# Automated Linkage of Free-text Descriptions of Patients with a Practice Guideline

Leslie A. Lenert, M.D., M.S. and Maria Tovar, M.S., Division of Clinical Pharmacology,
Stanford University School of Medicine, Stanford, California

## ABSTRACT

*The process of applying a practice guideline to a patient requires a great deal of clinical data. AAPT (Appropriateness-Assessment Processing from Text) is an experimental computer program that can assess the appropriateness of coronary-artery bypass grafting surgery (CABG) in patients with coronary-artery disease (CAD) and chronic stable angina from the admission summaries of those patients. The AAPT architecture combines natural-language processing (NLP) and probabilistic inference. The NLP module identifies single clinical concepts of interest in the free-text document. The probabilistic inference module, a Bayesian belief network, estimates values for variables not specifically mentioned. AAPT produces a patient's summary of CAD that is similar to a manually generated clinical summary. Work is ongoing to improve AAPT and evaluate it as a tool to assist in the dissemination of guidelines and as a tool to encourage adherence to practice guidelines.*

## INTRODUCTION

Practice guidelines are an evolving tool of increasing importance in limiting unnecessary variations in clinical practice and improving the overall appropriateness of care. Automated application of practice guidelines has been shown to be quite useful for decision support [10] and may be useful for screening the appropriateness of medical services in managed-care environments. However, applying practice guidelines is often difficult. Before a guideline can be applied a great deal must be known about the clinical context of care. An important source of detailed clinical data that is beyond the reach of most medical systems are free-text admission and discharge summaries.

To explore methods for using natural language processing to acquire data for application of a practice guideline, we designed and implemented AAPT (Appropriateness-Assessment Processing from Text). AAPT operates in the domain of assessment of the appropriateness of coronary-artery bypass grafting surgery (CABG) as a treatment for chronic stable angina. We have chosen this domain because v worked out constructs of appropriateness exist have been implemented in practice guidelines AAPT takes natural language descriptions of patie severity of coronary-artery disease (CAD) from admission's history and physical (H&P) of patie that have undergone or are about to undergo CA surgery, and processes those descriptions to develc clinical summary of the variables relevant to application of a practice guideline for CABG surgei

## BACKGROUND

Many researchers have developed systems ı analyze and extract information automatically fı narrative medical documents. The applications aı include document retrieval, automated summarizaı of patient records, and clinical database construcı [2, 3, 4, 12]. Application of NLP for qua assurance has been advocated by Sager and Gabı [3, 12]. The feasibility of processing narrative med records was demonstrated as early as 1981 Hisrchman who described a system that analyzed evaluated a patient's discharge summary [6]. Lyn Sager, and Tick have recently described using NLI derive summaries stored in a relational database Subsequent queries to the database identify cases wł appropriate procedures were potentially not follow Gabrieli's research has focused on systems that exti and summarize medical facts from medical charts His work has included a algorithm for identify charts that show problems with quality of care.

## DESIGN CONSIDERATIONS

The purpose of AAPT was twofold: to automa method for applying a practice guideline for CA surgery, and to explore probabilistic methods to iı missing data. In particular, it was not enough know that a patient has CAD (that may be obvi from a coded admission or discharge diagnosis). AA had to characterize the severity of the patients v regard to what is known about the pathophysiolı and treatment of CAD.

0195-4210/92/$5.00 © 1994 AMIA, Inc.

Missing information is a critical problem in this characterization. While a free-text admission summary or discharge summary usually contains a summary of the patients pathophysiology at admission, there is no guarantee that a document will contain the particular data required to classify a patient with regard to a particular guideline. The problem of "missing" information is ubiquitous in free-text processing. Several approaches to the problem have been attempted within the medical domain but in other contexts.

