A COGNITVE MODEL OF MEDICAL RECORD CODING: IMPLICATIONS FOR UNDERSTANDING INTER-RATER AGREEMENT

By

Emily M. Campbell

A DISSERTATION

Presented to the Department of Medical Informatics and Clinical Epidemiology
And the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

July 30, 2009

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Ph.D. dissertation of

Emily M. Campbell

"*A Cognitive Model of Medical Record Coding: Implications for Understanding Inter-rater Agreement*"

has been approved

_____
Mentor/Advisor: Aaron Cohen, MD, MS

_____
Member: Wendy Chapman, PhD

_____
Member: Brian Hazlehurst, PhD

_____
Member: Dean F. Sittig, PhD

_____
Member: Thomas Stibolt, MD

TABLE OF CONTENTS

List of Figures

List of Tables

Acknowledgements

We are all on some sort of journey, and one path of mine both ends and begins here. I have the great honor of meeting and interacting with many new people along the road. I was also privileged to be accompanied by my family and longtime friends each step of the way. I hope I am able to adequately express my gratitude to you all.

I have been at OHSU since 1981, and have worked in many different roles. Several served as clinical mentors over the years I practiced nursing at OHSU, and I am sincerely indebted to them: Susan Fischer, RN, BSN (formerly, Director of Surgical Services), Katie Flynn, RN, BSN (operating room nurse *par excellence*), Alan Barker, MD (formerly, Director of Pulmonary Medicine), and Donald Trunkey, MD (retired Chief of Surgery). Each of these individuals modeled excellence in clinical practice always highly suffused with enormous amounts of humanity and integrity. I thank each of you for demonstrating to me what we can achieve when we practice being the best we can be.

I changed roles at OHSU in 1991 when armed with a shiny new Master's Degree in Computer Science, I asked to meet Bill Hersh, MD, then an Informatics Department of one, to see if there was some fit for my combined skills in this new area of "informatics." Bill guided me to my first work in the field. Bill, thanks for starting me on this path.

I wish to especially acknowledge Joan Ash, MLS, MBA, PhD for allowing me to work with the Physician Order Entry Team for three years while I worked on this degree. At the time, the team consisted of Dean Sittig, PhD, Richard Dykstra, MD, MS, Ken Guappone, MD, PhD, Jim Carpenter, MPH, and beautiful, wonderful, always kind Cody

Curtis, MBA. It was a mighty productive, creative, open, fun, jellybean fueled blast, my friends. Thank you!

I would like to acknowledge John Stull, MD, MPH, in the Department of Public Health for being the finest instructor I have ever experienced. If I can ever be half as good as you at communicating knowledge, I will have accomplished a great deal. Socratic Method lives!

In DMICE, several administrative staff step up to the plate every time and do it with unfailing patience and good humor. Linda Slattery, I am eternally grateful for your friendship and loyalty. We will always be sisters. Andrea Ilg has been my surrogate mother, confidant, therapist, and once, a kind roommate who shared her space, when I inadvertently checked myself out of a conference one day early. Andrea, you are the best! I have succeeded because of you. Lynne Schwabe, I do not know how you do what you do – everything detail perfect, with kindness, humor, and humility. Lynne, you rock, girl! Diane Doctor makes walking the maze of forms and permissions, deadlines and responsibilities as effortless as it can be, and never have I seen her do anything with less than an enthusiastic smile. Diane, your kindness and helpfulness are phenomenal. Felicity Fields: I have known you for only a year, but I tell you, the best part of my day at work is the smile on your face when I walk in the door. Thank you.

Although I managed to run through three committee chairs in three years, I wish to acknowledge each of them for their distinct contributions to my dissertation. Kent Spackman, MD, PhD, encouraged met both to focus and to stand on my own two feet. Dean Sittig, PhD stepped in after Kent left OHSU, and helped me rope a dissertation

topic, before he, too, left Portland for the humid heat of Houston, Texas. It was a final and perfect storm than landed me in the capable sphere of Aaron Cohen MD, MS. Aaron, I do not have words to express my gratitude for your detailed and considered feedback at every step along the way. I am really proud to be the first PhD you mentored. Yours was precisely the mentoring I needed. You were there, always. Thank you, immensely.

Even though Dean Sittig stepped down as my Chair when he moved, he remained on my Committee. Dean, thank you for sticking with me, encouraging me to see it through, and calming me down when it got to me. I will always thank my lucky stars you passed through Portland when you did. You were the one person I could count on to tell me the absolute, honest truth, when I needed to hear it. I will not be able to thank you adequately for your help. Oh, and thanks for letting me write those papers for you!

