An Intelligent Distributed Medical Database via Data Superhighways

Thomas D. Coates, M.D.[1] *tom@hemonc.usc.edu*
Edward Chow, Ph.D.[2] *tlc@gator.jpl.nasa.gov*
Usama M. Fayyad, Ph.D.[2] *fayyad@aig.jpl.nasa.gov*
Randall W. Hill, Jr., Ph.D.[2] *hill@negev.jpl.nasa.gov*

[1]*Children's Center for Hematology-Oncology, Children Hospital Los Angeles, Los Angeles CA 90027.*

[2]*Jet Propulsion Laboratory, Pasadena, CA 91109.*

A. INTRODUCTION

> *A five year old child in rural Montana was diagnosed with leukemia. The nearest center for treatment of childhood leukemia is at the Children's Hospital of Los Angeles thousands of miles away. At great tax payer's expense, the child will be sent to Los Angeles for the complete one year treatment process. His parents are also faced with the dilemma of quitting their jobs to take care of her in Los Angeles.*

Healthcare is one of the most important social challenges facing the United States. Costs are skyrocketing and access to quality care is unevenly distributed throughout regions of the country. Current methods of dissemination of medical information result in substantial variation in treatment approach and costs among healthcare providers in different parts of the country. In order to reduce the cost of care critical statistical analysis of the medical and financial outcomes of various therapy options must be conducted on a large scale. Based on this information, expensive and ineffective treatments can be eliminated and feedback on effective treatment plans can be immediately communicated to all medical practitioners. We propose to make use of Artificial Intelligence techniques for data collection, management, and analysis coupled with high-speed networks (of the types that will become common as the Information Superhighway is constructed) to provide an innovative and badly needed solution to this important problem.

At the present time all electronically analyzable recording of medical practice is done by coding specialists who understand little about the human disease process. Thus, the data in the vast majority of hospital medical information systems contains coded representation of words written by physicians but none of the critical information which relates the coded words to each other, e.g., diagnosis to the treatment and outcome. Health care professionals understand little about the electronic recording or analysis of medical data. They record only minimal information in medical charts because of the laboriousness of manual entry. Hence, most of the critical relations between diagnosis and treatment are carried in their heads, cued only by the few words written in the record. Therefore, the critical medical information is unavailable for statistical scrutiny and certainly not available for innovative technological approaches to medical data analysis.

A.1 The Application Domain

The clinical database of the Children's Cancer Group (CCG) is an example of a rare set of medical information which preserves relations between diagnosis, treatment and outcome of treatment. The CCG is a consortium of approximately 28 universities and 128 hospitals involved in treatment of Childhood Cancer. Though rare, cancer is the most common disease which kills children in the United States. By instituting uniform treatment protocols, providing continual statistical analysis of pooled data, the CCG has brought the prognosis of childhood leukemia from zero in 1965, when the group started, to almost 80% cure rate in 1994. In spite of this success, all data entry and access is done by manual methods. Hundreds of data clerks, at great cost, collect this data and mail it in hard copy form to the associated institutions.

The first priority of the National Information Infrastructure Network (NIIN) project is to link all the hospitals in the U.S. before year 2000. NASA is playing the key role of linking rural area hospitals to NIIN with advanced satellite technology. A system based on NIIN which provides easy universal access to the outcomes of common treatment plans would normalize the vast differences in various regions of the country and reduce health care cost by emphasizing effective therapy and identifying ineffective therapy. The CCG has an excellent set of well described medical data in electronic form that could serve as a model to test artificial intelligence approaches to medical data management. Furthermore, if the problems of data entry could be solved using intelligent agents to assist in data collection/profile specification, this quality control and treatment protocol model could easily be scaled to develop nation-wide approaches to other disease processes.

B. OBJECTIVE

The long range goal of this project is to develop an AI-based workstation that allows health-care providers to efficiently and accurately document the practice of medicine in the course of their every-day work as well as provide automated tools for automatically analyzing and easy retrieval of the collected data. This involves recording diagnoses, treatments, outcomes and critical relationships between these data elements in a common national medical distributed data/knowledge base.

The specific goal of this project is to develop a pilot AI-based distributed database system on NIIN that provides efficient capture of data relevant to the study of pediatric cancer using the institutions of the CCG as an initial testbed. The system will also disseminate treatment protocol information and facilitate capture of critical data and relations between data elements in the course of practice of pediatric oncology. Furthermore, we will use the existing well-defined data set to explore innovative approaches to the analysis of image data collected by CCG institutions.