While a variety of approaches have been applied to solve the problem of missing information, none is completely satisfactory. For example, frames have been used to resolve incomplete information. This approach is applied in SPRUS, an NLP system for radiology reports developed by Haug and colleagues at Latter-day Saints Hospital [4]. Within SPRUS, frames and heuristic rules were combined to infer missing single elements of information. Suppose the system found the word "infiltrate" in a patient's chart. This word could mean "interstitial infiltrate," "diffuse alveolar infiltrate," or "localized alveolar infiltrate." SPRUS can disambiguate among these three terms by using the diagnosis mentioned at the end of the report to identify the frame that should be instantiated. This frame specifies the appropriate missing modifier for infiltrate. For example the diagnosis "pneumocystis pneumonia" would allow SPRUS to infer that the frame for pneumocystis pneumonia should be instantiated and the correct datum in the frame was "diffuse interstitial (infiltrate)."

Other researchers have used semantic networks to generalize information and infer facts not specifically stated in a document. A *semantic network* is a graph structure, where nodes represent concepts of interests, or objects, and an edge (a line) between two nodes represents a relationship between concepts. For example, SAPHIRE [5], a system for conceptually-based literature searching, takes advantage of the hierarchical relations in its vocabulary: the user can add the children. or parents of a node in the semantic network to broaden or narrow a search. Wagner has also applied this approach to predicting missing information using a UMLS-derived network [13].

A third approach has been to specify the distinction between findings not mentioned and findings mentioned but specified not to be present. Identification of this distinction can be combined with the application of an algorithm that can infer the meaning from absence of mention of specific terms. This approach has been applied by Zingmond and Lenert [15] to the interpretation of chest radiographs. Statistically derived decision rules treat the absence of mentioning a finding as positive information in classification of radiographs with regard to their need for oncologic follow up.

One critical distinction is the difference between information that is missing, and information that is implicit in the document but not explicitly stated. With implicit information, the document has all the data necessary to infer a particular concept, but the concept is not explicitly stated. The problem of inferring implicit information from medical narrative can be viewed as a "diagnostic" problem or as a classification problem. For example, suppose a physician wants to determine a patient's severity of angina according to published guidelines. The medical narrative may contain facts about this patient's severity of angina, but it may not use the guideline terminology to describe what the severity of angina is. The physician may use these observed findings to subjectively determine the patient's angina classification. The process is similar to that of applying a set of findings to formulate a differential diagnosis. Since findings about a patient are imprecisely defined in a medical chart, we need some method to manage the uncertainty inherent in the classification task.

## IMPLEMENTATION

### The AAPT Clinical Model

The clinical description of patients generated by AAPT is adapted from the variables used to describe patients in the Coronary Artery Study (CASS) [11]. The CASS study, a randomized trial of CABG surgery for chronic angina, remains the foundation for assessment of the appropriateness of CABG surgery. The CASS model includes age, gender, the number of obstructed coronary arteries, concurrent diseases which are known to increase risk of surgery, the clinical severity of congestive heart failure, the degree of left ventricular wall motion abnormality, and a measure indicating the clinical severity of angina based upon the Canadian Heart Association (CHA) classification score, a widely used clinical metric. With the exception of age, each variable is categorical describing the condition at a number of intensities. For example, the clinical level for congestive heart failure is based upon the history of congestive heart failure, the presence of a third heart sound (S3) on physical examination, and use of diuretics, digoxin and angiotensin-converting enzyme inhibitors for treatment congestive heart failure.

## The AAPT Inference Engine

AAPT uses both categorical and probabilistic reasoning to create a summary of a patient's CAD based on relevant indicators from the CASS model. Some of the variables in the clinical model are explicitly mentioned in the H&P, such as the patient's age and gender, which are almost universally documented. The number of obstructed coronary-artery segments, the number of concurrent diseases, and the congestive-heart-failure score are estimated using categorical reasoning. For example, to calculate the number of obstructed segments, the NLP component of AAPT extracts from the H&P which coronary-artery segments are obstructed. AAPT applies this information to the set of rules which encode knowledge about coronary anatomy to count the number of obstructed vessels.

