Application and Evaluation of Unified Medical Language System Resources

to Facilitate Patient Information Acquisition

through Enhanced Vocabulary Coverage

by

Eric M. Mills, D.V.M.

Dissertation submitted to the Faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY
in
Veterinary Medical Sciences

Jeff R. Wilcke, D.V.M., M.S., Diplomate, A.C.V.C.P.
(Chairman, Major Professor)

Holly S. Bender, D.V.M., Ph.D. Diplomate, A.C.V.P.

Deborah Hix, Ph.D.

William R. Chickering, D.V.M., Ph.D.

Harold A. Kurstedt, Ph.D.

Ludeman A. Eng, M.A., Ph.D.

John C. Lee, Ph.D.

Allen W. Hahn, D.V.M., Ph.D., Dipl. ACVIM
External Examiner

April 1998

Blacksburg, Virginia

Keywords: Patient Information Acquisition Model, PIAM, controlled medical vocabulary, lexical matching, lexical strength, explicit synonym disagreement, Systematized Nomenclature of Medicine and Veterinary Medicine, SNOMED

Application and Evaluation of Unified Medical Language System Resources to Facilitate Patient Information Acquisition through Enhanced Vocabulary Coverage

Eric M. Mills, D.V.M.

Committee Chairman: Jeff R. Wilcke
Veterinary Medical Science

(Abstract)

Two broad themes of this research are, 1) to develop a generalized framework for studying the process of patient information acquisition and 2) to develop and evaluate automated techniques for identifying domain-specific vocabulary terms contained in, or missing from, a standardized controlled medical vocabulary with emphasis on those terms necessary for representing the canine physical examination.

A generalized framework for studying the process of patient information acquisition is addressed by the Patient Information Acquisition Model (PIAM). PIAM illustrates the decision-to-perception chain which links a clinician's decision to collect information, either personally or through another, with the perception of the resulting information. PIAM serves as a framework for a systematic approach to identifying causes of missing or inaccurate information.

The vocabulary studies in this research were conducted using free-text with two objectives in mind, 1) develop and evaluate automated techniques for identifying canine physical examination terms contained in The Systematized Nomenclature of Medicine and Veterinary Medicine (SNOMED), version 3.3 and 2) develop and evaluate automated techniques for identifying canine

physical examination terms not documented in the 1997 release of the Unified Medical Language System (UMLS).

Two lexical matching techniques for identifying SNOMED concepts contained in free-text were evaluated, 1) lexical matching using SNOMED version 3.3 terms alone and 2) Metathesaurus-enhanced lexical matching. Metathesaurus-enhanced lexical matching utilized non-SNOMED terms from the source vocabularies of the Metathesaurus of the Unified Medical Language System to identify SNOMED concepts in free-text using links among synonymous terms contained in the Metathesaurus.

Explicit synonym disagreement between the Metathesaurus and its source vocabularies was identified during the Metathesaurus-enhanced lexical matching studies. Explicit synonym disagreement occurs, 1) when terms within a single concept group in a source vocabulary are mapped to multiple Metathesaurus concepts, and 2) when terms from multiple concept groups in a source vocabulary are mapped to a single Metathesaurus concept. Five causes of explicit synonym disagreement between a source vocabulary and the Metathesaurus were identified in this research, 1) errors within a source vocabulary, 2) errors within the Metathesaurus, 3) errors in mapping between the Metathesaurus and a source vocabulary, 4) systematic differences in vocabulary management between the Metathesaurus and a source vocabulary, and 5) differences regarding synonymy among domain experts, based on perspective or context. Three approaches to reconciling differences among domain experts are proposed. First, document which terms are involved. Second, provide a mechanism for selecting either vocabulary-based or Metathesaurus-based synonymy. Third, assign a “basis of synonymy” attribute to each set of synonymous terms in order to identify the perspective or context of synonymy explicitly.

The second objective, identifying canine physical examination terms not documented in the 1997 release of the UMLS was accomplished using lexical matching, domain-specific free-text, the Metathesaurus and the SPECIALIST Lexicon. Terms contained in the Metathesaurus and Specialist Lexicon were removed from free-text and the remaining character strings were presented to domain experts along with the original sections of text for manual review.

This dissertation is dedicated to:

Dr. Richard B. "Dick" Talbot. Dick was my first major professor who began the program in veterinary informatics and died tragically in a plane crash in 1994. What can you say about someone who never wanted to grow up? His boyish enthusiasm and ability to see far into the future was excelled by no one. He saw technology (and people) in the light of what could be, not what was. He always had a smile when the work was good, but had the "special smile" when the work was great. Dick had a laugh that I can still hear, and hope I always do. His laugh reminds me this work is supposed to be fun.

My family. My wife, Georgeann, is every bit the Proverbs 31 woman. As for my daughters, Becca and Bonnie, they have taught me about grace, forgiveness and unconditional love. It is only in the context of my family, that my work has meaning.

Holly Bender. A colleague such as Holly is a rare gift. Her wisdom, insight, tenacity, friendship and patience are indispensable. She has "glued my wing back on" more times than I can count.

Drs. Patrick A. Brett and Layne B. Brett. Pat and Layne have been great friends and have managed my animal hospitals with diligence and integrity in my absence. We have grown both personally and professionally together. I thank God for bringing us together.

My parents. Thank you for always fostering my curiosity and creativity. My first research grants came from you.

Acknowledgements

Many people deserve my thanks for their help in this endeavor. I'll begin with my major professor, Dr. Jeff Wilcke. A graduate degree is the result of a partnership between a student and a major professor. Neither can do their job without the other. I am continually amazed by the quickness of Jeff's mind and his editing ability. I have learned a great deal from him and will forever be grateful. I would also like to thank Duane Steward, a fellow informatics graduate student at MIT. No one has mastered the art of sleep deprivation and interrogation like Duane. He has continually challenged me until the wee hours of the morning with new ideas. I am particularly thankful to Dr. Steve Holladay for his unwavering support from the very beginning. He and Susan opened their home to me when I began commuting to Virginia Tech. He has always been there with a word of encouragement when I needed it most. I especially appreciate Siegfried "Ziggy" Hill and Betsy Blythe for their friendship and countless hours of stimulating conversation. I would also like to thank Linda Price from Research and Graduate Studies for answering all of my questions with pleasantness and patience. (Especially when the same questions were asked multiple times.) Finally, I would like to thank my good friends, Jimmy and Terrie Coates. Their integrity, faith and encouragement are a constant inspiration to Georgeann and me. Terrie was the one who told me, "If anyone should get out of private practice, it is you." After my confused look, she clarified herself. "So many people are stuck and can't follow their dreams, but you can. You owe it to yourself and to everyone else to do this."

LIST OF ABBREVIATIONS

CBC	Complete Blood Count
CUI	Unique Concept Identifier
HIS	Hospital Information System
MSM	Management System Model
NEC	Not Elsewhere Classified
NLM	National Library of Medicine
NOS	Not Otherwise Specified
PIAM	Patient Information Acquisition Model
SNOMED	Systematized Nomenclature of Medicine and Veterinary Medicine
UMLS	Unified Medical Language System

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Quality patient care requires clear and unambiguous communication of patient information among healthcare providers. The primary mechanism of this communication is the medical record. Therefore, the medical record should contain a complete and accurate description of what was observed, thought, and done on behalf of the patient. Studying and improving the process of perception, interpretation, documentation and presentation of patient information is essential to improving patient care.

A medical record is by nature incomplete and inaccurate. For that matter, any description is incomplete and inaccurate, as it is only a representation of reality. Thus, the aim of the medical record is to document observations, thoughts and actions in a manner that approximates the patient's reality. Incompleteness and inaccuracy must be reduced to the point that patient care and outcome are not compromised.

Completeness within the medical record is supported by requiring the entry of specific patient data. Unfortunately, requiring specific data for all patients is viewed by some as undesirable, since what is complete for one patient may be unnecessary for another. On the other hand, what was thought unnecessary for a patient initially, may be discovered later as essential.

Incompleteness and inaccuracy can enter the medical record at many points. The first and most obvious entry point for incompleteness is the decision on whether or not to make an observation. More mistakes are made when one “fails to look,” than when one “looks and fails to see.” Once an observation is made, the next entry point for incompleteness and inaccuracy is the perception and conversion of the observation into written text. Finally, when the written text is read, additional incompleteness and inaccuracy can result from the reader’s experience and frame of reference. In summary, not only must observations be seen and comprehended, their results must be recorded and presented in a manner that accurately communicates the original event to the reader.

Consistency within the medical record is supported through the use of structured data entry and a standardized controlled medical vocabulary. Structured data entry provides a fixed set of choices for recording observations. As an example, a veterinarian might want to describe the finding, “slight swelling around the stifle joint.” A data collection instrument could have an entry for **stifle joint**, along with the observation of **swelling** and the options of **slight**, **moderate**, and **severe**. This approach would certainly simplify data entry, however, not all swollen stifle joints can be grouped into one and only one of three unique categories. Issues such as symmetry of the swelling or discoloration also need to be recorded. Thus, consistency achieved through a limited set of choices can be artificial, accomplished at the expense of accuracy or expressiveness, and can even compromise the patient’s healthcare.

The immediate purpose and future use of this research is to develop a computerized structured data collection instrument for recording the canine physical examination. This instrument will support clear and accurate communication as practiced in the medical record by enforcing completeness and consistency, minimizing inaccuracies, while also providing expressiveness for the rich and subtle variations of medical data.

Before the issues of completeness, consistency and accuracy can be resolved, the issue of terminology must be addressed. Accurate communication begins with agreement on the topic being communicated. This agreement is best achieved through the use of a standardized nomenclature that provides names and their synonyms for the topics of discourse. This need for standardized names is reflected in the Chinese saying: the beginning of wisdom is getting things by their right name.[1]

Unfortunately, establishing a standardized nomenclature is not a trivial task. One must first decide which concepts (topics) will be included and what terms (names) will be used to represent those concepts. A term differs from a concept in that a term is a language-based label for a thought-based concept.[2] A further complication for veterinary medicine is that concepts and their related terms must reflect differences among species. As an example, the terms “knee” and “stifle” are considered synonymous by some, with “knee” preferred in human medicine and “stifle” preferred in veterinary medicine. An additional complication is that a layperson may use the term “knee” as a synonym for “carpus” in the horse. Thus, an effective standardized nomenclature must be dynamic and account for variation among species, variation among users, be adaptable to future changes in veterinary medicine, while also maintaining the ability to document past practices.

The Systematized Nomenclature of Medicine and Veterinary Medicine (SNOMED) has been chosen by the American Veterinary Medical Association as the standardized controlled vocabulary for veterinary medicine. SNOMED contains approximately 100,000 unique concepts represented by over 144,000 terms used in human and veterinary medicine. Each concept in SNOMED is assigned an alphanumeric code referred to as a termcode. Multiple terms representing a given concept are assigned the same termcode. As an example, the terms “Lacrimal gland function, NOS” and “Lacrimation” are considered synonyms for the same concept and therefore share the same termcode, F-F2900.

There are many standardized nomenclatures representing various domains. In an effort to facilitate information integration and retrieval from among multiple biomedical information sources using various vocabularies, the National Library of Medicine (NLM) established the Unified Medical Language System (UMLS).[3] Among the resources that comprise the UMLS are the Metathesaurus and the SPECIALIST Lexicon. The Metathesaurus is a compiled list of terms from over 40 autonomously controlled medical vocabularies known as, source vocabularies. The Metathesaurus links synonymous terms from among these source vocabularies. The 1997 version of the Metathesaurus contains approximately 330,000 unique concepts represented by approximately 740,000 terms. The SPECIALIST Lexicon contains lexical variations of medical terms in addition to non-medical terms. Between them, the Metathesaurus and SPECIALIST Lexicon contain all of the terms that are documented in the UMLS.

Themes of the research

The two broad themes of this research were, 1) to develop a generalized framework for studying the process of patient information acquisition in order to reduce incompleteness and inaccuracy in the medical record and 2) to develop and evaluate automated techniques for identifying domain-specific vocabulary terms contained in, or missing from a standardized controlled medical vocabulary.

A generalized framework for studying the process of patient information acquisition is addressed in Chapter 2 entitled, "The Patient Information Acquisition Model: A Framework For Identifying Causes of Missing or Inaccurate Information." The Patient Information Acquisition Model (PIAM) illustrates the decision-to-perception chain which links a clinician's decision to collect information, either personally or through another, with the perception of the resulting information. PIAM is useful as a diagnostic tool for improving the quality of patient information by serving as a framework for a systematic approach to identifying causes of missing or inaccurate information. Each component and activity of PIAM is associated with characteristic patterns of information corruption, and breakdowns in communication. In addition, PIAM also provides the context for the vocabulary studies described in Chapters 3,4 and 5, as each component and activity of PIAM relies on a standardized controlled medical vocabulary for optimal performance.

The vocabulary studies focus on identifying terms necessary for recording the canine physical examination and are conducted with two objectives in mind, 1) develop and evaluate automated techniques for identifying canine physical examination terms contained in SNOMED version 3.3, and 2) develop and evaluate automated techniques for identifying canine physical examination terms not documented in the 1997 release of the UMLS.

The first objective, development and evaluation of techniques for identifying canine physical examination terms contained in SNOMED is addressed in Chapter 3, "Metathesaurus-enhanced Lexical Matching in Domain-Specific Free-text." In this chapter, two complementary lexical matching techniques are evaluated, 1) lexical matching using SNOMED alone and 2) Metathesaurus-enhanced lexical matching. Both techniques are evaluated using a textbook on the physical examination in companion animals.[4]

The first technique, lexical matching using SNOMED alone, is limited to terms contained in SNOMED. Thus, concepts can be missed in the textbook when the terms used in the textbook differ from the selection of terms contained in SNOMED. As an example, the term "Tear production," which was used in the textbook did not result in a SNOMED match, as SNOMED represents that concept with the terms, "Lacrimation" and "Lacrimal gland function, NOS." The second technique, Metathesaurus-enhanced lexical matching utilizes non-SNOMED terms from the source vocabularies of the Metathesaurus to identify SNOMED concepts in free-text that are not identified using lexical matching and SNOMED alone.

Source vocabularies in the Metathesaurus contain unique terms in addition to terms held in common with other vocabularies. A set of unique concepts along with the preferred lexical form of each (a specific term) has been identified from among the terms contained in the source vocabularies. These concepts are the core components of the Metathesaurus. Each concept has been assigned a unique concept code referred to as a unique concept identifier (CUI). The CUI is used to link synonymous terms from among the individual source vocabularies. Terms from a source vocabulary are assigned the CUI of the Metathesaurus concept the term represents. As an example, the term "Lacrimal gland function, NOS" from SNOMED is designated by the UMLS as the preferred form of the concept and is assigned the CUI of C0234678. The terms "Tear production," "Lacrimation," and "Lacrimal Gland Functions" contained in other source vocabularies are considered synonyms by the Metathesaurus and therefore are assigned the same CUI, C0234678.

Metathesaurus-enhanced lexical matching used non-SNOMED terms when searching the physical examination textbook. Each matching non-SNOMED term was checked for a SNOMED equivalent using the links among synonymous terms contained in the Metathesaurus. Thus, returning to the previous example, the term, "Tear production" was found in the textbook and is a non-SNOMED term. The Metathesaurus-based link (CUI) between the term, "Tear production" and the equivalent SNOMED terms, "Lacrimation" and "Lacrimal gland function, NOS" was used to document the presence of the concept in SNOMED. Thus, the SNOMED representations of the concept were identified as a result of lexical matching, even though the character strings of the specific SNOMED terms were not contained in the textbook.

During analysis of the Metathesaurus-enhanced technique, it was observed that a single Metathesaurus concept was occasionally mapped to two distinct SNOMED concepts. An example was the Metathesaurus concept “Retinal Arteries” C0035301 mapped to the SNOMED concepts, “Retinal artery, NOS” termcode T-AA380 and “Central Retinal Artery” termcode T-45430. Thus, SNOMED considers the concepts as distinct, while the Metathesaurus does not. If there was complete agreement between a source vocabulary and the Metathesaurus regarding the synonymy of terms, all synonyms for a given concept in a source vocabulary would share the same vocabulary-based concept code and the same CUI. Conversely, all terms considered not to be synonymous in a source vocabulary would have differing vocabulary-based concept codes and differing CUIs. This disagreement regarding synonymy is referred to in this work as, explicit synonym disagreement and is addressed in Chapter 4 entitled, “Explicit Synonym Disagreement between the Metathesaurus of the Unified Medical Language System and its Source Vocabularies.”

The second objective, development and evaluation of automated techniques for identifying canine physical examination terms not documented in the UMLS was addressed in Chapter 5 entitled, “Use of the Metathesaurus and SPECIALIST Lexicon of the Unified Medical Language System, Lexical Matching and Domain-Specific Free-Text to Identify Undocumented Vocabulary.” This technique uses lexical matching in domain-specific free-text to locate and remove terms contained in the Metathesaurus or SPECIALIST Lexicon. The remaining character strings of text are undocumented in the UMLS and are presented to a domain expert along with their original sections of text for manual review. This technique is based on the premise that medical literature is a rich source of existing terminology and serves as an early sentinel for new terminology.

SUMMARY

This research begins with the Patient Information Acquisition Model (PIAM) which provides a generalized overview of the information gathering process. Inherent in the components and activities of PIAM is the need for a complete standardized controlled medical vocabulary. Therefore, this research is directed towards developing automated techniques for identifying documented and undocumented vocabulary. As a byproduct of these studies, explicit synonym disagreement between the Metathesaurus and its source vocabularies is identified and characterized.

REFERENCES

1. Cimino JJ. Saying What You Mean and Meaning What You Say: Coupling Biomedical Terminology and Knowledge. *Academic Medicine* 1993;68:257-260.

2. Moorman PW, van Ginneken AM, van der Lei J, et al. A model for structured data entry based on explicit descriptional knowledge. *Methods Inf Med* 1994;33:454-63.

3. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. *Methods Inf Med* 1993;32:281-91.

4. Rijnberk A, de Vries HW. Medical History and Physical Examination in Companion Animals. Dordrecht,, The Netherlands: Kluwer Academic Publishers, 1995;376.

CHAPTER 2

The Patient Information Acquisition Model: A Framework for Identifying Causes of Missing or Inaccurate Information

ABSTRACT

The Patient Information Acquisition Model (PIAM) illustrates the decision-to-perception chain which links a clinician's decision to collect information, either personally or through another, with the perception of the resulting information. PIAM applies to a wide range of information acquisition scenarios such as acquiring a CBC, performing a cardiac auscultation, a third party observation and a patient history question. PIAM is useful as a diagnostic tool for improving the quality of patient information by serving as a framework for a systematic approach to identifying causes of missing or inaccurate information. Each component and activity of PIAM is associated with characteristic patterns of information corruption, and breakdowns in communication which can be diagnosed through associated questions. PIAM is also useful during the design and evaluation phase of a data collection instrument. Each component of the instrument may be examined for conformity to the principles illustrated by the components and activities of PIAM. As PIAM is technology-independent, it is useful for designing and evaluating paper-based as well as computer-based data collection instruments. Finally, PIAM serves as an orientation framework for research involving areas such as, communication skills, form design, training, inventory management, standards, quality assurance, human-computer interaction, data storage, multi-media presentation, decision support, and cognitive science.

INTRODUCTION

The goal of case management is to reduce the uncertainty surrounding a patient's condition. Uncertainty is reduced through a repeating cycle of decisions, actions, and information acquisition. The cycle begins when the clinician realizes that additional information is needed and decides on an action that should result in generating the desired information. When the action is taken, newly generated information serves as feedback to the clinician. This new information is used to evaluate the previous decisions and to provide support for following decisions. Fluid therapy for severe dehydration provides a suitable example. An evaluation leads to a detailed description of a severely dehydrated patient. The description provides the information to support the decision regarding the selection of fluids and additives as well as amounts and rates of administration. The clinician then decides that fluid intake/output monitoring is indicated to reduce uncertainty regarding fluid therapy adjustments. Each healthcare decision is followed by an action that results in new information for the purpose of reducing uncertainty.

