify all patients who are above 50 years old and clinically presented with mass effect*, the proportion of patients within each group who meet those criteria is highlighted. The

subgroups can be further partitioned (*e.g.*, patients that experience long-term neurological deficit or no deficit). Other queries that incorporate treatment information (*e.g.*, whether the number of embolizations performed) may be used to determine what combination of treatments result in improved outcomes.

V. Discussion and Future Work

We present a case-based retrieval framework that incorporates natural language processing to structure clinical text, a disease model to represent extracted concepts and capture contextual information, and a population-based visualization that facilitates exploratory analysis using the model to group patients and perform case-based reasoning. We show how the framework is used to explore a database of ICA patients. While we explored the use of cTAKES as part of this implementation, we are also in the process of adapting an NLP system that has been developed in-house for neuro-oncology to perform this task [14]. In addition to performing named entity recognition, this system attempts to characterize identified problems and findings along four axes: location, time, existence, and causality. These properties can also be used as additional attributes when posing a query. In particular, time is an important notion in medicine, and much of the medical data collected during clinical care are time-stamped. Our population-based visualization can be enhanced by taking time into account by adding functionality to animate how the patient population changes over time (*e.g.*, with respect to an outcome variable) and providing temporal filters (*e.g.*, highlight patients who had a failed surgery followed by embolization). While frameworks such as i2b2 provide similar functionality that searches across the entire clinical enterprise [26], our approach is designed to be lightweight, focusing on providing tools and an interactive visualization that are Java-based, quickly deployable, intuitive to use, and extensible.

One challenge in extracting and mapping concepts from clinical narratives is the vagueness and inconsistency of how some findings are reported in clinical documents. Adoption of semi-structured reporting would help improve the extraction tool's precision and recall. Recent guidelines on standardizing how variables and attributes are recorded are a step in the right direction; [24] provides a list of recommendations for the ICA domain.

Clinicians could potentially use this application to explore trends within their patient populations. In addition, this system is not tied to a particular domain. By providing the appropriate model, the system can be generalized to other diseases. Additional future work includes a refinement of the mapping tool by exploring more sophisticated methods for matching concepts beyond a simple rule-driven knowledge base. We are also examining whether the disease model, when incorporating relationships specified in published literature, can be used to generate explanations, allowing the visualization to supplement visual patterns with potential explanations (theories) of why the data appears that way. A limitation of the current work is the lack of a formal usability study. We are conducting an evaluation on the framework that is comprised of two phases: the first phase involves assessing the performance of the case-based reasoning system in identifying similar patient cases in the database using metrics such as sensitivity and specificity. Additional evaluations

of the extraction tool will be performed using a larger pool of manually annotated patient documents. The second phase is to perform a usability study that involves a convenience sample of clinical users drawn from residents and attending physicians in interventional neuroradiology and neurosurgery. A neuroradiologist who is not participating in the study was asked to generate a set of ten tasks that require participants to identify trends in the patient population. Participants are instructed to pose queries and interpret the results using the population-based visualization. The system will be evaluated based on the accuracy of the participants' answers for each task and the quantitative scores from a user interaction and satisfaction questionnaire. Results from these studies will inform future revisions to the framework.

Acknowledgments

The authors would like to acknowledge Drs. Nestor Gonzalez, Joshua Dusick, and Andres Vinuela for their time and efforts in identifying ICA variables and manually reviewing patient records. The authors would like to also thank Juan Anna Wu and Albert Chern for their assistance in implementing the visualization. This research was supported by NLM R01-LM009961 and NIBIB R01-EB000362.

References

1. Safran C, Bloomrosen M, Hammond W, et al. Toward a national framework for the secondary use of health data: An American Medical Informatics Association white paper. *Journal of the American Medical Informatics Association*. 2007; 14(1):1–9. [PubMed: 17077452]
2. Weiner MG, Embi PJ. Toward reuse of clinical data for research and quality improvement: The end of the beginning? *Annals of Internal Medicine*. 2009; 151(5):359–360. [PubMed: 19638404]
3. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*. 2010; 17(5):507–513. [PubMed: 20819853]
4. Rosenbloom ST, Denny JC, Xu H, et al. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association*. 2011; 18(2):181–186. [PubMed: 21233086]
5. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*. 2009; 42(5):760–772. [PubMed: 19683066]
6. Rector AL, Nowlan WA, Kay S, et al. A framework for modelling the electronic medical record. *Methods Inf Med*. 1993; 32(2):109–119. [PubMed: 8321129]
7. Coden A, Savova G, Sominsky I, et al. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. *Journal of Biomedical Informatics*. 2009; 42(5):937–949. [PubMed: 19135551]
8. Bui AA, Taira RK, El-Saden S, et al. Automated medical problem list generation: towards a patient timeline. *Stud Health Technol Inform*. 2004; 107(1):587–591. [PubMed: 15360880]
9. Min H, Manion FJ, Goralczyk E, et al. Integration of prostate cancer clinical data using an ontology. *Journal of Biomedical Informatics*. 2009; 42(6):1035–1045. [PubMed: 19497389]
10. Ming Z, Hong Z, Tjandra D, et al. DBMap: a space-conscious data visualization and knowledge discovery framework for biomedical data warehouse. *Information Technology in Biomedicine, IEEE Transactions on*. 2004; 8(3):343–353.
11. Bell DS, Sobolevsky S, Day FC, et al. The stadium diagram, a web-based tool for visualizing the expected outcomes of alternative clinical management strategies. *AMIA Annu Symp Proc*. 2005:36–40. [PubMed: 16778997]