AAPT uses belief networks where probabilistic reasoning is required. There are two applications for belief networks in AAPT: to refine variable estimates, and to infer complex medical concepts. For example, AAPT's model requires a left-ventricular wall-motion score which must be derived from descriptions of the wall motion of the heart during cardiac catheterization. The formula for calculating this score uses descriptions of the quality of motion for the diseased wall segments from the admission summary. Some physicians describe wall motion from a particular projection of the ventriculogram, without describing the segments involved. Given a particular projection view, AAPT can infer which segments are visible, and can use this information to calculate an initial wall-motion score. There is a direct correspondence between the description of a segment-motion abnormality and a numerical score. The overall score is obtained by adding the contribution of the segments in at least one view. AAPT also uses a belief network to refine the left-ventricular wall-motion score based on the previously calculated score and other data mentioned in the summary such as the ejection fraction, whether there is a heart-valve insufficiency, and the cardiac output.

AAPT also uses belief networks for inference of values that cannot be encoded easily in a few rules, that require diagnostic evaluation, or that are based on relationships that are inherently imprecise. One example is assignment of the CHA angina classification score. The CHA algorithm assigns one of four severity grades to a patients CAD, based on the presence or absence of symptoms and the degree to which symptoms limit activity of the patient. Angina class is determined whether there is pain at rest (Anrest) or pain with exertion at different activity levels (PActivity). All pain is not presumed to be chest pain--only pain that is clinically consistent with angina pectoris. Various clinical descriptors can influence characterization of the pain. Quality of the pain (PainQual) helps determine the whether the symptoms are relevant to CHA class. Quality is assigned one of three levels: classic, typical, or atypical. Location, the type of sensation, radiation and the presence of autonomic symptoms determine the value of PainQual. (Autonomic) node refers to the presence or absence of autonomic symptoms with chest pain, which increase the likelihood that the described pain represents angina pectoris. Each node in the net work takes values determined by the concepts observed in text. If the class node is not observed, its value is calculated by the algorithms implemented in the network.

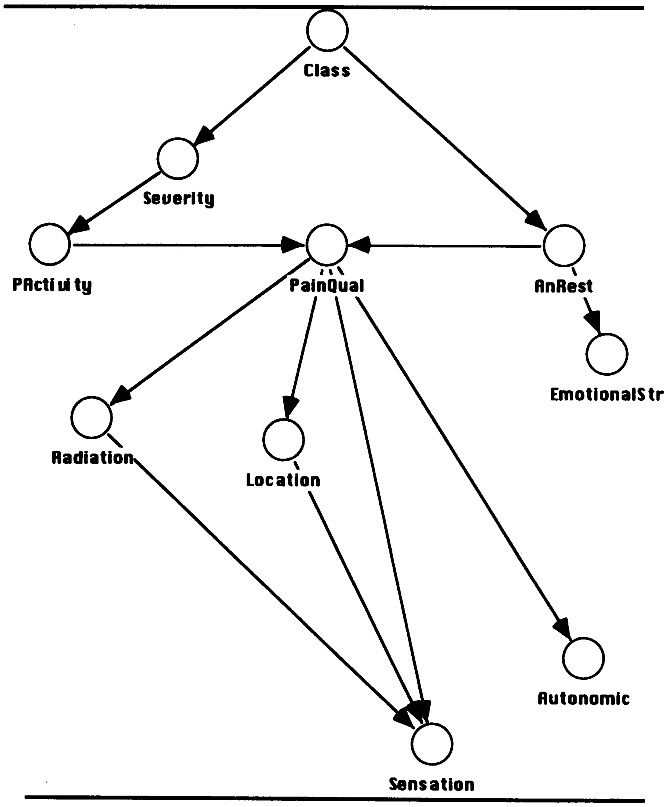

Figure 1. Belief network for Canadian Heart Association Angina Classification (see text for explanation).