C. PROPOSED APPROACH

a. **High speed network:** JPL has been awarded the California Research and Education Network (CALREN) project grant by Pacific Bell telephone company. This will be a high speed fiber optics wide area network. The California Medical Network (CALM Net), which connects several major hospitals throughout California (LA General Hospital, San Francisco General Hospital, etc.) will be based on the CALREN. JPL is also serving as the west coast satellite station for the NASA ACTS advanced satellite network. The combination of CALREN, CALM Net, and NASA ACTS will be a perfect model of the nation's NIIN. We will use this platform to serve as our testbed between geographically distributed hospitals.

b. **Treatment and data acquisition workstation:** The treatment and data acquisition workstation has two purposes: (1) to assist clinicians in the administration of the Multi-disciplinary Action Plans (MAPs) during patient treatment, and (2) to assist in the acquisition of data related to the diagnosis, treatment, and outcomes of patient illness. It is envisioned that this can be accomplished by an Intelligent Clinician's Associate Program (ICAP) that would provide an easy-to-use work environment where the clinician can encode a diagnosis using standard symbols and terminology, obtain the relevant treatment plan (MAP) template, record each step of the treatment, receive automated assistance in administering the MAP through the use of situated plan attribution techniques, and record the outcomes of the treatment. ICAP would thereby serve both the intended purposes of the treatment and acquisition workstation. Furthermore, it is envisioned that ICAP would model each user and adapt itself accordingly, and it would provide tutoring on-demand to new users. We propose to transfer technology from JPL's research on intelligent tutoring and intelligent assistance in procedural domains (i.e., NASA's Deep Space Network) to implement ICAP.

c. **Automatic classification and Intelligent Data Retrieval:** medical databases, both in relational and image formats are growing at a rate that makes manual analysis or case-based query match usage infeasible. We propose the application of machine learning classification techniques to automate the analysis and to allow for intelligent and useful case retrieval for diagnosis purposes. We plan to target the area of tumor detection and recognition as a proof of concept problem.
Microscopic images of human tumors are the fundamental data upon which cancer treatment decisions depend. The interpretation of these images depends heavily upon the personal experience of the pathologist. Certain pathologists develop particular expertise in certain kinds of tumors and receive slides from all over the country by mail. These tumor images are difficult to describe, let alone write dedicated programs for detecting them. Therefore, they represent an appropriate domain for a learn-from-example type approach. Training data for the learning algorithms will be provided by an experienced pathologist looking through a microscope.
CHLA is a national referral center for pathologic study of neuroblastoma specimens. This is a devastating cancer which spreads early to bones of infants and results in a prolonged, painful death. The tumor prognosis depends upon the visual appearance of the microscopic tissue section.

d. **Machine Learning for Non-intrusive Diagnosis of Brain Tumors:** Malignant brain tumors are one of the most common and most devastating childhood cancers. Frequently, it is not possible obtain pathologic samples of the tumor without causing increased neurological deficit to the child. High resolution x-rays, CAT scans and MRI, or PET scans (positron emission tomograms) of these tumors are always available before biopsy.
We propose to gather digital images of sections of tumors and express the images in n-dimensional feature space and determine if prognoses as determined by the existing survival data, correlates with any feature or combination of features. We will further classify tumors by clusters of features in n-D space and determine if optical features of the image, not discernible by human consideration of only 2 or 3 dimensions (features), are important determinants of prognosis. These studies will set the stage for a national repository of tissue images as well as establish telepathology technology for the CCG.
By coupling features measured from high resolution x-ray 3-d images with the prognosis as determined in the CCG survival database, we hope to automatically extract rules to predict outcome or pathologic tissue type as determined at biopsy or autopsy. If this is successful, new mathematical image features may allow accurate determination of tumor type without the need for invasive biopsy. This would directly minimize the need for invasive procedures, reducing a lot of pain, risk, and huge costs. This proposed solution is analogous to the one developed in SKICAT for recognizing sky objects that are beyond human recognition capabilities. In the latter case, we used high resolution CCD images to obtain training data. In this case the biopsy, autopsy, or later proper determination of the correct outcome provides the training data.

D. BENEFITS OF PROPOSED WORK

1. **Short Term Benefit:** The efficiency of data capture and analysis for the CCG will be greatly increased furthering the fight against the number one disease killer of children. New analytic approaches to image data will result by comparing features of microscope slides and radiographs to data vectors from a well-defined database of outcomes. *For example, the child from Montana will be able to get treatment from a local hospital which acquires the ideal treatment process from our database through NASA NIIN. His parents can take care of her at home.*

2. **Long Term Benefit:** This system will markedly improve the delivery and reduce cost of healthcare by providing easy access to the most common treatments used for the diseases as well as the outcomes of treatments to healthcare providers in all regions of the nation. Since

recording of diagnosis, treatment and outcome will be accomplished with a highly intelligent workstation, efficiency of documentation will be greatly increased at the same time that important information regarding therapy is provided. This will free health care providers to see more patients instead of writing notes. Since this information is national in scope, the system will result in a decrease in the variances in treatment seen between regions of the country. Overall medical cost will drop as clinicians select the most efficacious treatments[1]. The automated data analysis and diagnosis capabilities would provide the means to make proper use of the data collected for decision support, aid in case retrieval for diagnosis or treatment justification, and allow for "query by content" type operations on the large medical image databases. The latter will be extremely useful for research work: e.g., a new pattern is detected, a researcher wants hundreds of thousands of images scanned for the new pattern and cataloged appropriately for further statistical analysis/examination.

The work proposed represents the development of a new generation of medical data analysis tools. It puts together AI and high-speed networks to target the development of a system that is well-suited for demonstrating the power and utility of the planned national information superhighway.

[1]This has been shown and published: just printing the cost of a drug next to the selection on the computer screen reduces medical cost by 30%