Healthcare decisions will always be made under of two types of uncertainty; those uncertainties which can be controlled and those uncertainties which cannot be controlled. We have developed a model to help decision makers focus on two causes of uncertainty which can be controlled: missing information and inaccurate information. Missing or inaccurate information may be the result of improper requests, inadequate acquisition, incomplete recording, or inappropriate presentation. In addition to increased uncertainty, missing or inaccurate information decreases the validity and reliability of other related information. In our model, decision making by healthcare providers is represented by an information feedback cycle, which we will refer to as the decision-to-perception chain.

The decision-to-perception chain is the sequence of events beginning with a decision to take action and ending with the decision maker's perception of the resulting information. Perception in this instance refers to the detection of information by the decision maker, and is independent of analysis and interpretation. The decision-to-perception chain is a natural process used by everyone on a daily basis. As an example, the chain is employed during cooking when a series of taste tests are used to evaluate the need for additional ingredients or seasoning. Decisions are made, actions are taken, and information is acquired that results in new decisions with additional actions.

The decision-to-perception chain has numerous opportunities for failure. First, the decision-to-perception chain is subject to personal bias because the chain begins and ends within the decision maker. Second, an action cannot be assumed to have been performed as expected, especially if the action was performed by another individual. Third, the appropriate data from the action cannot be assumed to have been generated, captured, and recorded. Fourth, the resulting information cannot be assumed to have been accurately portrayed to and perceived by the decision maker.

Information is vitally important to the healthcare system; and a systematic approach to reducing uncertainty by identifying and correcting the causes of missing or inaccurate information is needed. The Patient Information Acquisition Model (PIAM) in Figure 2.1, provides a framework for this systematic approach.

BACKGROUND

Information feedback cycles depicting the quantity and quality of information resulting from an action are not unique and are described in the literature from various fields of study.   In the field of medicine, Card and Good[1] describe an information acquisition process as an "information channel" between a patient event and its corresponding recorded indicant. An event is any feature experienced by the patient having a bearing on health. Examples would include intestinal pain, dizziness, or nausea. An indicant, according to Card and Good is the entry in the medical record that describes the patient's event or experience. ("Indicant" is also used by some to refer to the finding itself, rather than its related entry in the medical record.) The example used by Card and Good dealt with heartburn as an event that occurred within a patient, resulting in nerve impulses traveling to the brain and causing the patient to experience an unpleasant sensation. The unpleasant sensation was coded by the patient's brain into nerve impulses leading to the utterance of sound that was a description of the sensation. The ear of the doctor picked up the sounds that caused impulses to be sent to her mind for comparison and analysis. The doctor's mind responded by sending nerve impulses to the hand to record the event. The authors point out, "There are many weak links in such a channel, and only a degree of probability that an event taking place within a patient is followed by a correct recording of the corresponding indicant. Analysis of such probabilities leads to a description of error." This error is missing or inaccurate information.

In the field of human-computer interaction, Norman[2] introduced a seven stage model of user activities that describes the performance of a task along with an information feedback mechanism. The activity stages are goals, intention, action specification, execution, perception, interpretation, and evaluation. The premise of his model is that someone has a goal to accomplish with a computer such as open a file, delete a file, or exit a program. The person decides how to accomplish the goal, makes the attempt, monitors the outcome in the form of the computer's feedback response, and decides if the goal was accomplished.

Two models of information management from the field of industrial engineering, the information-feedback system[3] and the Management System Model (MSM)[4,5] had a strong influence on the development of PIAM.   The parallel relationships among these two models and PIAM are shown in Table 2.1. The information-feedback system in Figure 2.2 consists of three components, each of which is also found in PIAM: a decision, an action, and information. This model illustrates the cyclical relationship between a decision, its corresponding action, and the resulting information that is produced in support of the next decision.

The second model that is fundamental to the development of PIAM is the Management System Model (MSM) shown in Figure 2.3. The MSM depicts a management system using three basic components: "Who manages," "What is managed," and "What is used to manage." The "Who manages" component is anyone who makes decisions regarding an area of responsibility. The "What is managed" component represents the physical items for which a decision maker is responsible. Examples would include personnel, equipment, and supplies. The “What is used to manage” component is a management tool that converts data into information such as surgery schedules, plans, and procedures. A surgery schedule is a management tool in that it converts data such as names, times, and locations into information such as who is supposed to be where, and when they are supposed to be there.

The MSM uses three interfaces to represent the interactions among the three components. The three interfaces which are also shown in Figure 2.3 are decision/action, measurement/data, and information portrayal/information perception. The activities represented by these three interfaces are critical for maintaining the integrity of the decision-to-perception chain. A decision must first be converted into action. This conversion is represented by the decision/action interface. The action is performed on the “What is managed” component and results in measurable changes. The measurable changes must be converted into data. This conversion is represented by the measurement/data interface. The resulting data are stored in the “What is used to manage” component and must be transformed into information and portrayed to the decision maker. The decision maker must then perceive the information. The difference between information as it is portrayed and information as it is perceived is represented by the information portrayal/information perception interface.

The MSM contains the same components and the same sequence of information flow as the information-feedback system, but the layout is reversed. This reversal was done for two reasons. First, logically the decision maker (one who initiates the decision-to-perception chain) is placed in the upper left corner of the model because it follows the more familiar western version of literary flow, left to right and top down. Second, the foundational importance of information is better emphasized by moving the information related component to the bottom of the model.

The MSM differs from the information-feedback system in several ways. First, it identifies additional components and activities. Second, the three main components of the MSM (Who manages, What is managed, and What is used to manage) provide concrete links to physical objects and individuals rather than to the products of their activities (decision, action, and information). Third, the decision, action, and information components of the information-feedback system were moved to interfaces between objects in the MSM. Fourth, the information component of the information-feedback system has been expanded in the MSM to include both information portrayal and information perception. Finally, the MSM emphasizes information management tools as opposed to the information itself. A more in-depth discussion of the MSM, its extensions and applications can be found in the following references.[4-7]

The functionality of each of these models is limited for use in illustrating the patient information acquisition process and diagnosing causes of missing or inaccurate information. Card and Good do not represent the healthcare provider as a proactive decision maker nor do they address information portrayal to, and perception by others as is necessary in a feedback cycle. Norman's seven stage model of user activities is task oriented and does not represent data storage or the use of third-party individuals to perform tasks. The information-feedback system is elegant in its simplicity, but too general for specific application. The Management System Model (MSM) is simple and robust, but was designed for building and using management tools, and does not include the necessary detail at the interfaces for identifying the causes of missing or inaccurate information. Kurstedt has noted the possibility of expanding two of the three interfaces. He identifies the possible role of an "actuator" at the decision/action interface, and the possibility of another component including a conversion process at the measurement/data interface. These additions were not made to the MSM, as they introduced complexity that did not contribute to the model's purpose.

The Patient Information Acquisition Model (PIAM) differs from previous work in that it is general enough to represent many different information gathering scenarios and detailed enough to assist healthcare providers and information systems designers in the identification of specific causes of missing or inaccurate information. PIAM addresses the Management System Model (MSM) interfaces explicitly by replacing each interface with an additional component, adding new activities and redefining others. Most notably, PIAM places decision and perception internal to the decision maker.

The Patient Information Acquisition Model

The Patient Information Acquisition Model (PIAM) is a generalized representation of the decision-to-perception chain and serves as a framework for a systematic approach to diagnosing the cause or causes of missing or inaccurate information. Each component and activity of PIAM serves to focus the real-world knowledge an individual has of the healthcare organization and its processes on a specific information gathering scenario.

The components and activities of PIAM are described using the scenario of acquiring a patient's complete blood count (CBC). This scenario is a typical example of the decision-to-perception chain and is described first in general, then in detail. Characteristic causes of missing or inaccurate information associated with each component and activity are identified and presented as they relate to maintaining the decision-to-perception chain. Three additional information gathering scenarios are described to illustrate PIAM's broad application. They are 1) a cardiac auscultation where the clinician is the sole actor, 2) a third-party observation where another individual is responsible for an observation, and 3) a patient history question where a question and its answer are the equivalent of a test and its result.

CBC Scenario

The CBC scenario begins when a clinician decides that additional patient information contained in a CBC is needed and initiates a request. A phlebotomist receives the request, draws a blood sample from the patient and delivers the sample to the laboratory. A CBC is performed on the blood sample, the results are stored, and a report is issued. The laboratory report contains data acquired from the hematology instrument along with the morphologic findings from the medical technologist and clinical pathologist. The results are stored in a computer-based or paper-based medical record system. The clinician locates the information by finding the medical record, querying the database, or contacting the laboratory for a voice report. The clinician perceives the CBC information in the report and formulates a plan of action in response. Table 2.2 illustrates the CBC scenario, the corresponding components and activities of PIAM, and characteristic causes of missing or inaccurate information that can be associated with each component and activity.

The oral or written order (**Requesting an action**) for a CBC initiates the decision-to-perception chain and usually entails a personal communication or the completion of a form. Forms must be intuitive, especially in the case of a test that is requested infrequently. Clear and unambiguous communication with the individual performing the action, whether in person or written, is essential.

The individual who performs an action (**Who acts and with what**) must be adequately trained and possess the necessary equipment and materials. The way an action is performed (**Performing the action)** can have a dramatic effect on the resulting information. In this case, a phlebotomist requiring multiple attempts using poor technique may acquire a sample more indicative of the patient's high anxiety or response to trauma, than the patient's true condition.

The patient's cooperation (**What is acted on)** is critical to the integrity of the decision-to-perception chain. The lack of cooperation from the patient can be a major cause of missing or inaccurate information. In this example, a patient who resists or even refuses venipuncture will affect the test results.

Each action performed on a patient has a set of potential **(Observations)** that can be detected and converted into data. In a CBC, these potential observations are blood cell types, numbers, and morphologic characteristics. It is important to note that according to PIAM, observations are present, modified, or absent independent of their detection and measurement.

The (**Detection and conversion of observations to data**) is a process that must be standardized in order for the data to be valid and reliable. First, the observable parameters must be detected and measured accurately. Second, the observation must be converted into data in a standardized and consistent manner; i.e., the automated blood cell counter must have the ability to distinguish cells from debris and a medical technologist must observe and interpret the observations diligently and competently. Standardization of instruments is well documented and reasonably straight forward, but standardization among human observers is more difficult. Each observer has internalized criteria that over time are subject to variation within themselves and among their colleagues. Even if observers could have identical perceptions, the vocabulary used to describe patient findings does not always evoke the same mental representations in others.

Once the data have been generated, they need to be recorded (**Data recording**). The blood cell counter may enter data directly into a computer-based system while the medical technologist or clinical pathologist may document their findings manually. Instrumented data entry is the least prone to error due to automation. Manual data entry on the other hand is more prone to error and should utilize picklists and validation routines whenever possible to ensure data integrity.

The (**Where results are recorded**) component of PIAM determines the structure, format, and expressiveness of the recorded data. Data elements range from structured and discrete to free text narratives, graphics, sound and video. Examples of structured and discrete data are dates, cell types, and cell counts. An example of free text data would be a medical technologist's description of blood cell morphology. Errors occur when the data storage structures constrain the expressiveness of the data. An example of this type of constraint is a fixed length field that limits the number of characters used in a morphologic description. Even paper-based systems with limited areas for data entry are not immune to this problem.

Next, the clinician composes a query (**Query of selected data**) to select those data that are of interest from all available data in the database. The constraints of a query determine the quality and quantity of the resulting information. A query represented by the statement, "Give me all white blood cell counts for patient x that were performed in the last thirty days," has the constraints of white blood cell counts, patient x, and those counts performed within the last thirty days. If the disease process produces a trend in the levels of white blood cells that occurs during a longer interval than thirty days, this query may produce inadequate information. The capabilities of the query engine can also be a significant cause of missing or inaccurate information. In a paper-based system, searching is limited largely to a visual scan of the medical record or use of an index.

The (**Report**) is the medium or vehicle by which results of a query are communicated. The report is not the results of the query. A report can be paper-based, computer-based, or oral. This distinction is important because it allows one to address causes of missing or inaccurate information associated with the method of reporting, apart from the implementation. The implementation is the organization and presentation of data contained in the report. As an example, an email-based laboratory reporting system has the potential strength of immediacy, but has a potential weakness in that the system depends on the reliability of the network and the promptness of the clinician in retrieving email.

Once a report medium is chosen, the specific characteristics of how the data are presented is represented in PIAM by (**Information portrayal**). Clinicians vary in their preference for information portrayal; some prefer text or tables while others prefer graphical displays. A mismatch between information portrayal and clinician preference may result in the “appearance” of missing information or an inaccurate perception of information. It is important to note that preference in this context refers to inherent information processing skills rather than mere aesthetics. A reporting system used by multiple clinicians should be flexible enough to portray information in a manner preferred by each. In addition, information portrayal in a report can include supporting information to assist interpretation. For example, reference values may accompany the CBC results, or the results of serial samples may be displayed using a graph in addition to the raw data.

In conclusion, the CBC scenario illustrates a delayed information-feedback system that is characteristic of many types of data acquisition scenarios such as radiology, microbiology, histopathology, etc. The delay could range anywhere from minutes to weeks. Turnaround time between sampling and reporting is critical, as the patient’s condition changes continually. Information delayed is information denied.

Cardiac auscultation scenario

The cardiac auscultation scenario is unique in that the clinician is the sole actor. This scenario begins with the clinician deciding to listen to a patient's heart, doing so, and recording the results in the medical record. Table 2.3 illustrates the cardiac auscultation scenario, the corresponding components and activities of PIAM, and the characteristic causes of missing or inaccurate information that can be associated with each component and activity.

The cardiac auscultation scenario highlights the critical role of the clinician in acquiring this type of information. Not only must the clinician perform the auscultation properly, she must be able to hear the heart sounds adequately and convert her observations into a complete standardized description using vocabulary that is consistent in its meaning among multiple individuals with varying backgrounds and training.

Third-party observation scenario

The third-party observation scenario is one in which another individual is asked to make an observation on behalf of the clinician. The third-party could be anyone from a trained professional to an untrained layperson. An example is that of a client whose pet is a patient. Here, the clinician asks the client to monitor the consistency of the pet's stool while the pet is receiving an antibiotic. The client is the **(Who acts and with what)**, in the same manner as the phlebotomist in the CBC scenario. However, in this example the action is in the form of an observation and is noninvasive.

The client **(Who acts and with what)** must know how an observation is made (**Performing the action**) and what should be observed (**Observations**). If the clinician says, “Monitor your pet for any diarrhea.”, the clinician must be sure the client understands what is meant by the term, “diarrhea.” The client must also know how to interpret, record, and describe observations to the clinician. These tasks are represented by the chain of components and activities in PIAM beginning with (**Who acts and with what**) and ending with (**Information presentation**). The principles in this scenario apply in a human medical setting to self-observation by a patient as well. As a side note, the clinician should also be certain that the client knows what to do or whom to notify if diarrhea is observed. The carefully-observed and well-documented disaster is not foreign to veterinary medicine.

Patient history question

Eliciting a patient’s history illustrates the role of questioning as an action, listening as an observation, and interpreting and recording answers as converting observations into data. Each question serves the same function as a laboratory test, with the answer being the result.

Questioning as a “testing” procedure is difficult to standardize. The communication skills, mental states, physical health, and recent circumstances of participants can have a dramatic effect on the quality of information that results from the dialog. Questioners must avoid leading questions, and questions that are perceived as threatening. In order to maintain data integrity, the questioner should look for discrepancies among a patient’s answers, as patients may be deceptive either intentionally or unintentionally. The patient’s voice tone or body language can be a rich source of information, but care should be taken not to over interpret. Recording patient responses can be a source of missing or inaccurate information, as lengthy and involved answers can lose meaning when reduced to a manageable medical entry. Failure to record the context of a patient's statement can also lead to missing and inaccurate information. A medical entry that seems clear to the author who is currently immersed in the context of an encounter can often become ambiguous to others or even the author later on.

DISCUSSION

The Patient Information Acquisition Model (PIAM) applies to a wide range of information acquisition scenarios.   The CBC scenario depicts a clinician who is not personally involved in the data acquisition process.   The cardiac auscultation and patient history scenarios depict the clinician as solely responsible for the data acquisition process, with the former scenario being a physical intervention and the latter being verbal. In the third-party observation, another individual is collecting, interpreting, and recording patient data on behalf of the clinician. All four scenarios illustrate how the general principles of PIAM can be used to identify causes of missing or inaccurate information by associating characteristic questions with each component and activity. Table 2.4 lists the components and activities of PIAM along with their characteristic questions.

PIAM is applicable to scenarios within scenarios. In the CBC scenario, the phlebotomist has her own implementation of PIAM and becomes the decision maker when deciding which blood vessel to use, how to approach the patient, and what types of restraint may be indicated. The phlebotomist evaluates the results of the approach and may enter the information into the medical record as feedback for the next time a blood sample needs to be drawn.

PIAM is also useful for identifying global information acquisition problems within a healthcare organization. A global problem is indicated when the same general component or activity of PIAM is associated with several breakdowns from different information gathering scenarios. If healthcare providers consistently make mistakes on several different forms, the problem may be related to the design of the forms overall, or the diligence or training of the healthcare providers. Patterns of information acquisition breakdowns such as this, indicate that a problem exists at the organizational level and interventions should be directed at this level rather than at the individual scenario alone.

The components and activities of PIAM can be used as checkpoints during the design and evaluation phase of a Hospital Information System (HIS). Each information gathering process and sub-process can be examined according to the components and activities of PIAM to insure that the HIS supports information acquisition according to the principles of the Patient Information Acquisition Model (PIAM). Since PIAM is technology-independent, it is useful with paper-based as well as computer-based HIS.

PIAM implies that the locus of control is with the clinician, but many clients have a differing view. Many clients view themselves as the (**Who decides on an action**) component, with the clinician being the (**Who acts and with what**) component. These clients reject the saying, "The successful outcome of this case may well be determined by which one of us gets to be the doctor." In fact, the client usually is the one who decides to initiate and maintain a healthcare relationship. Therefore, both the client and the clinician experience their own versions of PIAM. A successful healthcare relationship depends how the locus of control is either shared or transferred.

One might question where a HIS would fit into PIAM. The HIS, whether paper-based or computer-based, supports the entire patient information acquisition process as shown in Figure 2.4. The HIS can assist the (**Who decides on an action**) with expert systems, practice guidelines, adverse drug event monitoring, and educational resources. The HIS can process the (**Requesting an action**), assist the (**Who acts and with what**) with procedures and protocols, and participate in the (**Performing the action**) component. Standards and criteria for (**Detecting and converting observations into data**) can also be maintained by the HIS. Finally, data entry forms, data storage structures, queries, and reports with adaptive capabilities for information portrayal can also be supported by the HIS.

Finally, PIAM serves as an orientation for specific areas of research such as, communication skills, form design, training, inventory management, standards, quality assurance, human-computer interaction, data storage, multi-media presentation, decision support, and cognitive science. PIAM can also be used to organize a course on information management.