A node in the belief network is instantiated when a canonical concept is identified in text that is equivalent to the node is identified in text and has a context that indicates the presence of the concept (as opposed to the negation of the concept). Nodes can in this way be instantiated by a variety of equivalent phrases in the text.

### Natural-Language Processing

NLP processing in the AAPT system is handled by a modified version of the Canonical Phrase Identification System (CAPIS) program [8]. CAPIS applies a concept-based free-text processing algorithm to extract matched findings automatically from a dictated H&P. This architecture completely separates the knowledge-acquisition component from the parser, and therefore, CAPIS can be rapidly custom-tailored to analyze medical narrative for a particular domain. The knowledge base consists of two separate lexicons: a findings list and a thesaurus. The findings list contains the canonical phrases (medical concepts) that the user wants to extract from the H&P. The thesaurus contains the synonyms for each of the word in the findings list. CAPIS has a grammar implemented in a finite state machine that allows CAPIS to break sentences into single subject phrases, to track propagation of negations in a sentence, and to complete phrases with missing fragments.

## PROGRAM STATUS

AAPT is implemented in C on a SUN IPC workstation. A single application processes text and performs probabilistic inference. We use Ergo© (Noetic Software, 1989) for probabilistic inference, which is a general-purpose probabilistic reasoning system employing the Lauritzen–Spiegelhalter algorithm [7] that can be linked with other C programs. AAPT is capable of processing a two page admission summary in about 30 seconds.

### Validation studies

We performed a preliminary evaluation of the quality of the abstraction and interpretation of the CASS model variables produced by AAPT with those produced by manual abstraction. Rates of agreement for the components of the clinical model between manual abstraction and the software were compared.

We selected at random 40 patients who had undergone CABG surgery from a database of patients who had undergone CABG surgery in 1985 and 1986 at Stanford Medical Center. We obtained photocopies of these patients admission summaries and converted the summaries to electronic form by optical character recognition. Errors in recognition were corrected by manual review. A medical student, working on a separate project who was not familiar with the workings of the program selected 26 cases that represented a wide range of appropriateness as assessed by the predictions of a survival model under his study. He abstracted the 26 remaining admission summaries to obtain the values for the 5 variables of the CASS model. The abstractions were reviewed by a board-certified internist. We used 8 of the 26 patients' H&P as a training set for AAPT. These patients' admission summaries helped us refine the Bayesian belief networks and the rule base for inference and interpretation. We then evaluated the accuracy of the clinical descriptions generated by AAPT for the remaining 18 cases. To evaluate the accuracy of the description, for each element in the clinical description we compared the category assigned by AAPT with those assigned by manual abstraction using an Cohen's unweighted Kappa statistic .

Overall, the Kappa statistic shows there was fair agreement with regard to the description of the clinical variables except for the CHA score. This was the most complex inference. Kappa statistics for all variables are shown in Table 1.

| Variable | Kappa |
|---|---|
| CHA score | 0.34 |
| LV Score | 0.57 |
| CHF score | 0.625 |
| Vessels | 0.46 |
| Diseases | 0.51 |

Table 1 Unweighted Kappa statistic estimate for variables that comprise the clinical description of the patient in the CASS framework.

## DISCUSSION

The AAPT program combines abstraction and interpretation of free-text data to allow the definition of a clinical scenario. This is the foundation for application of a clinical guideline or other methods to assess the appropriateness of care. Guidelines are often expressed by saying treatment X is indicated for disease Y when conditions A, B, and C exist. The attributes used in the guideline may or may not be explicitly stated in an admission summary. Often, however, even when not stated, it may be relatively easy for a physician to infer what they ought to be. We have attempted to simulate this inference in the AAPT program.

AAPT's processing is a form of automated coding of free-text that is custom tailored to the knowledge representation of CAD in the CASS study. Because the obstacles in developing universally applicable systems for coding medical data are significant, free-text processing systems for high-value areas, such as CABG surgery, may be an important interim method to bring practice guidelines to the bedside. AAPT's performance, while not yet adequate for individual decision support, may be acceptable for screening the appropriateness of care. Given the high incidence of potentially inappropriate application of CABG surgery [14], AAPT may be useful.