12. Wang T, Plaisant C, Shneiderman B, et al. Temporal summaries: supporting temporal categorical searching, aggregation and comparison. *Visualization and Computer Graphics, IEEE Transactions on*. 2009; 15(6):1049–1056.
13. Klimov D, Shahar Y, Taieb-Maimon M. Intelligent visualization and exploration of time-oriented data of multiple patients. *Artificial Intelligence in Medicine*. 2010; 49(1):11–31. [PubMed: 20303245]
14. Hsu, W.; Arnold, CW.; Taira, RK. A neuro-oncology workstation for structuring, modeling, and visualizing patient records. *Proceedings of the 1st ACM International Health Informatics Symposium; Arlington, Virginia, USA*. 2010. p. 837-840.
15. Riesbeck, CK.; Schank, RC. *Inside case-based reasoning*. Lawrence: Erlbaum Associates, Inc; 1989.
16. Kolodner JL. An introduction to case-based reasoning. *Artificial Intelligence Review*. 1992; 6(1): 3–34.
17. Aamodt A, Plaza E. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*. 1994; 7(1):39–59.
18. Bichindaritz, I.; Marling, C. "Case-based reasoning in the health sciences: Foundations and research directions,". In: *Studies in Computational Intelligence*. Bichindaritz, I.; Vaidya, S.; Jain, A., et al., editors. *Computational Intelligence in Healthcare 4*. Berlin / Heidelberg: Springer; 2010. p. 127-157.
19. Melton GB, Parsons S, Morrison FP, et al. Inter-patient distance metrics using SNOMED CT defining relationships. *Journal of Biomedical Informatics*. 2006; 39(6):697–705. [PubMed: 16554186]
20. Wiebers DO, Whisnant JP, Huston J 3rd, et al. Unruptured intracranial aneurysms: natural history, clinical outcome, and risks of surgical and endovascular treatment. *Lancets*. 2003; 362(9378):103–110.
21. Brisman JL, Song JK, Newell DW. Cerebral aneurysms. *New England Journal of Medicine*. 2006; 355(9):928–939. [PubMed: 16943405]
22. Juvela S, Porras M, Poussa K. "Natural history of unruptured intracranial aneurysms: probability of and risk factors for aneurysm rupture. *J Neurosurg*. 2008; 108(5):1052–1060. [PubMed: 18447733]
23. Boeker M, Stenzhorn H, Kumpf K, et al. The @neurIST ontology of intracranial aneurysms: providing terminological services for an integrated IT infrastructure. *AMIA Annu Symp Proc*. 2007:56–60. [PubMed: 18693797]
24. Meyers PM, Schumacher HC, Higashida RT, et al. Reporting standards for endovascular repair of saccular intracranial cerebral aneurysms. *AJNR Am J Neuroradiol*. 2010; 31(1):E12–E24. [PubMed: 20075104]
25. Bui AAT, Weinger GS, Barretta SJ, et al. An XML gateway to patient data for medical research applications. *Annals of the New York Academy of Sciences*. 2002; 980(1):236–246. [PubMed: 12594093]
26. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*. 2010; 17(2):124–130. [PubMed: 20190053]

Figure 1.
Diagram summarizing system architecture and flow of information.

Figure 2.

Screen capture of the user interface. (a) The population pane shows the entire dataset of cases represented as an array of person icons. Icons are highlighted red based on whether they match a specific query. Icons can also be selected individually to retrieve details or perform case-based reasoning. (b) The query input pane is comprised of fields that allow users to specify values of attributes. (c) The query processing pane shows the constraints being used to filter the population. (d) The population pane can also be stratified into related groups. (e) The result pane shows information about a selected patient (highlighted in purple) or, as shown here, other similar cases found through a case-based reasoning query.

Figure 3.

Subset of the disease model showing the variable "hemorrhage", its subconcepts, and states.

Figure 4.

A population pane that is stratified by Fisher Grade. Highlighted patients are those that match the additional constraints of being over 50 years of age and presenting with mass effect.

TABLE I

Summary of Variables in ICA Model

Category	Variable	Data source
Demographics	Age Gender Clinical presentation	Admission note
Morphology	Number of aneurysms Dome size Neck size Anatomical location Shape indices	CT/MR angiography; radiology reports
Other features	Calcification Thrombosis Remodelling	
Embolization	Coil type Stent placement Technical complications	Surgical reports
Surgical	Exclusion reason Type of craniotomy Type of aneurysm clip Temporary artery clipping Bypass type Technical complications	
Outcome	Immediate outcome Long term outcome Neurological deficit	Discharge summary, physician consult

TABLE II

Summary of Extraction Results

Category	Example	Precision	Recall
Anatomical location	...left PCOM aneurysm...	92%	77%
Thrombosis	...progressive thrombosis within the aneurysm dome...	67%	100%
Aneurysm shape	...giant saccular aneurysm...	89%	78%
Aneurysm size	...aneurysm measures 9 mm × 10.3 mm × 11.9 mm	95%	97%
Calcification	Calcifications are seen in the cavernous ICA segments.	91%	100%