LIMITATIONS

All models have limitations, and PIAM is not without its share. First, PIAM is a closed-loop model and does not represent the initiating piece of information such as, "There is a patient in exam room 1 that has been vomiting occasionally for 2 weeks." Second, PIAM does not represent external factors that can influence decision making explicitly, such as client preferences or financial constraints. Perhaps PIAM can be placed within a larger model that takes these factors into account. Third, even though PIAM models the decision-to-perception chain, it does not model the decision or perception processes explicitly, inasmuch as these processes are internal to the decision maker. Models of cognition and perception are best suited to explore these issues and fit well within the (**Who decides on an action**) component.[8]

FUTURE WORK

It is not the goal of PIAM to transform the healthcare provider into a systems analyst or an organizational engineer. PIAM provides the framework for a systematic approach to diagnosing deficiencies within information systems and inefficiencies within organizations. Each component and activity of PIAM serves to focus the real-world knowledge an individual has of the healthcare organization and its processes on a specific information gathering scenario. Future work involving the Patient Information Acquisition Model will center around establishing the model's integrity and demonstrating its application. Integrity studies would entail asking clinicians to document incidents of missing or inaccurate information in the clinical setting and evaluate PIAM's ability to represent each incident. In addition, the list of characteristic questions associated with each component and activity of PIAM needs to be evaluated for completeness in identifying these incidents. PIAM also needs to be evaluated for effectiveness as a design model by using its framework to develop medical or non-medical information systems. Third, PIAM should also be evaluated for completeness when applied as a framework for organizing topics in a course on information management. It is anticipated that PIAM would provide an effective conceptual framework to support student learning.

Table 2.1 A mapping of the components and activities of the Information-Feedback System, the Management System Model, and the Patient Information Acquisition Model.

INFORMATION-FEEDBACK SYSTEM	**MANAGEMENT SYSTEM MODEL**	**PIAM COMPONENTS/ACTIVITIES**
	Who manages	**Who decides on an action**
Decision	**Decision**	*Decision is internal to the decision maker
		Requesting an action
		Who acts and with what
Action	**Action**	**Performing the action**
	What is managed	**What is acted on**
	Measurements	**Observations**
	Data	**Detection and conversion of observations to data**
		Data recording
	What is used to manage	**Where results are recorded**
		Query of selected data
		Report
Information	**Information portrayal**	**Information portrayal**
	Information perception	*Perception is internal to the decision maker

Table 2.2 A mapping between the Patient Information Acquisition Model, the CBC scenario, and potential causes of missing or inaccurate information.

PIAM Components/Activities	**CBC Scenario Counterparts**	**Potential Causes of Missing or Inaccurate Information in the CBC scenario**
Who decides on an action	the clinician	inappropriate decision
Requesting an action	verbal order or request form	the CBC was not requested
		the CBC was requested improperly
Who acts and with what	the phlebotomist	the “Who acts” is unqualified
		the lack of necessary supplies or instruments
Performing the action	drawing the sample	task execution was ineffective
	sample handling and delivery	the blood sample was mishandled
What is acted on	the patient	the patient was uncooperative or unavailable
Observations	the observable parameters of the blood	the anticipated observations were not present or were altered in the sample
Detection and conversion of observations to data	the hematology instrument	inaccurate or poorly maintained instruments
	the technologist or clinical pathologist	inaccurate or poorly trained personnel
		use of non-standard terminology
		undetected observations
Data recording	automated data entry	software/hardware failure
	manual data entry	inaccurate data entry
		non-intuitive methods of data entry
Where results are recorded	the medical record system	mismatch between storage structure and data
Query of selected data	locate the CBC report	inadequate skills for querying the database
		data storage structures preclude effective queries
		misplaced records (paper-based systems)
Report	the tangible CBC report	incomplete reports
Information presentation	CBC information layout	confusing presentation of information

Table 2.3 A mapping between the Patient Information Acquisition Model, the cardiac auscultation scenario, and potential causes of missing or inaccurate information.

PIAM Components/Activities	**Cardiac Auscultation Counterparts**	**Potential Causes of Missing or Inaccurate Information in the Cardiac Scenario**
Who decides on an action	clinician	not applicable
Requesting an action	clinician	not applicable
Who acts and with what	clinician	the clinician is unqualified
Performing the action	clinician	task execution was ineffective
What is acted on	patient	the patient was uncooperative
Observations	clinician hears the heart sounds	not applicable
Detection and Conversion of observations to data	clinician interprets the heart sounds	poorly trained clinician
		use of non-standard terminology
		unobserved observations
Data Recording	clinician enters the findings	incomplete data entry
		inaccurate data entry
Where results are recorded	medical record system	data storage structure does not match the data
Query of selected data	locate the report	inadequate skills for querying the database
		the data storage structures preclude effective queries
Report	report form	incomplete reports
Information portrayal	results of cardiac auscultation	confusing presentation of information
*Information perception internal to the decision maker		lack of diligence by the clinician

Table 2.4 Components and activities of the Patient Information Acquisition Model along with associated questions for diagnosing the cause or causes of missing or inaccurate information.

PIAM COMPONENTS AND ACTIVITIES	QUESTIONS ASKED WHEN DIAGNOSING THE CAUSE(S) OF MISSING OR INACCURATE INFORMATION
Who decides on an action	Was the action indicated?
Requesting an action	Was the action requested?
	Was the action requested correctly?
	Was the person performing the action notified?
Who acts and with what	Was the person performing the action trained?
	Was the person performing the action equipped?
Performing the action	Was the action executed?
	Was the action executed correctly?
	Did random variability affect the results?
What is acted on	Did the patient cooperate?
Observations	Did the desired observations occur?
Detection and conversion of observations to data	Were the observations observed?
	Were observations observed correctly?
	Was the conversion to data done properly?
Data recording	Was the data entered?
	Was the data entered correctly?
Where results are recorded	Are the data storage structures adequate?
Query of selected data	Was the query requested?
	Was the query requested correctly?
	Was the information available for query?
Report	Was the report generated?
Information portrayal	Was the information contained in the report?
	Was the information presented clearly?

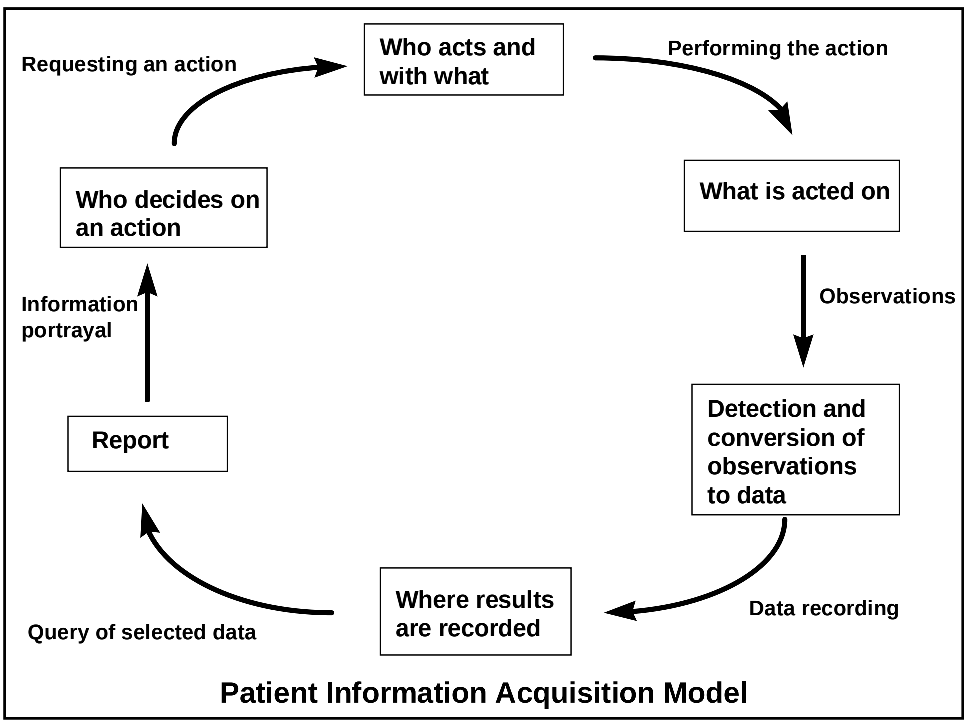


Figure 2.1 Patient Information Acquisition Model (PIAM)

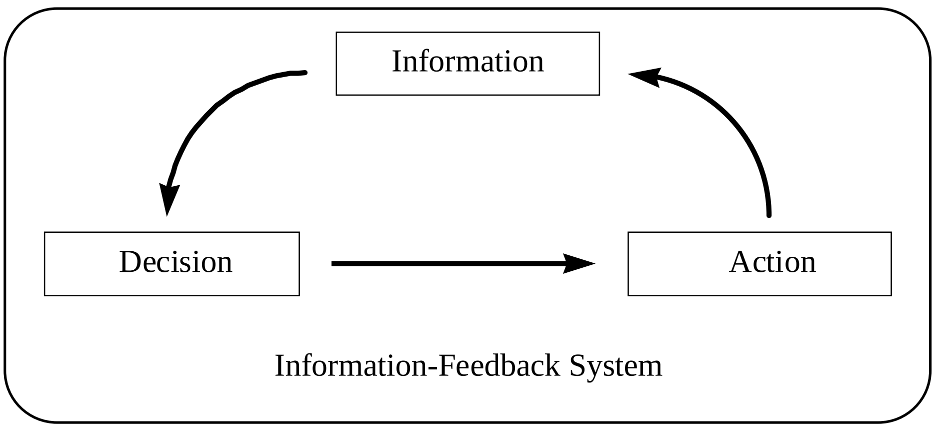
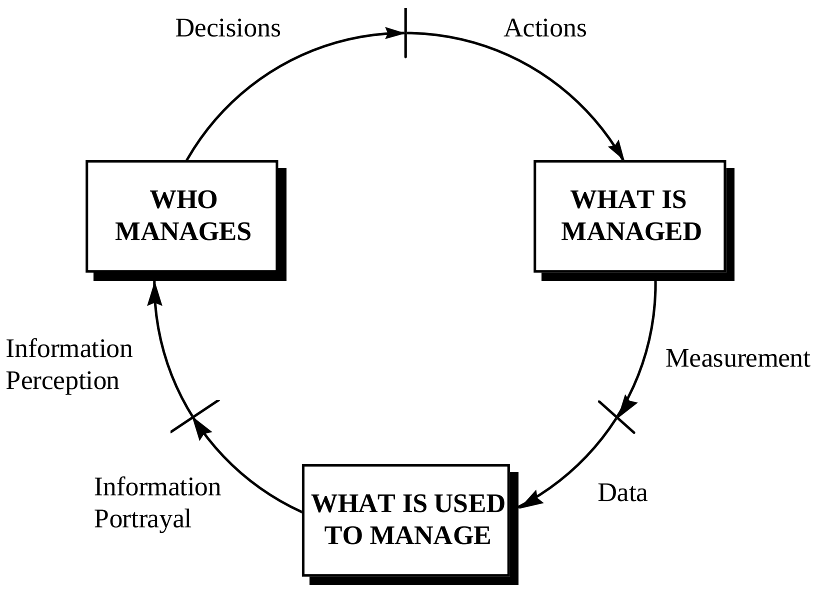


Figure 2.2 Information-Feedback System[3].



Management System Model (MSM)

Figure 2.3 Kurstedt's Management System Model (MSM)[4].

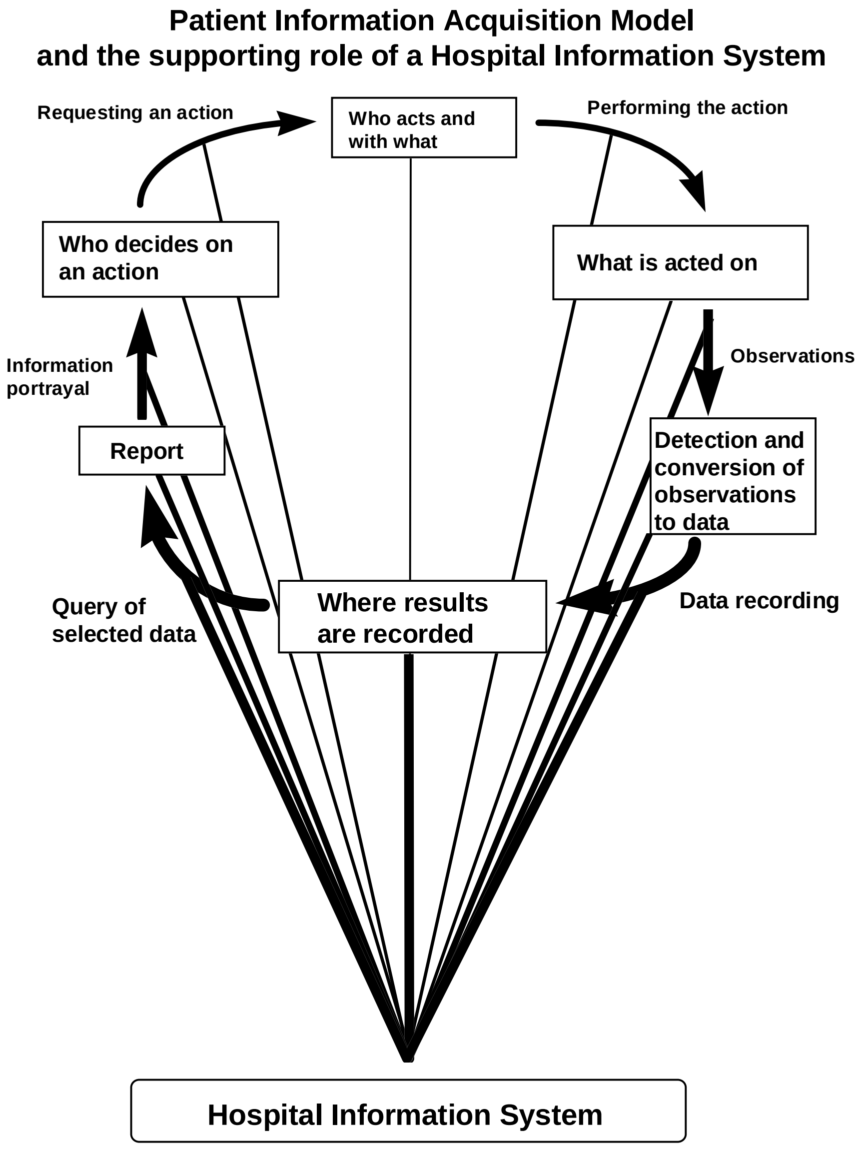


Figure 2.4 The Hospital Information System (HIS) supports information acquisition according to the principles of the Patient Information Acquisition Model (PIAM).

REFERENCES

1. Card WI, Good IJ. A logical analysis of medicine In: R. Passmore and J. S. Robson, eds. *A Companion to Medical Studies*. Oxford Edinburgh: Blackwell Scientific Publications, 1973.

2. Norman DA, Draper SW. User centered system design: new perspectives on human-computer interaction. Hillsdale, N. J.: Lawrence Erlbaum Associates, 1986.

3. Forrester JW. *Industrial Dynamics*. Cambridge, MA: MIT Press, 1961.

4. Kurstedt HA, Jr.,, Mendes PM, Lee KS. Engineering Analogs in Management, in *Proceedings of the Ninth Annual Conference*. American Society for Engineering Management 1988;197-202.

5. Kurstedt HA, Jr.,. Course Notes: Management Systems Theory, Applications, and Design, 1994.

6. Mallak LA. Applying the Management System Model to a Federal Government Organization. *Industrial Engineering and Operations Research*. Blacksburg Virginia: Virginia Polytechnic Institute and State University, 1986;203.

7. Kleiner BM. An Integrative Framework for Measuring and Evaluating Information Management Performance. *International Journal of Computers and Industrial Engineering* 1997;(in press).

8. Elson EB, Faughnan JG, Connelly DP. An Industrial Process View of Information Delivery to Support Clinical Decision Making: Implications for Systems Design and Process Measures. *J Am Med Inform Assoc* 1997;4:266-278.

CHAPTER 3

Metathesaurus-Enhanced Lexical Matching in Domain-Specific Free-Text

ABSTRACT

In this research we examined two complementary lexical matching techniques for identifying concepts from the Systematized Nomenclature of Medicine and Veterinary Medicine (SNOMED) contained in free-text, 1) lexical matching using SNOMED version 3.3 terms alone and 2) Metathesaurus-enhanced lexical matching using non-SNOMED terms from the source vocabularies of the Metathesaurus of the Unified Medical Language System (UMLS). Lexical matching using SNOMED alone resulted in 2212 matched terms representing 1802 unique concepts. A random sample of 331 matched terms resulted in 312 (94%) word-sense appropriate terms, of which 308 (99%) were relevant to veterinary medicine with 193 (62%) relevant to the canine physical examination in particular. Unfortunately, lexical matching using SNOMED alone missed concepts used in free-text when terms used in the text differed from the selection of terms contained in SNOMED. The second technique, Metathesaurus-enhanced lexical matching utilized the combined lexical strength of non-SNOMED terms from the source vocabularies of the Metathesaurus to identify SNOMED concepts in free-text that were not identified using lexical matching and SNOMED alone. Matching non-SNOMED terms were traced to their SNOMED equivalents using links among synonymous terms provided by the Metathesaurus. Metathesaurus-enhanced lexical matching identified 208 additional SNOMED concepts. A random sample of 136 of these additional SNOMED concepts and the corresponding non-SNOMED term of each was evaluated in two steps. First, the matching non-SNOMED term for each concept was evaluated for word-sense appropriateness. One hundred eleven (82%) of the 136 non-SNOMED terms were judged as word-sense appropriate. Second, of these 111, 97 (87%) were linked to SNOMED concepts considered concept-sense appropriate with the original text. Of the 97

concept-sense appropriate concepts, 97 (100%) were judged relevant to veterinary medicine in general with 57 (59%) judged relevant to the canine physical examination in particular. Forty of the 97 suggested SNOMED concepts were discovered through the match of non-SNOMED terms that were plural forms of the suggested SNOMED concept.

Examination of the 14 SNOMED concepts that were not concept-sense appropriate, but had been derived from word-sense appropriate matches, resulted in the discovery of explicit synonym disagreement between the Metathesaurus and its source vocabularies. Explicit synonym disagreement occurs, 1) when terms within a single concept group in a source vocabulary are mapped to multiple Metathesaurus concepts, and 2) when terms from multiple concept groups in a source vocabulary are mapped to a single Metathesaurus concept.

INTRODUCTION

This research originated from a desire to build a standardized data collection instrument for documenting the canine physical examination. In order to build such an instrument, one must first identify a set of terms for use in acquiring physical examination data. The Systematized Nomenclature of Medicine and Veterinary Medicine (SNOMED) has been chosen as the standardized controlled vocabulary for veterinary medicine by the American Veterinary Medical Association. Thus, the focus of this work is on the identification of canine physical examination terms contained in SNOMED.

The approach used initially was lexical matching of SNOMED terms in a physical examination textbook. However, this approach is limited in that concepts can be missed when terms used in the textbook differ from the nomenclature (selection of terms) contained in SNOMED. As an example, the term "Tear production," which was used in the textbook did not result in a SNOMED match, as SNOMED represents that concept with the terms, "Lacrimation" and "Lacrimal gland function, NOS." Thus, lexical matching is limited by the selection of terms (language-based) for a given set of concepts (thought-based)[1].

Synonyms have been shown to account for differences in nomenclature when using lexical matching to map one vocabulary to another.[2]. Kannry et al. added synonyms to pharmacy terms prior to lexical matching. Our research uses existing mappings among synonyms from multiple vocabularies to enhance lexical matching by combining their nomenclatures. This approach is referred to as Metathesaurus-enhanced lexical matching, and is presented as a complementary technique to lexical matching using SNOMED terms alone.

Metathesaurus-enhanced lexical matching utilizes non-SNOMED terms from the source vocabularies of the Metathesaurus of the Unified Medical Language System[3] to identify SNOMED concepts in free-text that were not identified using lexical matching and SNOMED alone. Every non-SNOMED term that results in a lexical match is checked for a SNOMED equivalent using links among synonymous terms contained in the Metathesaurus. Thus, returning to the previous example, the Metathesaurus-based link between the non-SNOMED term, "Tear production" and the equivalent SNOMED terms, "Lacrimation" and "Lacrimal gland function, NOS" was used to document the presence of the concept in SNOMED. Therefore, the SNOMED representations of the concept were identified as a result of lexical matching, even though the character strings of the specific SNOMED terms were not contained in the textbook.