Potential areas for improvement of performance include refinement of the natural language processing system (CAPIS) to improve its sensitivity and precision. More extensive modeling of the relationships between findings identified by CAPIS and clinical states that are part of the guideline may also be fruitful. The belief networks that capture those relationships could be refined and expanded.

Further research is also needed to help describe the relationship between what physician see and what they document. Clearly, this is not a random relationship, but is based on their clinical training , experience and their assessment of the relevance of the findings to prognosis. Lapses in documentation also occur from time to time. The same methods used in analysis could be applied feedback information to physicians on the quality of their documentation and improve documentation of key attributes.

## ACKNOWLEDGMENTS

We are grateful to Lyn Dupré for reviewing earlier versions of this manuscript and acknowledge the contributions of Eric Stark, Blackford Middleton, and Naomi Sager to this project. This work supported by a grants from the Pharmaceutical Manufacturers Association and the Institute for Clinical and Biological Investigation. We also acknowledge the support of the CAMIS facility and the evironment of the Section on Medical Informatics at Stanford University.

## REFERENCES

1. ACC/AHA guidelines and indications for coronary-artery bypass graft surgery. *Circulation.* **83**(3):1125–73, 1991.
2. K. Canfield, B. Bray, S. Huff, and H. Warner. Database capture of natural language echocardiographic reports: an unified medical language approach. *Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care*, Washington, D.C., IEEE Press, 559-63, 1988.
3. E. Gabrieli, G. Pollock, E. Beltrami, and R. Soldano. Automated screening for quality of care in the hospital—a feasibility study. *J Clin Comput.* **17**: 5-6, 1989.
4. P. Haug, D. Ranum, and P. Frederick. Computerized extraction of coded findings from free-text radiologic reports. *Radiology*, Feb. **174**(2): 543-8, 1990.
5. W. Hersh and R. Greenes. Information retrieval in medicine: state of the art. *M.D. Comp.* **7**(5):302-11, 1990.
6. L. Hirschman, G. Story, E. Marsh, M. Lyman, and N. Sager. An experiment in automated health care evaluation from narrative. *Comput Biomed Res.* **14**(5): 447-63, 1981.
7. S. Lauritzen and D. Spiegelharter. Local Computations with Probabilities in Graphical Structures and their Application to Expert Systems. *Royal Statistical Society B.* **50**(2): 157-224, 1988.
8. R. Lin, L. Lenert, B. Middleton, and S. Schiffman: A free-text processing system to capture physical examination findings: CAPIS. In: Clayton P.D., ed. *Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care*, McGraw-Hill, 843-47, 1991.
9. M. Lyman, N. Sager, L. Tick, and J. Scherrer. The application of natural language processing to healthcare quality assessment. *Med Decis Making.* **11**(suppl): S65-S68, 1991.
10. C. McDonald, S. Hui , D. Smith, et al. Reminders to physicians form an introspective computerized medical record. A two-year randomized trial. *Ann Int Med* **100**(1):130-8, 1984.
11. Principal Investigators of CASS and their Associates. The National Heart, Lung, and Blood Institute Coronary Artery Surgery Study (CASS). *Circulation.* **63**(SI): I1-I67, 1981.
12. N. Sager, C. Friedman, and M. Lyman. "Medical language processing: computer management of narrative data." 1987, Addison Wesley, Menlo Park.
13. M. Wagner. An automatic indexing method for medical documents. *Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care*, Washington, D.C., 1011–16, 1991.
14. C. Winslow, J. Kosecoff, M. Chassin, D. Kanouse, and R. Brook. The appropriateness of performing coronary artery bypass surgery. *JAMA* **260**(4):505–9,1988.
15. D. Zingmond and L. Lenert. Monitoring free-text data using medical language processing. *Computers in Biomedical Research (in press).*