BACKGROUND

Systematized Nomenclature of Medicine and Veterinary Medicine (SNOMED)

SNOMED version 3.3 contains approximately 100,000 unique concepts represented by over 144,000 terms used in human and veterinary medicine. Each concept in SNOMED is assigned an alphanumeric code referred to as a termcode. Terms that represent a given concept are assigned the concept's termcode. Thus, the terms "Lacrimal gland function, NOS" and "Lacrimation" mentioned previously are considered synonyms for the same concept and therefore share the same termcode, F-F2900.

Throughout this paper a group of synonymous terms will be referred to as a concept group. The authors of SNOMED designate one term from each concept group as the preferred term for representing the concept. The term designated as preferred, is not always the lexical string found most commonly in free-text. (e.g., The term “Lacrimal gland function, NOS” is designated the preferred term over “Lacrimation” even though “Lacrimation” is more likely to appear in free-text.)

Unified Medical Language System (UMLS)

The UMLS, established in 1986 by the National Library of Medicine (NLM), was designed to “facilitate the retrieval and integration of information from multiple machine-readable biomedical information sources. The sources of interest include: descriptions of the biomedical literature, clinical records, factual databanks, knowledge-based systems, and directories of people and organizations.” [4] The UMLS is organized into four Knowledge Sources which are the Metathesaurus, Semantic Network, Information Sources Map, and SPECIALIST Lexicon. Only the Metathesaurus was used in this research.

The 1997 version of the Metathesaurus contains approximately 330,000 unique concepts represented by approximately 740,000 terms from over 40 autonomously controlled medical vocabularies known as, source vocabularies. Each source vocabulary contains unique terms in addition to terms held in common with other vocabularies. A set of unique concepts along with the preferred lexical form of each (a specific term) has been identified from among the terms contained in the source vocabularies. Each concept is assigned a unique concept code referred to in the Metathesaurus as a unique concept identifier (CUI). The CUI is an eight place character string beginning with the letter C followed by 7 numbers.

The CUI is used to link synonymous terms from among the individual source vocabularies. Terms from a source vocabulary are assigned the CUI of the Metathesaurus concept the term represents. As an example, the term “Lacrimal gland function, NOS” from SNOMED is designated by the UMLS as the preferred form of the concept and is assigned the CUI of C0234678. The terms “Tear production,” “Lacrimation,” and “Lacrimal Gland Functions” contained in other source vocabularies are considered synonyms by the Metathesaurus and therefore are assigned the same CUI, C0234678. The 1997 version of the Metathesaurus contains approximately 330,000 unique concepts represented by approximately 740,000 terms.

Lexical Matching

Lexical matching identifies one character string within another and has been noted as a practical, powerful, and productive way to begin construction of a thesaurus. [5] While lexical matching is quick and efficient, it has two main weaknesses. First, the character string of a target term must be an exact character for character match (and possibly case-sensitive) with the search term. Thus, plurals and other lexical variations of a term are missed unless they are represented explicitly. Second, lexical matching does not guarantee semantic equivalence.[2] Lexical matching cannot resolve word-sense ambiguity such as occurs with the term “cold” which can refer to an illness or the temperature, or the uppercase character string “COLD” which is an abbreviation for the term, “Chronic Obstructive Lung Disease.” Semantic processing has been used to address these types of ambiguities. [6,7]

Domain-specific free-text

Domain-specific free-text is a rich source of terminology. Several researchers have used domain-specific free-text along with commercially available natural language processing tools for vocabulary discovery[8,9]. Other researchers have used domain-specific free-text with manual evaluation by domain experts to test the completeness of a controlled medical vocabulary or demonstrate the preference of one vocabulary over another [10-16]. While expert evaluation is effective, it is more costly in terms of time and personnel than lexical matching with manual review. A textbook on the physical examination in companion animals [17] was used in this research as a source of domain-specific free-text. This text was thought to be a good source of physical examination terms and a good pre-filter of terms not related to the canine physical examination.

Two complementary techniques

In this research we examine two complementary lexical matching techniques for identifying SNOMED concepts contained in free-text, 1) lexical matching using SNOMED terms alone and 2) Metathesaurus-enhanced lexical matching. The first technique identifies specific SNOMED terms found in free-text, while the second technique identifies additional SNOMED concepts contained in the Metathesaurus through the use of non-SNOMED terms and their CUI-based links to SNOMED terms.

METHODS

Term list preparation

Two sets of terms were used in this research, 1) non-clerical terms from SNOMED version 3.3[*] and 2) English terms from the 1997 release of the UMLS Metathesaurus. Both sets of terms were imported from their distribution files into relational databases using Visual Foxpro™. A subset of uniquely spelled terms (case-insensitive) from each set was extracted and converted to uppercase.

Some terms in SNOMED and the Metathesaurus contained characters used for classification or notation that would prevent their matching in free-text. Examples of such characters from either source are underlined, “Abdomen, NOS” meaning Not Otherwise Specified or “Cold <1>” and “Cold <2>” with the number used to denote unique concepts. These classification characters and others such as “(NOS)”, “ NEC”, “(Non MeSH)”, and “MeSH Category” were removed using a “pruning” filter technique developed in-house based on ideas put forward by Sherertz te al.[18] Thus, “Abdomen, NOS” was converted to “ABDOMEN” and “Cold <1>” and “Cold <2>” were converted to “COLD.”

[*] SNOMED assigns termcodes to headings and clerical terms used for organizational purposes such as, “CHAPTER 0 DISEASES OF THE SKIN AND SUBCUTANEOUS TISSUES.” These terms were excluded from this research.

Unfortunately, in some instances the removal of clerical and classification characters created ambiguous character strings. An example, is the removal of <1> and <2> from the terms “Cold <1>” and “Cold <2>.” The character string “COLD” would return both concepts when using case-insensitive matching. In addition, some original character strings in SNOMED and the Metathesaurus were also duplicated among multiple concepts. An example from the Metathesaurus is the term “urine,” which is mapped to two distinct concepts, “urine” the body substance and “urine” the diagnostic procedure, as in, “Get me a urine on that patient.” Character strings from SNOMED or the Metathesaurus known to be mapped to multiple concepts, created by pruning or otherwise, were excluded from this study.

Both sets of terms were sorted in order of decreasing word count as in six word terms, then five word terms, and so on prior to matching. A word was defined as any alphanumeric text string delimited by spaces and/or punctuation. Matching in order of decreasing word count prevents multi-word terms from being eliminated or fragmented by single word terms such as would occur with the multi-word term, “abnormal behavior” and the single word terms, “abnormal” and “behavior.”

Text source preparation

The textbook, Medical History and Physical Examination in Companion Animals [17] was used with kind permission from Kluwer Academic Publishers. The Index and chapters relating to the canine physical examination were scanned using a flatbed scanner and converted to a text file using optical character recognition software.†

The resulting text file was spellchecked for scanning errors using Microsoft Word for Windows 95 version 7.0a©. A custom veterinary dictionary was created by parsing the CD-ROM version of the Merck Veterinary Manual [19] into a list of unique words. This list of unique words was saved as an ASCII text file and registered with the Word® spellchecker as a custom dictionary.

Next, the textbook file was converted to a table, one record for each sentence, phrase or heading delimited by punctuation, tab character, or carriage return. Handling sentences, phrases and headings in this manner provided the ability to link a matched term with its text source. Barrier words [20,21] such as "of" and "the" were not used as delimiters between terms, as many vocabulary terms contain typical barrier words. (e.g., "Cupping of the optic disk") Non-alpha characters (numbers and publishing characters) were trimmed from the beginning and end of each sentence, phrase or heading. Numbers imbedded within a character string were not removed.

Matching Techniques

Two complementary lexical matching techniques were performed on separate copies of the textbook file, 1) lexical matching using SNOMED version 3.3 terms alone and 2) Metathesaurus-

† TypeReader® Professional 3.0XA for Windows™ by ExperVision, Inc.

enhanced lexical matching using non-SNOMED terms from the UMLS Metathesaurus to identify additional SNOMED concepts. Flow diagrams of both techniques are shown in Figure 3.1.

In the first technique, each sentence, phrase and heading of the textbook was searched using case-insensitive lexical matching and terms contained in SNOMED version 3.3. When a term was located in the text, the term, its termcode and the unique identifier of the sentence, phrase or heading that contained the term was placed in a results table. The termcode was used to document the identification of the term's concept group, and the unique identifier of the sentence, phrase or heading was used to track the context of the match. Finally, the character string within the sentence, phrase or heading that matched the term was replaced with a meaningless token character.

In the second technique, Metathesaurus-enhanced lexical matching, each sentence, phrase and heading from the textbook was searched using non-SNOMED terms from the Metathesaurus source vocabularies. The CUI of each matching non-SNOMED term was checked electronically to determine if any synonymous SNOMED terms shared the same CUI. When a synonymous SNOMED term was identified, its termcode was compared against the list of SNOMED terms discovered using lexical matching and SNOMED alone. If the termcode of the synonymous SNOMED term was not found, the SNOMED concept was counted as an additional SNOMED concept identified by the Metathesaurus-enhanced technique. All SNOMED terms for the newly identified concept were copied to a results table for manual review along with the non-SNOMED term that had resulted in a lexical match, and the unique identifier of the sentence, phrase or heading that contained the match. All SNOMED terms for the discovered concept were evaluated manually, as there is no way of automatically determining which of the terms representing the concept is preferred for veterinary medicine. On the other hand, if the termcode of the synonymous SNOMED term was found when compared against the list of SNOMED terms discovered using lexical matching and SNOMED alone, the SNOMED concept was not counted as an additional SNOMED concept identified by the Metathesaurus-enhanced technique.

Figure 3.2 illustrates scenarios in which the Metathesaurus-enhanced technique either 1) discovers an additional SNOMED concept or 2) does not discover an additional SNOMED concept. In the first scenario, the term "Tear production" from the Read Thesaurus is a non-SNOMED term that was matched in the textbook. The CUI of "Tear production" is shared by the SNOMED synonyms, "Lacrimal gland function, NOS" and "Lacrimation" which are associated with the termcode, F-F2900. None of the SNOMED terms found in the textbook using lexical matching and SNOMED version 3.3 alone had a termcode of F-F2900. Therefore, the SNOMED concept having the termcode F-F2900, and represented by the terms "Lacrimal gland function, NOS" and "Lacrimation," was discovered as the result of Metathesaurus-enhanced lexical matching.

In the second scenario, the term "Slow pulse" from the Dxplain expert diagnostic system is a non-SNOMED term that was matched in the textbook. The CUI of "Slow pulse" is shared by the SNOMED synonyms, "Bradycardia, NOS" and "Slow heart beat, NOS" which are assigned the termcode, F-33160. However, the term "Bradycardia, NOS" had already been discovered using lexical matching and SNOMED version 3.3 alone. Therefore, the SNOMED concept having the termcode F-33160, was not discovered as the result of Metathesaurus-enhanced lexical matching.

Sampling and Analysis

The first technique, lexical matching using SNOMED alone resulted in a list of 2212 SNOMED terms representing 1802 unique concepts. A random sample of terms was obtained in the following manner.

Each term was assigned a unique random number using the Rand() function of Visual Foxpro™. The list was sorted on the randomly assigned numbers with sampling done sequentially based on sorted order. Sample size was determined using a table of sample sizes from finite populations[22]. The table was constructed based on the assumption of a population proportion of .50 (½ of the terms would belong to a given category, ½ would not) which yields the maximum possible sample size required such that the proportion in each category determined for the sample would represent the proportions in the population with a 95% level of confidence.

The random sample of terms was evaluated for 3 criteria, 1) word-sense appropriateness which determined if the definition of the matching term represented the author's intent, 2) relevance of the term to veterinary medicine and 3) relevance of the term to the canine physical examination.

Word-sense appropriateness required the matching term have a single meaning and represent the author's intent in all occurrences in the text. This gives us a measure of the inherent ambiguity of lexical matching in this domain. Thus, the term “back” was not judged word-sense appropriate as the author used the term to refer to a dog’s “back,” and in the sentence, “Try to bring an over-talkative client back to the point by …”

A term was considered relevant to veterinary medicine if a single conceivable use for the term could be imagined. On the other hand, relevance to the canine physical examination was more restrictive in that it had to pertain to experiencing or describing the physical examination directly. Thus, terms having to do with seeing, touching, hearing, smelling or descriptions of their results were considered valid. Terms representing procedures or topics such as physical examination, deep palpation, illumination, or ophthalmic examination were not considered a part of the canine physical examination directly. In addition, terms requiring a conclusion on the part of the observer were excluded such as, “arthritis” and “ileus.” One can palpate crepitus or distended loops of bowel, but must make a conclusion before attributing such findings to “arthritis” or “ileus.”

The second technique, Metathesaurus-enhanced lexical matching resulted in two lists 1) a list of non-SNOMED terms that matched in the text and 2) a list of the SNOMED concepts suggested as a result of their link with the matching non-SNOMED terms. A random sample of the suggested SNOMED concepts along with the matched non-SNOMED term of each was obtained using the same randomization process and table of sample sizes from finite populations mentioned above.

The non-SNOMED terms that matched in the textbook were evaluated for word-sense appropriateness. The list of suggested SNOMED concepts was evaluated for 3 criteria, 1) concept-sense appropriateness 2) relevance of the concept to veterinary medicine and 3) relevance of the concept to the canine physical examination. Concept-sense appropriateness is used rather than word-sense appropriateness, as the meaning of a concept is being evaluated rather than the meaning of an individual term. All evaluations were made by (EMM) with review by (JRW) to reach consensus.

RESULTS

Lexical matching using SNOMED alone (summarized in Table 3.1) resulted in 2212 matched terms representing 1802 unique concepts. Evaluation of a random sample of 331 terms representing 317 concepts, resulted in 312 (94%) terms judged as word-sense appropriate. Of the 312 word-sense appropriate terms, 308 (99%) representing 295 concepts were judged relevant to veterinary medicine in general, with 193 (62%) terms representing 185 concepts judged relevant to the canine physical examination in particular.

Metathesaurus-enhanced lexical matching (summarized in Table 3.2) resulted in 208 suggested SNOMED concepts. A random sample of 136 SNOMED concepts and the corresponding non-SNOMED term of each was evaluated in two steps. First, the matching non-SNOMED term for each concept was evaluated for word-sense appropriateness. One hundred eleven (82%) of the 136 non-SNOMED terms were judged as word-sense appropriate. Of these 111, 97 (87%) were linked to concepts considered concept-sense appropriate with the original text. This meant the matching non-SNOMED term and its corresponding SNOMED concept were judged synonymous in all but 14 occurrences. Of the 97 concept-sense appropriate concepts, 97 (100%) were judged relevant to veterinary medicine in general with 57 (59%) judged relevant to the canine physical examination in particular. Forty of the 97 suggested SNOMED concepts were discovered through the match of non-SNOMED terms that were plural forms of the suggested SNOMED concept.

DISCUSSION

Metathesaurus-enhanced lexical matching is based on the notion that links among synonymous terms in the UMLS Metathesaurus can be used to leverage the combined lexical strength of its source vocabularies. We use lexical strength to mean the capacity of a vocabulary to recognize concepts in free-text using synonyms. Thus, the more synonyms and therefore, character strings a vocabulary maintains for a given concept, the more likely it is for the concept to be recognized in free-text using that vocabulary.

Lexical strength can apply to a vocabulary's ability to represent a single concept or to the vocabulary in general. At the single concept level, one vocabulary may contain a greater number of synonyms for a given concept, and thus be lexically stronger for that concept. However, the lexical strength of a vocabulary depends on the breadth of its domain coverage as well as the synonym content for individual concepts.

Lexical strength has a quantitative component and a qualitative component. Not only is the number of synonyms important, but the likelihood that the character strings of the synonyms are used in free-text is important as well. Numerous obscure terms, or terms with convoluted structures that are rarely found in free-text are of lesser value in recognizing concepts than a few widely used terms with familiar structures. As an example, the terms “Kidney Failure, Chronic” and “Renal Failure, Chronic” are less likely to be found in free-text than the terms “Chronic Kidney Failure” and “Chronic Renal Failure.”

The Metathesaurus-enhanced lexical matching technique resulted in the identification of 208 (11.5%) additional SNOMED concepts over those discovered using lexical matching and SNOMED version 3.3 alone. In addition, the percentage of these additional concepts relevant to the canine physical examination 57/97 (59%) was consistent with that obtained by lexical matching using SNOMED alone 193/312 (62%).

Forty of 97 (41%) suggested SNOMED concepts were due to plural forms of non-SNOMED terms found in the text. Examples include, “Vocal Folds” suggesting “Vocal fold, NOS,” “Laryngeal Muscles” suggesting “Laryngeal muscle, NOS” and “Hyoid Bones” suggesting “Hyoid bone, NOS” The singular forms of these suggested SNOMED terms were not used in the textbook.

Word-sense appropriateness was not as critical for vocabulary discovery as it would be for automated text analysis, as inappropriate matches with manual review often resulted in the discovery of relevant concepts. In text analysis, the software would need to discern in each case the specific meaning of a character string for which multiple meanings existed. When word-sense appropriateness was ignored, an additional 11 veterinary medical terms and 5 canine physical examination terms were identified using SNOMED alone and an additional 38 veterinary medical concepts and 27 canine physical examination concepts were identified using the Metathesaurus-enhanced technique.

Metathesaurus-enhanced lexical matching does not have to be used as a complementary technique. One could use Metathesaurus-enhanced lexical matching without comparing matches against a list of previously matched SNOMED terms from a stand-alone version of SNOMED, providing the user is willing to accept the subset of SNOMED terms contained in the Metathesaurus that matched using lexical matching alone.

The textbook was an acceptable source of canine physical examination terms, though it was not as effective as a pre-filter of terms not related to the canine physical examination as we had hoped. The textbook often contained explanations that were tangential to the canine physical examination such as those referring to embryology, physiology and diagnoses. A more focused text source would have yielded a more specific list. In addition, including the index as a text source was not a good idea, as it contained non-canine related terms such as “wing” and “gizzard.”

LIMITATIONS

Metathesaurus-enhanced lexical matching was critically dependent on the accuracy of the links among synonyms in the Metathesaurus. Errors in these links can cause erroneous SNOMED concepts to be suggested even though the matching non-SNOMED term was word-sense appropriate. Fourteen of the matching non-SNOMED terms considered word-sense appropriate, were linked to SNOMED concepts that were not concept-sense appropriate. Thus, the suggested concept did not match the author's intent, even though the matching non-SNOMED term did.

These inappropriate links fell into one of three categories, 1) a term mapped to a narrower SNOMED concept, 2) a term mapped to a broader SNOMED concept or 3) an erroneous mapping. Examples of Metathesaurus terms mapped to narrower SNOMED concepts were "Aortic stenosis" mapped to "Aortic valve stenosis, NOS," "Pyloric Stenosis" mapped to "Acquired obstruction of pylorus," and "Retinal Arteries" mapped to "Central retinal artery." An example of a Metathesaurus term mapped to a broader SNOMED concept was "Cervical Vertebrae" mapped to "Cervical spinal column." This error in mapping also occurs within SNOMED, as "Cervical vertebra" and "Cervical spinal column" are considered synonymous. Mappings judged to be erroneous included, "Proteinuria" mapped to "Protein measurement, urine," "Malignant tumor" mapped to "Epithelioma, malignant," and "Vaginal Discharge" mapped to "Vaginal secretions, NOS."

During data analysis, it was observed that a single Metathesaurus concept was occasionally mapped to two distinct SNOMED concepts. One example was the Metathesaurus concept

“Scratching” C0311213 (the activity), which was mapped to the SNOMED concepts, “Scratch” M-14700 the injury, and “Scratches” D0-10163 the disease. The term “Scratching” is not contained in SNOMED. Another example was the Metathesaurus concept “Retinal Arteries” C0035301 mapped to the SNOMED concepts, “Retinal artery, NOS” T-AA380 and “Central Retinal Artery” T-45430. In both instances, SNOMED considers the concepts as distinct, while the Metathesaurus does not. Thus, these mappings indicate explicit synonym disagreement between the Metathesaurus and SNOMED.

As mentioned previously, character strings in SNOMED and the Metathesaurus for which word-sense ambiguity was known to occur were excluded from this research. However, additional ambiguous character strings such as “litter,” relating to a group of puppies and not a device used to carry a non-ambulatory patient, were discovered as a result of manual review.

Searches were limited to terms in the Metathesaurus and SNOMED containing 6 words or less since pilot studies using terms containing 6 words resulted in 0 matches. It was assumed that few significant terms would contain more than 6 words. In addition, we believe that any matching string over 6 words would likely be a pre-coordinated finding of the canine physical examination and would be better represented by the term's individual components. In fact, the only 5 word matching term identified in this study was “Measurement of central venous pressure.”

Despite the fact that UMLS and SNOMED use case to convey meaning, we chose to use case-insensitive matching, as many text documents do not enforce capitalization consistently. Some text in the textbook was in uppercase, and some abbreviations were in lowercase. Unfortunately, this decision led to some word-sense inappropriate matches such as, “ACT” (Activated Clotting Time) in the statement, “The act of regurgitation …,” and “ARC” (Aides-related complex) in the statement, “Probably the reflex arc…”

Ocassionally a term was suggested by the Metathesaurus-enhanced technique that was discovered previously using lexical matching and SNOMED alone. Closer examination of these terms revealed a discrepancy between the termcode of the term as reported in the Metathesaurus and SNOMED version 3.3. This problem could have been addressed a priori by comparing the character strings and termcodes of SNOMED version 3.3 terms with the SNOMED version 3.1 terms contained in the Metathesaurus.

Use of electronic sources of free-text for processing is highly recommended, as scanning the textbook introduced errors in the text document. Although we sought permission from several publishers, we could not obtain an electronic version of a suitable text.

The Metathesaurus-enhanced lexical matching technique has no mechanism for suggesting the veterinary-preferred term for the suggested concept. However, as none of the existing SNOMED synonyms resulted in a lexical match, consideration should be given to adding the matching non-SNOMED term to SNOMED. However, a primary benefit of the Unified Medical Language System is the potential to eliminate the need for a “terms race” among competing vocabularies. The National Library of Medicine is sponsoring studies to evaluate the combined coverage of multiple vocabularies [23,24]. The benefit of this NLM research is obvious for human medicine, but its impact on veterinary medicine is not yet established.

CONCLUSIONS

Redundancy in a standardized controlled medical vocabulary is undesirable if one is trying to standardize or control what can be said.[25] However, if one is using lexical matching to recognize what has been said in free-text, redundancy in the form of a synonym list is desirable. [5,18] This research demonstrated that the combined lexical strength of the Metathesaurus source vocabularies can be leveraged to identify concepts from an individual source vocabulary contained in free-text. Vocabularies with a high percentage of terms lacking literary flow such as, "Vomiting, projectile" will benefit most from this technique. In addition, manual review of lexically matched terms in context, is useful for identifying and cataloging ambiguous character strings for use by text processing applications. Finally, this research discovered instances of explicit synonym disagreement between SNOMED and the Metathesaurus.

FUTURE WORK

Metathesaurus-enhanced lexical matching needs to be evaluated more fully when applied to additional free-text sources, both formal and clinical. In addition, methods for compiling results over multiple free-text sources need to be developed and evaluated.

Finally, explicit synonym disagreement between the Metathesaurus and its source vocabularies needs to be further characterized and evaluated.

Table 3.1 Results from lexical matching using SNOMED alone

	Lexical matching using SNOMED alone
Matched terms	2212
Matched concepts	1802
Sample size	331
Word-sense appropriate	312/331 (94%)
Veterinary medical related	308/312 (99%)
Physical examination related	193/312 (62%)

Table 3.2 Results from Metathesaurus-enhanced lexical matching

	Metathesaurus-enhanced lexical matching
Suggested SNOMED concepts	208
Sample size	136
Word-sense appropriateness of non-SNOMED matching term	111/136 (82%)
Concept appropriateness of suggested SNOMED concept	97/111 (87%)
Veterinary medical related	97/97 (100%)
Physical examination related	57/97 (59%)

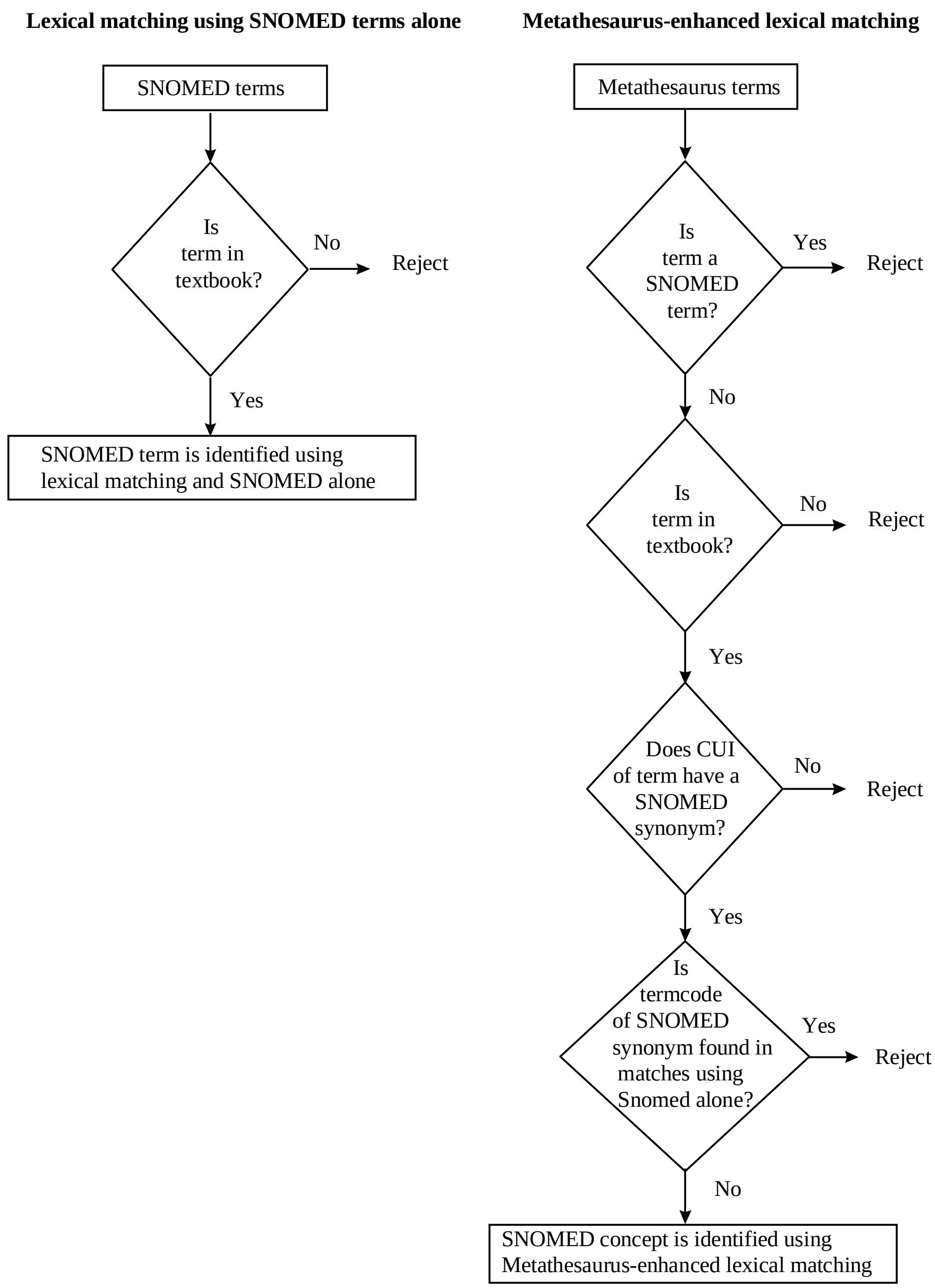


Figure 3.1 Flow diagrams of lexical matching using SNOMED terms alone and Metathesaurus-enhanced lexical matching.

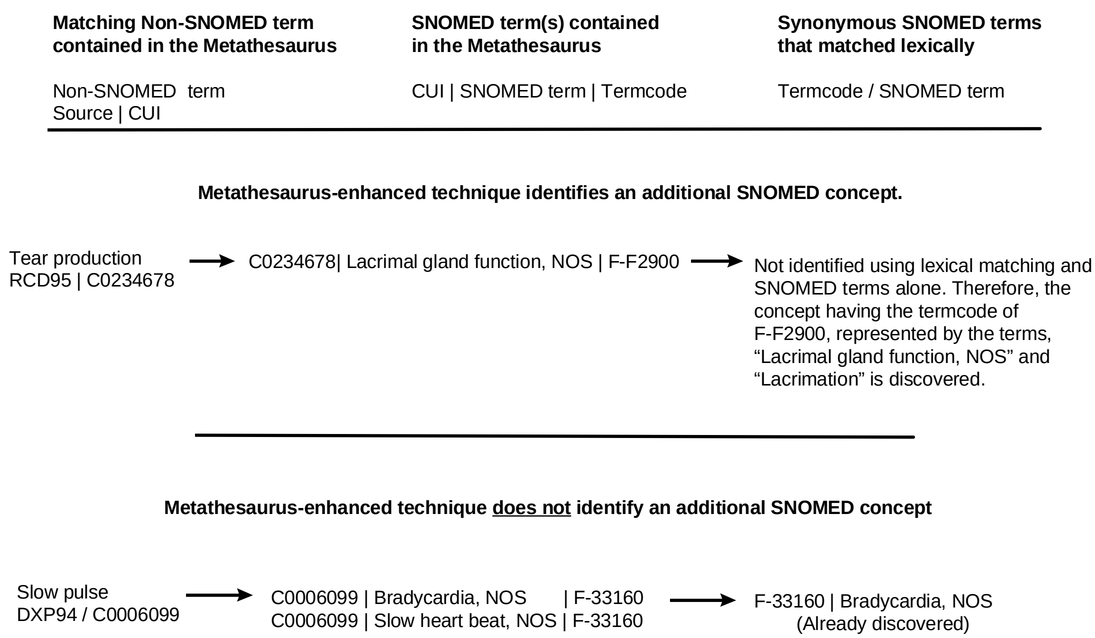


Figure 3.2 Scenarios in which SNOMED concepts were considered discovered or not discovered using Metathesaurus-enhanced lexical matching.

REFERENCES

1. Moorman PW, van Ginneken AM, van der Lei J, et al. A model for structured data entry based on explicit descriptional knowledge. *Methods Inf Med* 1994;33:454-63.

2. Kannry JL, Wright L, Shifman M, et al. Portability issues for a structured clinical vocabulary: mapping from Yale to the Columbia medical entities dictionary. *J Am Med Inform Assoc* 1996;3:66-78.

3. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. *Methods Inf Med* 1993;32:281-91.

4. National Library of Medicine. *Documentation, 8th Edition UMLS Knowledge Sources*. Bethesda, MD, 1997.

5. Sherertz DD, Tuttle MS, Blois MS, et al. Intervocabulary Mapping Within the UMLS: The Role of Lexical Matching. *Proc Annu Symp Comput Appl Med Care* 1988:201-6.

6. Rindflesch TC, Aronson AR. Semantic processing in information retrieval. *Proc Annu Symp Comput Appl Med Care* 1993:611-5.

7. Rindflesch TC, Aronson AR. Ambiguity resolution while mapping free text to the UMLS Metathesaurus. *Proc Annu Symp Comput Appl Med Care* 1994:240-4.

8. Hersh WR, Campbell EH, Evans DA, et al. Empirical, Automated Vocabulary Discovery Using Large Text Corpora and Advanced Natural Language Processing Tools. *Proc Annu Symp Comput Appl Med Care* 1996:159-63.

9. Spackman KA, Hersh WR. Recognizing Noun Phrases in Medical Discharge Summaries: An Evaluation of Two Natural Language Parsers. *Proc Annu Symp Comput Appl Med Care* 1996:155-58.

10. Henry SB, Campbell KE, Holzemer WL. Representation of nursing terms for the description of patient problems using SNOMED III. *Proc Annu Symp Comput Appl Med Care* 1993:700-4.

11. Henry SB, Holzemer WL, Reilly CA, et al. Terms used by nurses to describe patient problems: can SNOMED III represent nursing concepts in the patient record? *J Am Med Inform Assoc* 1994;1:61-74.

12. Lange LL. Representation of Everyday Clinical Nursing Language in UMLS and SNOMED. *Proc Annu Symp Comput Appl Med Care* 1996:140-44.

13. Henry SB, Holzemer WL. Can SNOMED International represent patients' perceptions of health-related problems for the computer-based patient record? *Proc Annu Symp Comput Appl Med Care* 1994:184-7.

14. Klimczak JC, Hahn AW, Sievert ME, et al. Comparing clinical vocabularies using coding system fidelity. *Proc Annu Symp Comput Appl Med Care* 1995:883-7.

15. Campbell JR, Payne TH. A comparison of four schemes for codification of problem lists. *Proc Annu Symp Comput Appl Med Care* 1994:201-5.

16. Chute CG, Cohn SP, Campbell KE, et al. The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures. *J Am Med Inform Assoc* 1996;3:224-33.

17. Rijnberk A, de Vries HW. Medical History and Physical Examination in Companion Animals. Dordrecht,, The Netherlands: Kluwer Academic Publishers, 1995;376.

18. Sherertz DD, Tuttle MS, Olson NE, et al. Lexical Mapping in the UMLS Metathesaurus. *Proc Annu Symp Comput Appl Med Care* 1989:494-9.

19. Fraser CM. The Merck Veterinary Manual. Seventh Edition ed. Rahway, NJ: Merck & Co., Inc., 1991.

20. Tersmetta KWF, Scott AF, Moore WG, et al. Barrier Word Method for Detecting Molecular Biology Multiple Word Terms. *Proc Annu Symp Comput Appl Med Care* 1988:207-11.

21. Moore GW, Berman JJ. Automatic SNOMED coding. *Proc Annu Symp Comput Appl Med Care* 1994:225-9.

22. Isaac S, Michael WB. *Handbook in Research and Evaluation*. Third ed. San Diego, Ca: EdITS, 1995.

23. Humphreys BL, Hole WT, McCray AT, et al. Planned NLM/AHCPR large-scale vocabulary test: using UMLS technology to determine the extent to which controlled vocabularies cover terminology needed for health care and public health. *J Am Med Inform Assoc* 1996;3:281-7.

24. Humphreys BL, McCray AT, Cheh ML. Evaluating the Coverage of Controlled Health Data Terminologies: Report on the Results of the NLM/AHCPR Large Scale Vocabulary Test. *J Am Med Inform Assoc* 1997;4:484-500.

25. Cimino JJ. Auditing the Unified Medical Language System with Semantic Methods. *J Am Med Inform Assoc* 1998;5:41-51.

CHAPTER 4

Explicit Synonym Disagreement between the Metathesaurus of the Unified Medical Language System and its Source Vocabularies

ABSTRACT

Explicit synonym disagreement between the Metathesaurus and its sources vocabularies occurs when, 1) terms within a single concept group in a source vocabulary are mapped to multiple Metathesaurus concepts, and 2) terms from multiple concept groups in a source vocabulary are mapped to a single Metathesaurus concept. Explicit synonym disagreement is examined using two source vocabularies, the Systematized Nomenclature of Medicine and Veterinary Medicine (SNOMED) and the Read Thesaurus. One thousand eight hundred and fifty-eight of 99,802 (1.9%) SNOMED concept groups contained terms mapped among multiple Metathesaurus concepts. One thousand thirty-one of 47,640 (2.2%) Read Thesaurus concept groups contained terms mapped among multiple Metathesaurus concepts. One thousand twenty-five of 331,756 (.3%) Metathesaurus concepts were mapped to terms from multiple SNOMED concept groups and 3044 (.9%) were mapped to terms from multiple Read Thesaurus concept groups.

Five causes of explicit synonym disagreement between a source vocabulary and the Metathesaurus were identified in this research, 1) errors within a source vocabulary, 2) errors within the Metathesaurus, 3) errors in mapping between the Metathesaurus and a source vocabulary, 4) systematic differences in vocabulary management between the Metathesaurus and a source vocabulary, and 5) differences regarding synonymy among domain experts based on perspective or context. Explicit synonym disagreement due to errors within source vocabularies, errors within the Metathesaurus or errors in mapping between the Metathesaurus and its source vocabularies will be identified with use, and corrected over time. Systematic differences in vocabulary management can be documented and anticipated as long as the differences remain

consistent. Reconciling differences among domain experts is more problematic. One approach is to simply document which terms are involved. Another approach is to provide a mechanism for selecting either vocabulary-based synonymy using vocabulary-based concept group identifiers or Metathesaurus-based synonymy using the unique concept identifier (CUI). Finally, a "basis of synonymy" attribute is proposed that would provide a mechanism for identifying the perspective or context of synonymy explicitly.

INTRODUCTION

Accurate communication begins with common agreement on exactly which concepts are being discussed. This agreement is made possible through an established set of terms which represent the concepts of discourse. A term is a language-based label, for a thought-based concept[1]. As an example, the term "dog" generates a mental image or concept of a prototypical animal, which differs from the mental image generated by the term, "puppy." Significance of these differences is determined by the degree to which these differences impact decisions and actions.

Unfortunately, agreement on exactly which concepts are being discussed is not achieved easily, as many concepts can be represented by multiple terms, and a term can have multiple meanings depending on context. As an example, in one controlled medical vocabulary the terms "Canine species," "Dog," "Bitch," and "Canine, NOS" are considered synonyms. However, "Canine species" actually refers to all types of canines including the wolf, fox, coyote, and Canis lupus familiaris (domestic dog). The term "Dog" usually refers to the domestic dog, but depending on context can also refer to only male members of various canine species (e.g., wolves, the domestic dog etc.) In the same manner, "Bitch" usually refers to the female domestic dog, but can also refer to only female members of various canine species. And finally, the term "Canine" is an adjective, not a noun. These distinctions are significant in only a few circumstances and are beyond pedantic in most. However, when handling terms and their related concepts, the computer is unequaled in its pedanticism.

The aggregation of patient data across multiple institutions or even within the same institution over an extended period of time is difficult when multiple terms for a given concept exist. This problem is exacerbated even further if no consistent standards for data entry are enforced. When performing a retrospective study across multiple institutions on the incidence of a particular disease, one would need to search on all possible synonyms used within each institution during the period of interest. Therefore, searching under these circumstances requires a mechanism for managing concepts and their synonyms.

Identification and management of concepts and their synonyms can be accomplished with a standardized controlled medical vocabulary that uses unique concept codes. Each unique concept code identifies an individual concept and is assigned to all terms that are synonyms for the concept. Thus, synonymous terms are linked through a shared concept code and are referred to in this chapter as a concept group.

Concept codes are used directly and indirectly when forming search strategies. If the resource to be searched contains concept codes explicitly, or the concept codes have been added through prior processing, the search strategy can employ the concept code directly. On the other hand, if the resource contains terms in the form of free-text or specialized lists, concept codes can be used to identify synonymous terms that may have been used in the resource to represent the concept of interest.

When using a single controlled medical vocabulary as a source of synonymous terms for searching, a potential problem exists in that the search results are dependent upon the vocabulary's breadth of domain coverage, and the quantity and quality of links among synonymous terms. A search is further complicated if the target documents were constructed or indexed using the synonymy of vocabularies other than the searching vocabulary.

Synonymy among multiple vocabularies

Concept codes are effective for addressing synonymy within a single vocabulary, but synonymy among terms from multiple vocabularies is more problematic. One approach is to link synonymous concept codes from multiple vocabularies. Since each concept code represents a group of synonymous terms from an individual vocabulary, linking synonymous concept codes from multiple vocabularies should result in the combination of all synonymous terms from each vocabulary. A problem with this approach is that vocabularies differ regarding which terms are synonymous, and therefore, one vocabulary may group four terms in a single concept group, while another may distribute the same four terms over multiple groups.

Another approach to establishing synonymy among multiple vocabularies is to create a new authority, with its own set of concept codes, and link synonymous terms from multiple vocabularies to the new set of concept codes individually. This is the approach taken by the Unified Medical Language System (UMLS). [2,3]

Unified Medical Language System (UMLS)

The UMLS, established in 1986 by the National Library of Medicine (NLM), was designed to "facilitate the retrieval and integration of information from multiple machine-readable biomedical information sources. The sources of interest include: descriptions of the biomedical literature, clinical records, factual databanks, knowledge-based systems, and directories of people and organizations." [4] The UMLS is organized into four Knowledge Sources which are the Metathesaurus, Semantic Network, Information Sources Map, and SPECIALIST Lexicon. [4] Only the Metathesaurus was used in this research.

The 1997 version of the Metathesaurus contains approximately 330,000 unique concepts represented by approximately 740,000 terms from over 40 autonomously controlled medical vocabularies known as, source vocabularies. Two source vocabularies were used in this research, the Systematized Nomenclature of Medicine and Veterinary Medicine (SNOMED), and the Read Thesaurus.

Each source vocabulary contains unique terms in addition to terms held in common with other vocabularies. A set of unique concepts along with the preferred lexical form of each (a specific term) has been identified from among the terms contained in the source vocabularies. Each concept is assigned a unique concept code referred to in the Metathesaurus as a unique concept identifier (CUI). The CUI is an eight place character string beginning with the letter C followed by 7 numbers.

The CUI is used to link synonymous terms from among the individual source vocabularies. Terms from a source vocabulary are assigned the CUI of the Metathesaurus concept the term represents. As an example, the term "Anterior Chamber" from one source vocabulary, is designated by the UMLS as the preferred form of the concept and is assigned the CUI of C0003151. The terms "anterior chamber," "Anterior Chambers," "Anterior chamber of eye," and "Anterior chamber (EYE) contained in other source vocabularies are considered synonyms by the Metathesaurus and therefore are assigned the same CUI, C0003151.

Systematized Nomenclature of Medicine and Veterinary Medicine (SNOMED)

SNOMED version 3.3 contains approximately 100,000 unique concepts represented by over 144,000 terms used in human and veterinary medicine. Each concept in SNOMED is assigned an alphanumeric code referred to as a termcode. Terms that represent a given concept are assigned the concept's termcode. Thus, the terms "Canine species," "Dog," "Bitch," and "Canine, NOS," are considered synonyms for the same concept and therefore share the same termcode, L-80700.

A term representing the preferred form of a concept is designated from among a group of synonymous terms. In this case, “Canine species” is given that designation.

Read Thesaurus

The Read Thesaurus contains approximately 220,000 terms used in human medicine. The Read Thesaurus is used widely in the United Kingdom and is maintained by the National Health Service of the United Kingdom. Each term in the Read Thesaurus is assigned two codes, a Read code and a Term ID. The Read code is a concept code and represents a unique concept in the same manner as the termcode in SNOMED and the CUI in the Metathesaurus. The Term ID is a unique identifier for each term. As an example, the Read Thesaurus contains the term, “Dog” having the Read code, “X79or” and the Term ID, “YM7Da”. No other synonyms for the term “Dog” are contained in the Read Thesaurus. However, if one was added, it would be assigned the Read code “X79or” and a unique Term ID.

Explicit synonym disagreement

Terms contained in the Metathesaurus for which synonymy is addressed by their source vocabulary are subject to two sources of synonymy, 1) vocabulary-based synonymy and 2) Metathesaurus-based synonymy. Vocabulary-based synonymy is established among terms within a source vocabulary using a vocabulary-based concept code such as the termcode in SNOMED or the Read code in the Read Thesaurus. Metathesaurus-based synonymy is established among terms from the source vocabularies using the CUI, which is assigned independent of the vocabulary-based assignment.

If there was complete agreement between a source vocabulary and the Metathesaurus regarding the synonymy of terms, all synonyms in a source vocabulary would share the same vocabulary-based concept code and the same CUI. Conversely, all terms considered not to be synonymous in a source vocabulary would have differing vocabulary-based concept codes and differing CUIs.

Unfortunately, this is not the case. Returning to the first example, in SNOMED the terms “Canine species,” “Dog,” “Bitch,” and “Canine, NOS” are considered synonymous as each term is assigned the same SNOMED termcode, L-80700. However, in the Metathesaurus only the terms “Canine species,” “Canine, NOS,” and “Dog” are considered synonymous as each term is assigned the same CUI, C0012984. The term “Bitch,” is considered by the Metathesaurus to be a separate concept and thus is assigned a different CUI, C324294. This disagreement in synonymy between the Metathesaurus and a source vocabulary is referred to here as, explicit synonym disagreement.

Explicit synonym disagreement should not be unanticipated. The Metathesaurus and each of its source vocabularies, reflect the biases of its designers and domain experts. In addition, synonymy is often based on a particular perspective or a limited set of attributes.[5] However, if one is to depend on the Metathesaurus for mappings of synonymous terms among source vocabularies, terms involved in explicit synonym disagreement must be documented.

This research describes a method for identifying concept groups in a source vocabulary for which member terms are subject to explicit synonym disagreement. Explicit synonym disagreement was identified if, 1) terms within a single concept group in a source vocabulary where mapped among multiple Metathesaurus concepts or 2) terms from multiple concept groups in a source vocabulary were mapped to a single Metathesaurus concept. This method uses only resources available in the UMLS Metathesaurus and can be applied to each version of the Metathesaurus independently. This method can be used with any source vocabulary of the Metathesaurus in which synonymy is addressed explicitly.

Five causes of explicit synonym disagreement were identified using this approach, 1) errors within a source vocabulary, 2) errors within the Metathesaurus, 3) errors in mapping between the Metathesaurus and a source vocabulary, 4) systematic differences in vocabulary management between the Metathesaurus and a source vocabulary, and 5) differences regarding synonymy among domain experts based on perspective or context.

Implications of explicit synonym disagreement are discussed and a proposal is made for an attribute to identify the "basis of synonymy" for a given concept group explicitly. The concept group and its "basis of synonymy" attribute are likened to a single level hierarchical parent/child relationship and its relationship attribute. The concept represents the parent, synonymous terms represent the children in the relationship and the basis attribute represents the relationship attribute which identifies the nature of the parent/child relationship.

METHODS

All English terms contained in the text files of the 1997 release of the UMLS Metathesaurus were imported into a relational database using Visual Foxpro™ and assigned a unique identifier (recnum). Table 4.1 illustrates fields from the database that were relevant to this research along with their definitions.

SNOMED Processing

Terms contained in the Metathesaurus were filtered such that only SNOMED terms version 3.1 (SNMI95) with a term type status of Designated Preferred name (PT), Designated Synonym (SY), or Adjective (AD) were eligible for processing. This filter excluded terms of the term type status, Hierarchical term (HT) which are used for classification purposes. The term "other" was excluded as it is mapped to 51 unique Metathesaurus concepts.

Identifying SNOMED concept groups whose members are mapped among multiple Metathesaurus concepts.

The set of SNOMED terms was sorted on **termcode** (contained in the Scd field) into SNOMED-based concept groups, meaning all terms with the same termcode were grouped together. (see Table 4.2, column labeled "Termcode") Terms were processed sequentially, beginning with the first. Three parameters were monitored for each term. First, the **termcode** of the first term in a concept group was stored in the termcode memory variable. Second, the **CUI** of the first term in a concept group was stored in the CUI memory variable. Third, the **recnum** of the first term in a concept group was stored in the recnum memory variable. Thus, when the first record in Table

4.2 was processed, the termcode memory variable contained L-80700, the CUI memory variable contained C0012984, and the recnum memory variable contained 51979.

As the SNOMED terms were processed sequentially, the termcode of each new term was compared to the termcode of the first term in the current concept group (stored in the termcode memory variable). If the termcode of the new term and the termcode stored in the termcode memory variable were the same, the new term belonged to the current SNOMED concept group. Otherwise, the new term belonged to the next concept group.

If the new term belonged to the current SNOMED concept group, the CUI of the new term was compared to the string of unique CUIs contained in the CUI memory variable from the previously processed terms in the concept group. If the CUI of the new term was unique from the CUIs contained in the CUI memory variable, as with the term “Bitch/C0324294” and the previously processed term “Canine, NOS/C0012984,” the CUI was appended to the CUI memory variable and the recnum was appended to the recnum memory variable. Thus, when the second record in Table 4.2 was processed, the termcode memory variable contained L-80700, the CUI memory variable contained C0012984|C0324294, and the recnum memory variable contained 51979|599552. The last two terms in the L-80700 concept group, “Canine species/C0012984” and “Dog/C0012984” would have no effect on the memory variables, as neither term is associated with an undiscovered CUI for the concept group.

When a new term did <u>not</u> belong to the current concept group, such as would be the case when termcode L-80A00 for "Feline species" was encountered, the current concept group was evaluated for multiple CUIs. If the Metathesaurus had assigned multiple Metathesaurus-based concepts (indicated by multiple CUIs), among terms in a single SNOMED-based concept group, as was the case with the L-80700 concept group, the termcode, string of CUIs contained in the CUI memory variable, and string of recnums contained in the recnum memory variable were copied to a results table. The L-80A00 concept group represented by the terms "Feline species," "Feline, NOS" and "Cat" would not result in an entry in the results table, as all three terms are associated with the same CUI, C0007450. After evaluating the current concept group for multiple CUIs, the CUI, termcode and recnum memory variables were reinitialized to contain the termcode, CUI and recnum of the first term in the new concept group.

Identifying Metathesaurus concepts mapped to terms from multiple SNOMED concept groups.

The SNOMED terms were sorted on **CUI** into Metathesaurus-based concept groups, meaning all SNOMED terms with the same CUI were grouped together (see Table 4.3, column labeled CUI). Terms were processed sequentially, beginning with the first. Three parameters were monitored for each term. First, the **CUI** of the first term in a concept group was stored in the CUI memory variable. Second, the **termcode** of the first term in a concept group was stored in the termcode memory variable. Third, the **recnum** of the first term in a concept group was stored in the recnum memory variable. Thus, when the first record in Table 4.3 was processed, the CUI memory variable contained C0015967, the termcode memory variable contained F-03003, and the recnum memory variable contained 65856.

As the SNOMED terms were processed sequentially, the CUI of each new term was compared to the CUI of the first term in the current concept group (stored in the CUI memory variable). If the CUI of the new term and the CUI stored in the CUI memory variable were the same, the new term belonged to the current Metathesaurus concept group. Otherwise, the new term belonged to the next concept group.

If the new term belonged to the current Metathesaurus concept group, the termcode of the new term was compared to the string of unique termcodes contained in the termcode memory variable from the previously processed terms in the concept group. If the termcode of the new term was unique from the termcodes contained in the termcode memory variable, as with the term “Hyperthermia/F-0A440” and the previously processed term “Fever/F-03003,” the termcode was appended to the termcode memory variable and the recnum was appended to the recnum memory variable. Thus, when the second record in Table 4.3 was processed, the CUI memory variable contained C0015967, the termcode memory variable contained F-03003|F-0A440, and the recnum memory variable contained 65856|65874. The last term in the C0015967 concept group, “Fever 106 degrees F or higher” would also be added to the termcode and recnum memory variables, as its termcode is also unique for the concept group. Thus, when the third record in Table 4.3 was processed, the CUI memory variable contained C0015967, the termcode memory variable contained F-03003|F-0A440|F-03009, and the recnum memory variable contained 65856|65874|65889.

When a new term did <u>not</u> belong to the current concept group, such as would be the case when CUI C0015970 for “Fever of unknown origin” was encountered, the current concept group was evaluated for multiple termcodes. If the Metathesaurus had assigned a single Metathesaurus-based concept to terms from multiple SNOMED-based concept groups (indicated by multiple termcodes), as was the case with the concept represented by C0015967, the CUI, string of termcodes contained in the termcode memory variable, and string of recnums contained in the recnum memory variable were copied to a results table. The C0015970 concept group represented by the terms “Fever of unknown origin” and “F.U.O.” would not result in an entry in the results table, as both terms are associated with a single termcode, F-03010. After evaluating the current concept group for multiple termcodes, the CUI, termcode and recnum memory variables were reinitialized to contain the CUI, termcode and recnum of the first term in the new concept group.

Read Thesaurus Processing

Terms contained in the Metathesaurus were filtered such that all Read Thesaurus terms version 3.1 (RCD95) were included except terms with a term status of Obsolete abbreviation (OA) or Obsolete preferred term (OP). Terms from the Read Thesaurus were processed in the same manner as SNOMED terms (Read code was used in place of the SNOMED termcode).

RESULTS

Results are reported according to concept groups. Even though terms were processed individually, and corrections must be made on a term-by-term basis, all terms in a concept group for which explicit synonym disagreement exists must be reviewed individually, as there is no way to know which terms are in error; only that differences exist.

SNOMED concept groups having terms mapped among multiple Metathesaurus concepts

1858 of 99,802 (1.9%) SNOMED concept groups contained terms that were mapped among multiple Metathesaurus concepts. Terms from 1858 SNOMED concept groups were mapped to a total of 4145 Metathesaurus concepts for an average of 2.2 unique Metathesaurus concepts per affected SNOMED concept group.

Metathesaurus concepts mapped to terms from multiple SNOMED concept groups

1025 of 331,756 (.3%) Metathesaurus concepts were mapped to terms from multiple SNOMED concept groups. 1025 Metathesaurus concepts were mapped to terms from a total of 2128 SNOMED concept groups for an average of 2.1 unique SNOMED concepts per affected Metathesaurus concept group.

Read Thesaurus concept groups having terms mapped among multiple Metathesaurus concepts

1031 of 47640 (2.2%) Read Thesaurus concept groups contained terms that were mapped among multiple Metathesaurus concepts. Terms from 1031 Read Thesaurus concept groups were mapped to a total of 2350 Metathesaurus concepts for an average of 2.3 unique Metathesaurus concepts per affected Read Thesaurus concept group.

Metathesaurus concepts mapped to terms from multiple Read Thesaurus concept groups

3044 of 331,756 (.9%) Metathesaurus concepts were mapped to terms from multiple Read Thesaurus concept groups. 3044 Metathesaurus concepts were mapped to terms from a total of 7043 Read Thesaurus concept groups for an average of 2.3 unique Read Thesaurus concepts per affected Metathesaurus concept group.

DISCUSSION

Explicit synonym disagreement

Explicit synonym disagreement between a source vocabulary and the Metathesaurus occurs, 1) when terms within a single concept group in a source vocabulary are mapped to multiple Metathesaurus concepts, and 2) when terms from multiple concept groups in a source vocabulary are mapped to a single Metathesaurus concept.

The first scenario is illustrated in Table 4.4. Examination of the middle column reveals 5 terms from the Read Thesaurus concept group, R056z, mapped among three unique Metathesaurus concepts (right column - C0001416, C0024208, C0362040). Thus, the Read Thesaurus considers all five terms to be synonyms, but the Metathesaurus does not.

The second scenario is illustrated in Table 4.5. Examination of the middle column reveals 10 terms from six unique Read Thesaurus concept groups (right-most column - Xa0Vr, 1692., XM07t, R056., R056z, and R0561) mapped to a single Metathesaurus concept, C0024208. Thus, the Read Thesaurus considers the 10 terms as members of 6 unique concept groups, while the Metathesaurus does not.

Five causes of explicit synonym disagreement

Five causes of explicit synonym disagreement between a source vocabulary and the Metathesaurus were identified in this research, 1) errors within a source vocabulary, 2) errors within the Metathesaurus, 3) errors in mapping between the Metathesaurus and a source vocabulary, 4) systematic differences in vocabulary management between the Metathesaurus and a source vocabulary, and 5) differences regarding synonymy among domain experts based on perspective or context.

Errors within a source vocabulary

Errors within a source vocabulary may be indicated when terms belonging to the same concept group in a source vocabulary are mapped to different Metathesaurus concepts. As an example, SNOMED considers the terms “Polydactyly, NOS” and “Hexadactyly” synonymous, but the Metathesaurus does not since “many” is not always “six.” An example of a more subtle distinction, SNOMED considers the terms, “Pleurisy, NOS” and “Pleurisy without effusion” to be synonyms while the Metathesaurus considers them unique. One could argue that adding the phrase “without effusion” was a form of specification and therefore the term is differentiated from the prior term which was designated “NOS” meaning “Not Otherwise Specified.”

Errors within the Metathesaurus

Errors within the Metathesaurus may be indicated when terms from a single concept group in a source vocabulary are mapped to multiple Metathesaurus concepts. As an example, three terms from the SNOMED concept group, T-A8780, “Accessory nerve, NOS,” “Spinal accessory nerve,” and “Eleventh cranial nerve,” are distributed over two apparently duplicate Metathesaurus concepts, “Accessory Nerve” (C0000905) and “Spinal Accessory Nerve” (C0282035), both of which refer to the eleventh cranial nerve.

Errors within the Metathesaurus may also be indicated when terms from multiple concept groups in a source vocabulary are mapped to a single Metathesaurus concept. As an example, the terms “Arm” (T-D8200) and “Upper extremity, NOS” (T-D8000), are members of two different SNOMED concept groups, but are mapped to the same Metathesaurus concept, “Arm” C0003792. It is incorrect to combine these two terms as, “Arm” refers only to the section of the “Upper extremity” from the shoulder joint to the elbow[6].

Errors in mapping between the Metathesaurus and a source vocabulary

Errors in mapping between the Metathesaurus and a source vocabulary may be indicated when terms from multiple concept groups in a source vocabulary are mapped to a single Metathesaurus concept. As an example, three SNOMED terms from unique concept groups, "COLD" (D2-60000) which is an abbreviation for "Chronic Obstructive Lung Disease", "Cold" (F-03230) which refers to temperature, and "Common cold" (DE-35210) are mapped to the same Metathesaurus concept, "Common cold" (C0009443). SNOMED recognizes these terms as distinct, while the Metathesaurus identifies them incorrectly as synonyms. A similar situation exists with the Read Thesaurus in that the Metathesaurus incorrectly maps the Read Thesaurus term "Cold" (XC06u), referring to temperature, to the Metathesaurus concept, "Common cold" (C0009443).

Errors in mapping between the Metathesaurus and a source vocabulary may be indicated when terms from a single concept group in a source vocabulary are mapped to multiple Metathesaurus concepts. As an example, in Table 4.6 the term "High blood pressure" from the SNOMED concept group D3-02000 meaning "Hypertension" is mapped to the Metathesaurus concept, C0085907 meaning "Elevated blood pressure," while other members of the concept group D3-02000 are mapped to the Metathesaurus concept, C0020538 meaning "Hypertension." This error occurs because the term, "High blood pressure" in SNOMED is a synonym for the disease state, while in the Metathesaurus, the term is mapped to the physiologic abnormality. Table 4.6 also illustrates the same error with the Read Thesaurus term, "BP-High blood pressure" (Read code

G2…). The intended meaning of a given term such as “Hypertension” can often be determined only by examination of synonymous terms in a vocabulary.

Mapping errors can even result from something as simple as overlooking the placement of a comma or an assumption regarding the plural form of a term. Examples include the terms “Deer mouse” and “Mouse, Deer” that are mapped to the Metathesaurus concept, “Mouse deer” (both mammals, but very different in size). And finally, the Read Thesaurus term “Crabs” (Read code AD22.) as in Phthiriasis pubis, is mapped to the Metathesaurus concept, “Crab” C0010260 the crustacean. These types of errors will be discovered and corrected through continued use of the Metathesaurus. In addition, errors will be discovered and corrected using lexical methods such as we employed and methods using semantics as described by Cimino.[7]

Systematic differences in vocabulary management

Systematic differences in vocabulary management between the Metathesaurus and a source vocabulary are consistent design and process issues which can be documented and anticipated. Both SNOMED and the Read Thesaurus exhibit these differences with the Metathesaurus.

SNOMED exhibits a systematic difference by assigning the same termcode to a noun and its adjectival form. The Metathesaurus considers these as distinct concepts. Examples of the noun/adjectival form disagreement are, Ingestion/Ingested F-50440, Larynx/Laryngeal T-24100 and Heart/Cardio-/Cardiac T-32000. There are 377 of these instances in the 1997 release of the Metathesaurus. It should be noted that within SNOMED, nouns and their adjectival forms are distinguished from one another by the “class” attribute, which contains an “05” for adjectives. The Metathesaurus designates adjectives with an “AD” in the TTY attribute field while also assigning each adjective a unique CUI.

The Read Thesaurus exhibits a systematic difference with the Metathesaurus regarding terms that have one of the following prefixes, “[D],” “[V],” “[X]” or “[M].” Many times these terms are duplicated in the Read Thesaurus with a term having an identical character string, excluding the prefix. The Read Thesaurus uses these prefixed terms for backward compatibility with previous Read versions to distinguish special circumstances for use of a particular character string. As an example, in accordance with ICD cross references, a term with a “[D]” prefix can be used to signify a symptom that is reported as a discharge diagnosis when a definitive diagnosis has not been made. Thus, “Cough” the symptom, Read code 171.. is differentiated from “[D]Cough” the diagnosis, R062.. The Metathesaurus often does not recognize this distinction, and in 134 instances, links both terms to the same Metathesaurus concept.

Similar special circumstance usage applies to terms with the prefix “[V],” “[X],” and “[M].” “[V]” is used to denote reasons for a patient encounter, “[X]” is used to denote new chapters for mental health, causes of accidents and a few pure classification terms and “[M]” is used to differentiate between a term as a clinical diagnosis or histologic finding. The number of Read concept groups in each category for which explicit synonym disagreements of this type exist are “[D]” 110, “[V]” 21, “[X]” 2 and “[M]” 1.

The developers of the Read Thesaurus do not recommend the use of these bracketed terms by new users of version 3 unless the previously mentioned functionality of backward compatibility or ICD cross referencing is required. However, these terms are contained in the Metathesaurus and existing medical documents, so their involvement in explicit synonym disagreement must be documented.

Differences among domain experts

Differences among domain experts refers to debatable differences of opinion regarding synonymy of terms. These differences are often based on an individual's perspective or what Sherertz refers to as, "degrees of similarity between concepts." As an example, Campbell cites that "Aspirin," "Ecotrin," and "Aspergum" are considered equivalent by the Metathesaurus and the reason is presumably because they share the same active ingredient.[5] A pharmacist, however, would obviously distinguish among USP tablets, an enteric coated trade name drug and a chewing gum.

Context can have a significant impact on synonymy. A "wrist" and "carpus" are considered synonyms when referring to a human being or a primate. However, a dog has a "carpus," not a "wrist." Another example involves the terms "knee" and "stifle." Both terms are synonyms though "knee" is preferred in human medicine, while "stifle" is preferred in veterinary medicine. To add more confusion, the term "knee" can refer to the carpus in the front limb of the horse.

IMPLICATIONS

The implications of explicit synonym disagreement vary depending on the particular use of the Metathesaurus-based concept groups. When searching free-text using lexical matching of terms, if one is using Metathesaurus-based concepts to identify a set of synonymous terms for a given concept, then the set of terms the Metathesaurus considers synonymous is critical. As an example, the Metathesaurus contains 16 lexical variations for the concept “Lymphadenopathy” CUI C0024208 (case insensitive with clerical notations removed). However, as shown in Table 4.5, 10 Read Thesaurus terms considered to be synonyms for “Lymphadenopathy” by the Metathesaurus, are distributed among 6 distinct Read concept groups. Thus, the results of a free-text lexical search can differ considerably depending upon whether one uses Metathesaurus-based synonymy or vocabulary-based synonymy.

If one is using Metathesaurus-based concept groups to identify synonymous concept groups among source vocabularies, explicit synonym disagreement can cause inconsistent results depending on which Metathesaurus-based concept group is used. As an example (see Table 4.6, first row, middle column), according to the Metathesaurus, the concept C0020538 “Hypertension” is synonymous with SNOMED terms having the termcode, D3-02000 and Read Thesaurus terms having the Read code, “G2…”. However, if one was interested in the concept C0085907 “Blood pressure, High,” and was unaware of the explicit synonym disagreement, it would appear that terms from both SNOMED concept groups D3-02000 and F-31003 are synonymous with the Read Thesaurus concept groups, G2… and XM02V.

SOLUTIONS

One might dispense with the issue of explicit synonym disagreement by choosing to use only Metathesaurus-based concepts and their mappings to identify synonymous terms among source vocabularies, thus ignoring the source vocabulary's determination of synonymy. A problem with this approach is that a set of documents constructed using assumptions regarding synonymy from a specific source vocabulary, may be inaccurately retrieved when searched using Metathesaurus-based synonymy. As an example, if one is searching for conditions involving female dogs, the Metathesaurus-based concept "Bitch" C0324294 indicates the equivalent SNOMED code is L-80700. However, documents that are coded according to SNOMED-based synonymy, use L-80700 for conditions affecting all dogs, including "Bitches."[‡]

One solution to the issue of explicit synonym disagreement is to simply document which terms are involved. Once these terms and their associated concept groups are documented, errors can be identified and corrected, or search engines can employ appropriate strategies to handle discrepancies. Documentation can be performed on each release of the UMLS individually.

A compromise to explicit synonym disagreement is to provide a mechanism for selecting either vocabulary-based or Metathesaurus-based synonymy. The Metathesaurus could store the vocabulary-based unique term identifier and the vocabulary-based concept group identifier with each term in the Metathesaurus. In this manner, one could utilize either the vocabulary-based synonymy by grouping terms according to the vocabulary-based concept unique identifier (eg.

termcode or Read code), or Metathesaurus-based synonymy by grouping terms according to the mappings between the CUIs and the source vocabulary's unique term identifier.[§]

Finally, an attribute could be used to identify the "basis of synonymy" for a given set of synonymous terms in a concept group. This approach is similar to that taken by the UMLS for managing hierarchical parent/child relationships. The concept is equivalent to the parent, synonymous terms are equivalent to the children, and the "basis of synonymy" attribute is equivalent to the parent/child relationship attribute. In essence, this synonymy unit is a single level, independent hierarchy having no continuity with, or inheritance from other members.

LIMITATIONS

The method described in this paper for identifying concept groups having terms involved in explicit synonym disagreement, applies only to source vocabularies in which synonymy is addressed explicitly. Furthermore, this method is limited to terms contained in the Metathesaurus, as many source vocabularies contribute only a subset of their terms. And finally, the set of concepts for which explicit synonym disagreement occurs will differ from the current state of source vocabularies due to the lag time between the release of the Metathesaurus and source vocabulary updates. The UMLS has partially addressed the issue of lag time between releases by providing Internet access to the most current version of the Metathesaurus at any given time.[8]

[‡] This assumes the more common use of the term "Bitch" to mean a female domestic dog.
[§] This compromise would not work with SNOMED terms as they are not assigned a unique term identifier.

CONCLUSIONS

Mapping of synonymous terms among multiple vocabularies is a formidable task and the Metathesaurus does an admirable job. Approximately 98% of SNOMED and Read Thesaurus concept groups contained in the Metathesaurus are **not** involved in explicit synonym disagreement. However, the remaining 2% must be identified and documented for those information systems that utilize Metathesaurus-based synonymy.

.

Table 4.1
Definitions of each field used for processing SNOMED terms.

Field	**Field Definition**
Recnum	Unique record number for each term (assigned in-house)
Term	Term
Sab	Source vocabulary
Tty	Term type status in the Metathesaurus
Scd	Concept code in the source vocabulary (e.g., Termcode in SNOMED; Read code in Read Thesaurus)
CUI	Metathesaurus concept code (Concept Unique Identifier)

Table 4.2
SNOMED terms sorted on SNOMED termcode.

Termcode	**Term**	**CUI**	**Recnum**
L-80700	Canine, NOS	C0012984	51979
L-80700	Bitch	C0324294	599552
L-80700	Canine species	C0012984	51980
L-80700	Dog	C0012984	51977
L-80A00	Feline species	C0007450	30090
L-80A00	Feline, NOS	C0007450	30089
L-80A00	Cat	C0007450	30086

Table 4.3
SNOMED terms sorted on Metathesaurus unique concept identifier (CUI).

CUI	**Term**	**Termcode**	**Recnum**
C0015967	Fever	F-03003	65856
C0015967	Hyperthermia	F-0A440	65874
C0015967	Fever 106 degrees F or higher	F-03009	65889
C0015970	Fever of unknown origin	F-03010	65902
C0015970	F.U.O.	F-03010	65915

Table 4.4
Terms of a source vocabulary concept group mapped among multiple Metathesaurus concepts.

Read Code	**Read Term**	**Metathesaurus CUI**
Read 56z	Adenitis NOS [D]Adenitis	C0001416
	[D]Lymph node enlargement NOS	C0024208
	[D]Enlarged submandibular lymph gland [D]Enlarg submand lymph gland	C0362040

Table 4.5
Terms of multiple source vocabulary concept groups mapped to a single Metathesaurus concept.

Metathesaurus CUI	**Read Term**	**Read Code**
C0024208	Lymphadenopathy	Xa0Vr
	Swollen glands	1692.
	Enlarged glands	1692.
	Enlarged lymph nodes	XM07t
	Swollen lymph nodes	XM07t
	Swollen lymph glands	XM07t
	Lymphadenopathy - swelling	XM07t
	[D]Lymph node enlargement	R056.
	[D]Lymph node enlargement NOS	R056z
	[D]Swollen glands	R0561

Table 4.6
SNOMED and Read Thesaurus concept groups split between 2 Metathesaurus concepts.

SNOMED Termcode	**SNOMED Term**	**Metathesaurus CUI/Term**	**Read Term**	**Read code**
D3-02000 **D3-02000** **D3-02000**	Hypertension Hyperpiesis Hyperpiesia	C0020538 Hypertension	- Hypertension - - High blood pressure disorder	**G2…** **G2…**
D3-02000	**High Blood Pressure**	**C0085907** Blood pressure, High	High blood pressure	XM02V
F-31003	Elevated blood pressure (not hypertension)		Elevated blood pressure	XM02V
F-31003	Increased blood pressure (not hypertension)		Raised blood pressure	XM02V
			BP - High blood pressure	**G2…**

REFERENCES

1. Moorman PW, van Ginneken AM, van der Lei J, et al. A model for structured data entry based on explicit descriptional knowledge. Methods Inf Med 1994;33:454-63.

2. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993;32:281-91.

3. Humphreys BL, Lindberg DA, Schoolman HM, et al. The Unified Medial Language System: An Informatics Research Collaboration. J Am Med Inform Assoc 1998;5:1-11.

4. National Library of Medicine. Documentation, 8th Edition UMLS Knowledge Sources. Bethesda, MD, 1997.

5. Campbell KE, Oliver DE, Shortliffe EH. The Unified Medical Language System: Toward a Collaborative Approach for Solving Terminologic Problems. J Am Med Inform Assoc 1998;5:12-16.

6. Taylor EJ. Dorland's Illustrated Medical Dictionary. 27 ed. Philadelphia: W.B. Saunders Co., 1988.

7. Cimino JJ. Auditing the Unified Medical Language System with Semantic Methods. J Am Med Inform Assoc 1998;5:41-51.

8. McCray AT, Razi AM, Bangalore AK, et al. The UMLS Knowledge Source Server: A Versatile Internet-Based Research Tool. Proc Annu Symp Comput Appl Med Care 1996:164-168.

CHAPTER 5

Use of the Metathesaurus and SPECIALIST Lexicon of the Unified Medical Language System, Lexical Matching and Domain-Specific Free-Text to Identify Undocumented Vocabulary

ABSTRACT

The technique described in this chapter identifies character strings not documented in the Unified Medical Language System (UMLS) using lexical matching within domain-specific free-text. Documented terms contained in the Metathesaurus or SPECIALIST Lexicon are removed and the remaining character strings are presented to domain experts along with the original sections of text for manual review. The key element of this technique is the SPECIALIST Lexicon which contains common words and known variations of medical terms. The removal of these words reduced the number of remaining character strings in the text from 12,075 to 574. Ninety-three percent of these remaining character strings resulted in the identification of undocumented vocabulary terms.

INTRODUCTION

Electronic processing of free-text medical information requires the identification of medical and non-medical terms, along with their lexical variations. This capability is often addressed using a controlled medical vocabulary and general lexicon, which serve as a repository of terms that are "known" to the system. A complete and well maintained repository is an essential component of an effective electronic information processing system. Unfortunately, establishing and maintaining a repository of terms and their lexical variations is difficult due to the breadth of the medical and non-medical domains, the high cost of personnel and resources, and the rapid changes in medicine with few formal mechanisms for reporting undocumented medical terms.

This research describes and evaluates an economical technique that uses lexical matching to identify medical and non-medical terms not documented in the Metathesaurus or SPECIALIST Lexicon [1] of the Unified Medical Language System (UMLS). [2] This technique capitalizes on the presence of undocumented terminology in domain-specific free-text and focuses manual review on character strings that remain in the text after documented terms contained in the Metathesaurus and SPECIALIST Lexicon have been removed. These remaining character strings are presented along with the original section of text that contained each string to a domain expert for review.

Unified Medical Language System (UMLS)

The UMLS, established in 1986 by the National Library of Medicine (NLM), was designed to "facilitate the retrieval and integration of information from multiple machine-readable biomedical information sources. The sources of interest include: descriptions of the biomedical literature, clinical records, factual databanks, knowledge-based systems, and directories of people and organizations." [3] The UMLS is organized into four Knowledge Sources which are the Metathesaurus, Semantic Network, Information Sources Map, and SPECIALIST Lexicon. [3] Two components of the UMLS were used in this work, the Metathesaurus and the SPECIALIST Lexicon. This discussion is limited to those aspects of the Metathesaurus and SPECIALIST Lexicon significant in this work.

The Metathesaurus (1997 release) is a compiled list of terms from over 40 autonomously controlled medical vocabularies known as, source vocabularies. The SPECIALIST Lexicon contains lexical variations of medical terms and non-medical terms. Examples of medical term variations are "hematopericardia" for "hematopericardium," "endoscopic" for "endoscope," and "radiographs" for "radiograph." Examples of non-medical terms are, "a," "and," "the," "but," and "for."

Some terms and variations are contained in both the Metathesaurus and the SPECIALIST Lexicon. An example would be the variations of the term "cough" as in "cough," "coughs," and "coughing." However, the term "coughed" is only found in the SPECIALIST Lexicon. Ocassionally, medical terms are found only in the SPECIALIST Lexicon such as the terms "obesity index," "obesity indexes," and "obesity indices." Combined, the Metathesaurus and SPECIALIST Lexicon include all terms currently "recognized" by the UMLS.

Domain-specific free-text

Domain-specific free-text is a rich source of terminology. Several researchers have used domain-specific free-text along with commercially available natural language processing tools for vocabulary discovery.[4,5] Other researchers have used domain-specific free-text with manual evaluation by domain experts to assess the completeness of a controlled medical vocabulary or demonstrate the preference of one vocabulary over another. [6-12] While expert evaluation is effective, it is more costly in terms of time and personnel than lexical matching with manual review. A textbook on the physical examination in companion animals [13] was used in this research as a source of domain-specific free-text.

Lexical Matching

Lexical matching identifies one character string within another and has been noted as a practical, powerful, and productive way to begin construction of a thesaurus.[14] While lexical matching is quick and efficient, it has two main weaknesses. First, the character string of a target term must be an exact character for character match (and possibly case-sensitive) with the search term. Thus, plurals and other lexical variations of a term are missed unless they are represented explicitly. Second, lexical matching does not guarantee semantic equivalence.[15] Lexical matching cannot resolve word sense ambiguity such as occurs with the term "cold" which can refer to an illness or the temperature. Furthermore, the uppercase character string "COLD" is an abbreviation for the term, "Chronic Obstructive Lung Disease." Semantic processing has been used to address these types of ambiguities. [16,17]

This work proposes an economical technique for identifying undocumented vocabulary terms, both medical and non-medical, using domain-specific free-text, lexical matching, and the Metathesaurus and SPECIALIST Lexicon of the UMLS. This technique focuses manual review on character strings in free-text that are not documented in the Metathesaurus or SPECIALIST Lexicon. No attempt is made to determine electronically, the meaning or context of these character strings. However, this technique addresses the issue of context by providing the reviewer with the original section of text containing the undocumented character string. Additional context is also provided through a searchable electronic copy of the entire text document.

METHODS

Term list preparation

Two sets of terms were used in this research, 1) English terms contained in the 1997 release of the Metathesaurus and 2) terms contained in the SPECIALIST Lexicon. Both sets of terms were imported from the UMLS distribution files into a relational database using Visual Foxpro™ and assigned a unique identifier. A subset of uniquely spelled terms (case-insensitive) from each set was extracted and converted to uppercase. Some terms in the Metathesaurus contained characters used for classification or notation that would prevent the term from being matched in free-text. Examples of such characters are underlined, "Abdomen, NOS" meaning Not Otherwise Specified or "Cold <1>" and "Cold <2>" with the number used to denote unique concepts. These classification characters and others such as "(NOS)", " NEC", "(Non MeSH)", and "MeSH Category" were removed using a "pruning" filter technique developed in-house based on ideas put forward by Sherertz et al.[18] Thus, “Abdomen, NOS” was converted to “ABDOMEN” and “Cold <1>” and “Cold <2>” were converted to “COLD.” Even though the “pruning” filter resulted in the creation of duplicate character strings, this was not an issue as the goal was to identify undocumented character strings in the textbook.

Both sets of terms were sorted in order of decreasing word count as in six word terms, then five word terms, and so on prior to matching. A word was defined as any alphanumeric text string delimited by spaces and/or punctuation. Searches were limited to terms containing 6 words or less since pilot studies using terms containing 6 words resulted in 0 matches. In fact, the only 5 word matching term identified in this study was “Measurement of central venous pressure.” Matching in order of decreasing word count was done to prevent multi-word terms from being eliminated or fragmented by single word terms such as would occur with the multi-word term, “abnormal behavior” and the single word terms, “abnormal” and “behavior.”

Text source preparation

The textbook, Medical History and Physical Examination in Companion Animals[13] was used with kind permission from Kluwer Academic Publishers. The Index and chapters relating to the canine physical examination were scanned using a flatbed scanner and converted to a text file using optical character recognition software.[**]

[**] TypeReader® Professional 3.0XA for Windows™ by ExperVision, Inc.

The resulting text file was spellchecked for scanning errors using Microsoft Word for Windows 95 version 7.0a©. A custom veterinary dictionary was created by parsing the CD-ROM version of the Merck Veterinary Manual [19] into a list of unique words. This list of unique words was saved as an ASCII text file and registered with the Word® spellchecker as a custom dictionary.

Next, the textbook file was converted to a table, one record for each sentence, phrase or heading delimited by punctuation, tab character, or carriage return. Handling sentences, phrases and headings in this manner provided the ability to link a matched term with its text source. Barrier words [20,21] such as "of" and "the" were not used as delimiters between terms, as many vocabulary terms contain typical barrier words. (e.g., "Cupping of the optic disk") Non-alpha characters (numbers and publishing characters) were trimmed from the beginning and end of each sentence, phrase or heading. Numbers imbedded within a character string were not removed.

Matching technique

Lexical matching (case-insensitive) was performed in two phases. In the first phase, each sentence, phrase and heading of the textbook was searched using terms contained in the Metathesaurus. When a term was located in the text, it was replaced with a meaningless token character. This search and replacement technique resulted in a set of sentences, phrases, and headings in which all lexically matched Metathesaurus terms had been removed.

In the second phase, the sentences, phrases and headings that remained from the first search, were searched again using terms contained in the SPECIALIST lexicon. When a term was located in the text, it was replaced with a meaningless token character. Thus, after the second phase search, all terms in the text contained in the Metathesaurus and the SPECIALIST Lexicon had been replaced with a token character. The remaining character strings were exported into a table, one record for each unique character string along with its frequency of occurrence.

Sampling and Analysis

Phase one searching, in which terms contained in the Metathesaurus were removed from the text, resulted in 12,224 remaining unique character strings, with 12,075 containing at least one alpha character and 149 containing numbers only. Phase two searching in which terms contained in the SPECIALIST Lexicon were removed from the text after the prior removal of terms contained in the Metathesaurus, resulted in 806 remaining unique character strings, with 574 containing at least one alpha character and 232 containing numbers only.

Unique random numbers were assigned to each character string that contained at least one alpha character. These character strings were sorted on the randomly assigned numbers with sampling done sequentially based on sorted order. The sample size was 230, based on a table of sample sizes from finite populations.[22] The table was constructed based on the assumption of a population proportion of .50 (½ of the character strings would belong to a given category, ½ would not) which yields the maximum possible sample size required such that the proportion in each category determined for the sample would represent the proportions of the population with a 95% level of confidence.

Character strings in the sample set were analyzed for usefulness in identifying undocumented vocabulary terms (medical or non-medical). A character string was considered useful if it represented a complete or partial term not contained in the Metathesaurus or the SPECIALIST Lexicon. Character strings for which usefulness was not immediately apparent (e.g., pseudo-pregnancy) were submitted for searching on the Internet using the Metacrawler search service.[††] If the character string was located in context on the Internet, it was judged useful in identifying undocumented vocabulary. Evaluation was performed by (EMM) with review by (JRW) to reach consensus.[12]

RESULTS

Of the 574 character strings containing at least one alpha character, 230 randomly selected strings were evaluated for usefulness in identifying undocumented vocabulary terms, either medical or otherwise. Two hundred thirteen of the 230 (93%) randomly selected character strings were considered useful.

[††] MetaCrawler is a World Wide Web search service developed in 1994 at the University of Washington by Erik Selberg, Oren Etzioni and Greg Lauckhart. MetaCrawler is now operated by go2net, Inc. MetaCrawler sends queries to several Web search engines, including Lycos, Infoseek, WebCrawler, Excite, AltaVista, and Yahoo. Results are returned to the user in a uniform format, and ranked by relevance. Metacrawler can be found at http://www.metacrawler.com/ .

Of the 17 character strings considered not useful in identifying undocumented vocabulary terms, 7 were uncorrected hyphenated words at the end of a text line (e.g., hypothyroid-ism), 5 were figure notations (e.g., 17-30a), 3 were proper names, 1 was an uncorrected scanning error and 1 was a temperature reading (e.g., 20°C)

DISCUSSION

The key element of this technique is the SPECIALIST Lexicon used to remove common words and known variations of medical terms from the text. The removal of these terms reduced the island count from 12,075, having an average word length of 4 (3.93) and an average character string length of 21.2 characters, to a more manageable 574, having an average word length of 1 (1.04) and an average character string length of 10.3 characters.

Manual analysis of the character strings that remained after phase two searching revealed complete terms, partial terms, proper names, errors in the original text, scanning errors and various combinations of the above. Complete terms included “unnoticed,” “unconditional,” “enophthalmia,” “enophthalmus” and “polypnea.” Multiple complete terms were found in some character strings such as “forefinger medially,” “gloved forefinger,” and “battery-powered microphones.”

Partial terms resulted when a documented term was removed from an undocumented multiword term in the text. As an example, the term “interoceptive receptors” is not documented in the

Metathesaurus or SPECIALIST Lexicon, but the term “receptors” is. Thus, when the term “receptors” was matched and removed from the text, “interoceptive” remained.

Partial terms can lead to the identification of multiple terms. Referring to the previous example, while “interoceptive” led to the identification of “interoceptive receptors” in the text, the investigation of “interoceptive” on the Internet led to the identification of additional terms containing the word “interoceptive” that were not contained in the Metathesaurus or SPECIALIST Lexicon. Examples of such terms are “interoceptive system,” “interoceptive cueing,” “interoceptive stressors” and “interoceptive stimuli.” It should also be noted, the term “interoceptive” is an independent term in and of itself.

Proper names in medical text pose a dilemma in that some are associated with medical terms such as, Wood’s lamp or Robert Jones bandaging, while others are not, as in the names of patients or medical personnel. One must not forget the possibility that Robert Jones could present for bandaging. For this work, only proper names associated with medical terms were considered to be undocumented terms.

Errors in the original text did occur, but most errors were due to scanning. The optical character recognition software frequently mistook the letters “r n” for the letter “m”, which led to the erroneous identification of Homer’s syndrome, rather than Horner’s syndrome. Incidentally, the Internet is not always a reliable measure of term authenticity, as Homer’s syndrome was found incorrectly at seven sites.

Hyphenated terms were particularly challenging. Many of the search engines used by MetaCrawler did not treat a hyphenated term as a phrase. Thus, a search using a hyphenated term often returned the same results as a search using both words. Some hyphenated terms such as "epiphyseal-metaphyseal" and "crescendo-decrescendo" contained terms found in the SPECIALIST Lexicon individually, but the hyphenated combination was not. Hyphens were also used to denote a specific intervertebral space such as "L3-L4." Ocassionally, there was disagreement among and within documents on the Internet regarding the use of a hyphen in terms such as, "pseudo-pregnancy," "pre-anesthetic," "sub-mucosa" or "two-thirds." Whether the use of a hyphen in a given circumstance is correct, incorrect, preferred, or not, these variant forms of hyphenated character strings occur in the medical literature and must be identified in order to be processed correctly. Some terms in the text were informal variations of documented terms in the Metathesaurus such as Boxer, for Boxer dog and Bouvier, for Bouvier des Flandres.

LIMITATIONS

The technique described in this paper is a quick and efficient first step that utilizes **what has been said** in existing medical information sources to identify undocumented vocabulary, both medical and non-medical. This technique is a "first cut" process, best suited for initial studies in a given domain. Although this technique is quick and efficient, it requires the acceptance of several limitations.

This technique will not identify multiword terms for which all words have been documented previously in the Metathesaurus or SPECIALIST Lexicon. Returning to a previous example, had the term “interoceptive” been documented previously, the undocumented term “interoceptive receptors” would not have been identified using this technique.

Matching terms in decreasing word count may have masked the discovery of significant individual terms that were contained within a multi-word term. Future study might include matching only single word terms contained in the Metathesaurus and SPECIALIST Lexicon with examination of the remaining undocumented character strings.

A punctuation character followed by a space was used to determined the end of a sentence. Unfortunately, a period followed by a space resulted in fragmentation of nerve and muscle abbreviations such as “n. facialis” and “m. temporalis.” In retrospect, examination of the text revealed that using a period followed by a space and an uppercase character would have eliminated this problem without introducing others.

Plurals are considered as distinct character strings by this technique and are only identified if contained in the Metathesaurus or SPECIALIST Lexicon explicitly. Thus, the term “semitendinosis muscles” did not result in a match, as only the singular form, “semitendinosis muscle” was contained in the Metathesaurus.

This technique does not deal well with a literary structure such as “anagen (growth period) and telogen (rest period) cycle” The separation of “anagen” and “telogen” from the word “cycle,” would result in these terms being suggested as undocumented, even though the terms "anagen cycle" and "telogen cycle" are contained in the Metathesaurus.

The technique as implemented did not exclude figure headings that contained an alpha character such as, 17-26a or 17-26b. Nor did it account for temperatures as in 20°C.
Finally, use of electronic sources of free-text for processing is highly recommended, as scanning the textbook introduced errors in the text document. Although we sought permission from several publishers, we could not obtain an electronic version of a suitable text.

CONCLUSIONS

Controlled medical vocabularies are better suited for establishing **what can be said**, than for determining **what has been said**. As a result, when one proclaims, "I can't use your vocabulary to say it my way." the response has often been, "Then say it the correct way." While this response makes a strong case for controlled data entry and keeps the size of controlled medical vocabularies somewhat manageable, it does little to assist in the electronic processing of existing medical information. Therefore, techniques are needed to improve medical vocabularies and general lexicons by identifying undocumented character strings, and presenting these character strings in context for manual review.

REFERENCES

1. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care 1994:235-9.

2. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993;32:281-91.

3. National Library of Medicine. Documentation, 8th Edition UMLS Knowledge Sources. Bethesda, MD, 1997.

4. Hersh WR, Campbell EH, Evans DA, et al. Empirical, Automated Vocabulary Discovery Using Large Text Corpora and Advanced Natural Language Processing Tools. Proc Annu Symp Comput Appl Med Care 1996:159-63.

5. Spackman KA, Hersh WR. Recognizing Noun Phrases in Medical Discharge Summaries: An Evaluation of Two Natural Language Parsers. Proc Annu Symp Comput Appl Med Care 1996:155-58.

6. Henry SB, Campbell KE, Holzemer WL. Representation of nursing terms for the description of patient problems using SNOMED III. Proc Annu Symp Comput Appl Med Care 1993:700-4.

7. Henry SB, Holzemer WL, Reilly CA, et al. Terms used by nurses to describe patient problems: can SNOMED III represent nursing concepts in the patient record? J Am Med Inform Assoc 1994;1:61-74.

8. Lange LL. Representation of Everyday Clinical Nursing Language in UMLS and SNOMED. Proc Annu Symp Comput Appl Med Care 1996:140-44.

9. Henry SB, Holzemer WL. Can SNOMED International represent patients' perceptions of health-related problems for the computer-based patient record? Proc Annu Symp Comput Appl Med Care 1994:184-7.

10. Klimczak JC, Hahn AW, Sievert ME, et al. Comparing clinical vocabularies using coding system fidelity. Proc Annu Symp Comput Appl Med Care 1995:883-7.

11. Campbell JR, Payne TH. A comparison of four schemes for codification of problem lists. Proc Annu Symp Comput Appl Med Care 1994:201-5.

12. Chute CG, Cohn SP, Campbell KE, et al. The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures. J Am Med Inform Assoc 1996;3:224-33.

13. Rijnberk A, de Vries HW. Medical History and Physical Examination in Companion Animals. Dordrecht,, The Netherlands: Kluwer Academic Publishers, 1995;376.

14. Sherertz DD, Tuttle MS, Blois MS, et al. Intervocabulary Mapping Within the UMLS: The Role of Lexical Matching. Proc Annu Symp Comput Appl Med Care 1988:201-6.

15. Kannry JL, Wright L, Shifman M, et al. Portability issues for a structured clinical vocabulary: mapping from Yale to the Columbia medical entities dictionary. J Am Med Inform Assoc 1996;3:66-78.

16. Rindflesch TC, Aronson AR. Semantic processing in information retrieval. Proc Annu Symp Comput Appl Med Care 1993:611-5.

17. Rindflesch TC, Aronson AR. Ambiguity resolution while mapping free text to the UMLS Metathesaurus. Proc Annu Symp Comput Appl Med Care 1994:240-4.

18. Sherertz DD, Tuttle MS, Olson NE, et al. Lexical Mapping in the UMLS Metathesaurus. Proc Annu Symp Comput Appl Med Care 1989:494-9.

19. Fraser CM. The Merck Veterinary Manual. Seventh Edition ed. Rahway, NJ: Merck & Co., Inc., 1991.

20. Tersmetta KWF, Scott AF, Moore WG, et al. Barrier Word Method for Detecting Molecular Biology Multiple Word Terms. Proc Annu Symp Comput Appl Med Care 1988:207-11.

21. Moore GW, Berman JJ. Automatic SNOMED coding. Proc Annu Symp Comput Appl Med Care 1994:225-9.

22. Isaac S, Michael WB. Handbook in Research and Evaluation. Third ed. San Diego, Ca: EdITS, 1995.

CHAPTER 6

CONCLUSIONS

Quality information is essential for quality patient care. The Patient Information Acquisition Model (PIAM) provides a generalized framework that is useful for improving the quality of patient information. PIAM relates the physical process of collecting data to its storage and presentation and facilitates the identification and correction of the involved processes.

The vocabulary studies in this research are based on the premise that quality information and its accurate communication requires a complete consensus-based controlled medical vocabulary. Complete should not be confused with excessive. Therefore, the first objective of the vocabulary studies was to develop techniques for identifying terms contained in a controlled medical vocabulary necessary for a given domain. For veterinary use, unnecessary terms can be excluded from a controlled medical vocabulary such as the Systematized Nomenclature of Medicine and Veterinary Medicine (SNOMED). The second objective was to develop efficient techniques for identifying undocumented vocabulary, thus addressing completeness from the more traditional approach.

Assembling a collection of terms is a relatively trivial task compared to establishing consensus as to the meaning of the terms and their synonymy. The National Library of Medicine is providing leadership and significant resources towards this effort by its sponsorship of the Unified Medical Language System (UMLS). This research contributes to their effort by describing and characterizing explicit synonym disagreement which is useful in identifying errors in synonymy as well as valid disagreements among domain experts based on perspective or context. The techniques used in this research can be used to audit synonym mappings within the Metathesaurus and its source vocabularies. Additionally, we propose a mechanism to address valid synonym disagreements when they occur.

Everyone agrees that patient data is import and there is great urgency to begin its collection. These sentiments motivated this research based on the goal of building a patient data collection instrument. However, urgency for information must not be allowed to compromise the quality of the information. At stake is the quality of decisions and the quality of patient care.

FUTURE WORK

Future work involving the Patient Information Acquisition Model will center around establishing the model's integrity and demonstrating its application. Integrity studies would entail asking clinicians to document incidents of missing or inaccurate information in the clinical setting and evaluate PIAM's ability to represent each incident. In addition, the list of characteristic questions associated with each component and activity of PIAM needs to be evaluated to determine if they would identify these incidents.

Second, PIAM needs to be evaluated for effectiveness as a design model by using its framework to develop a canine physical examination data collection instrument. In addition, PIAM's effectiveness for designing non-medical information systems can be evaluated as well.
Third, PIAM can be further evaluated for completeness when applied as a framework for organizing topics in a course on information management. It is anticipated that PIAM would provide an effective conceptual framework to support student learning.

The techniques developed in the vocabulary studies of this research can be adapted to facilitate maintenance of a standardized controlled medical vocabulary such as SNOMED. In addition, as many concepts in SNOMED are human related and not used in veterinary medicine, these techniques could also be used to facilitate development of a microglossary of SNOMED terms relevant to veterinary medicine or any individual discipline.

Metathesaurus-enhanced lexical matching needs to be evaluated more fully when applied to various sources of free-text, both formal and clinical. In addition, methods for compiling results over multiple sources of free-text need to be developed and evaluated.

Finally, it would be interesting to know what percentage of vocabulary terms discovered using the technique described in Chapter 5 represented new concepts or were synonyms of existing concepts. This would entail a panel of experts making the same judgements as are being made by the UMLS currently when adding a term to the Metathesaurus. A term representing a documented concept is associated with an existing CUI, while a term representing a new concept is assigned a new CUI.

Further research needs to verify the completeness of term lists generated using these techniques. One approach is to have reviewers judge the ability of the term list to represent concepts in a set of documents.[1] This approach begs the question of which set of documents should be used. The final verification however, will only occur when a data collection instrument is developed and tested in a clinical setting.

The effort to design a structured data collection instrument that supports accurate communication by achieving a balance among completeness, consistency, and expressiveness is well worth the necessary efforts. Improved quality of information contained in the medical record will facilitate accurate communications among healthcare providers and thus improve patient care. In addition, endeavors such as outcomes assessment and longitudinal research within and among institutions will benefit from complete and consistent patient data, supported by well-documented controlled medical vocabularies.

REFERENCES

1. Klimczak JC, Hahn AW, Sievert ME, et al. Comparing clinical vocabularies using coding system fidelity. *Proc Annu Symp Comput Appl Med Care* 1995:883-7.

VITA

Name: Eric M. Mills, DVM
Education: DVM, The Ohio State University 1979

Home address: 557 Cherokee Trail
Blacksburg, VA 24060
(540) 953-1146

Academic address: Virginia-Maryland Regional
College of Veterinary Medicine
Phase II Duck Pond Drive
Blacksburg, VA 24061
(540) 321-5564
email: rmills@vt.edu

Professional Experience

Associate Small Animal Practitioner	1979 - 1983
Owned and operated two small animal practices	1983 - current

Invited Presentations:

Mills, Eric M (1993). The Basic Components of a Practice Management System. Western States Veterinary Conference, Las Vegas, Nevada.

Mills, Eric M (1993). Selecting a Practice Management System. Western States Veterinary Conference, Las Vegas, Nevada.

Mills, Eric M (1993). Hardware Selection for a Practice Management System. Western States Veterinary Conference, Las Vegas, Nevada.

Mills, Eric M (1993). Preparing Your Staff, Your Clients and Yourself for a Practice Management System. Western States Veterinary Conference, Las Vegas, Nevada.

Mills, Eric M (1993). A Technicians Role in Selecting and Implementing a Practice Management System. Western States Veterinary Conference, Las Vegas, Nevada.

Mills, Eric M (1995). Selecting a Practice Management System. Texas Veterinary Medical Association Conference, Houston, Texas.

Mills, Eric M (1995). Preparing Your Staff, Your Clients and Yourself for a Practice Management System. Texas Veterinary Medical Association Conference, Houston, Texas.

Publications

1. Talbot RB and Mills EM. SNOMED International for veterinary medicine. JAVMA 205:1445-1446, 1994.
2. Veterinary Microglossary of SNOMED International 1995 Mills EM.

Grants

HS Bender*, BB Lockee*, EM Mills*, D Hix* *An Interdisciplinary Partnership in Biomedical Informatics*. A Support Program for Innovative Research Strategies (ASPIRES) Grant, 1998. $85,683

HS Bender*, BB Lockee*, EM Mills*, SD Forrester *Reaching Out to Veterinary Practitioners through Diagnostic Pathology Imaging: A Program for Clinical Service and Continuing Education*, ReachOUT Grant,1998 $7,500

HS Bender*, BB Lockee*, EM Mills* *Using Educational Technologies to Promote Active-Learning and Development of Skills in Clinical Problem-Solving and the Diagnostic Process*. CIL, 1998. $74,000

HS Bender*, BB Lockee*, EM Mills*, D Hix* *An Interdisciplinary Partnership in Biomedical Informatics*. Eli Lilly Company, 1998.$5000

Smith BJ*, Holladay SD, Mills EM, Klein BG, Jones, J. A computerized teaching aid for instruction in transverse-sectional anatomy of the sheep brain. *Computer Alternatives to Animal Use Fund*, 1996 $5675.00.

HS Bender*, J Burton, WR Chickering, EM Mills *Development of Deductive Reasoning Skills Using Interactive, Multimedia Programs in Veterinary Clinical Pathology*. Center for Excellence in Undergraduate Teaching (CEUT) Teaching-Learning Grant; 1996, $7500.

HS Bender*, J Burton, WR Chickering, EM Mills *Development of Deductive Reasoning Skills Using Interactive, Multimedia Programs in Veterinary Clinical Pathology*. Grant chosen for the Merck-AgVet Award for Creativity in Teaching-1996. $1000.

HS Bender*, WR Chickering, J Burton, EM Mills. *Case-Based, Interactive, Multimedia Programs for Teaching Clinical Pathology Using Naturally Occurring Diseases in Hospital Patients*. Computer Alternatives to Animal Use Fund. 1996 $4933.90.

Jones JC, Moon ML, Holladay SD, Smith BJ, Mills EM. A workstation providing access to computed tomographic images in a cross-platformed, internet-transferable format. 1996 $113,000.00

Holladay SD*, Mills EM, Jones JC, Smith BJ, Moon ML. *A computer-based teaching aid for instruction in canine transverse-sectional anatomy and computed tomography*. CEUT Teaching-learning grant. 1995; $4180.