The remaining members of my Committee were stellar. Tom Stibolt, MD provided me prompt, excellent feedback regarding asthma while serving as my clinical expert on this project. Tom thanks for your patience with my questions. Wendy Chapman, PhD encouraged me with kindness and insight. Wendy, I appreciate your soft touch and careful reading and consideration of my study work. Brian Hazlehurst, PhD exposed me to the field I had been looking for all my life: Cognitive Science. Thanks, Brian for teaching me the importance of cognitive modeling.

My fellow National Library of Medicine Fellows are the best! I will never forget the day Jayashree Kalpathy-Cramer put her hands on my shoulders to breathe with me. Then she

took me to India and opened my eyes. I learned more there in a month there than in a lifetime here. Jay, thanks for being my friend, mentor and colleague! Steven Bedrick, BS and I had the greatest, almost daily morning confabs, where we philosophized and solved numerous intractable problems. You amaze me with your knowledge, patience and kindness. Thanks also, Esteban, for your enormous help with Ruby on Rails as I tried to ramp up quickly. Joshua Richardson, MLS, MS has been my cube mate, fellow household improvement wizard, and gardener for three years. Josh, yer da man! And, everybody else: Adam Wright (for blazing the trail), Suzi Fei, Heather Hill, Alexey Panchenko, Zephy McKenna, Michael Mooney, Rose Campbell, and Sam Wang. What a phenomenal group of people you are!

I have worked for the Department of Anesthesiology for many years. I cannot say enough how much I appreciate the flexibility the Department has always allowed me to get my work done while working on this dissertation. Brenda Quint-Gaebel and I have run the gauntlet together. Brenda, I respect and love you enormously. Stephen Robinson, MD, Jeff Kirsch, MD, Mark Zornow, MD, and Peter Mollenholt, MD, PhD are phenomenal colleagues. You folks may ask me to jump, any time, and I promise I will jump as high as I can. I could not ask for better people to know and work with. Thank you.

Thanks also to my superb study subjects; you gave me the most insightful and detailed data to work with. Thank you for your prompt completion of the project and great feedback.

Then there is the inner circle: Diana Dutton, Stan Jones-Umberger, Chris Gruse, Stacey Fletcher, and Silke Vanderzanden. We go back a long ways, all of us. I could not have

collected a weirder or more wonderful set of tribe members, nor would I want any. We are stuck together like glue for always. What a circle of friends!

To my mother, Betsy Pless, you are and have always been unconditional love manifest. You are my greatest cheerleader and wise counsel. My sister, Patti Jones is one of my greatest life teachers; she looks for the win-win in everything and reminds what people mirror for me. Thanks Ralph, Mariah, and Ashley Jones for the encouragement and pretending I'm cool, when we all know the truth! I am so lucky you are my family.

And lastly, to my husband William: you remind me that there is good in me when I cannot find it, and help me find my way when I am stumbling in the dark. I am more than blessed to share this path with you. You are my absolute, steadfast foundation. I will never be able to thank you enough for the pure and constant faith you have in me.

Abstract

Objective

This study evaluated the level of agreement between clinicians (experts) and non-clinicians (lay persons) when answering questions and selecting supporting text from ambulatory care encounter notes. The study hypothesized that 1) clinicians would agree more often than non-clinicians across all documents and 2) agreement would be higher for both groups when subjects were asked to find explicit text in documents than when the subjects were asked to draw inferences from the text. The study was designed to shed light on the causes of disagreement among coders of clinical documents.

Methods

Eight clinical experts and eight non-clinicians reviewed 58 clinical encounter notes, answered questions about the notes, highlighted text in support of answers, and provided comments about the reasoning behind the answers and/or text selections. Study subjects interacted with a web-based data collection tool that displayed the documents and collected user input. The data were analyzed using quantitative measures of agreement for question answers and selected text as well as qualitative methods for content analysis of the comments data.

Results

The quantitative analysis revealed support for Hypothesis #1 though not for Hypothesis #2, likely due to confounders in study design. However, the qualitative analysis provided important information about how subjects search for information within clinical records

and attempt to resolve ambiguity, when present. Five general approaches emerged from the content analysis: 1) Explicit statements are best, if found, and lead to the highest agreement among subjects 2) all subjects utilize *ad hoc* heuristics based on the available data to reach conclusions, 3) poor temporal specificity creates ambiguity, 4) exceptions to common clinical presentations cause confusion among all codes, and 5) some ambiguity is irresolvable *post hoc*. Additionally all subjects in this study were able to identify relevant information in response to questions, regardless of clinical training. Finally, subjects appeared to disagree for predominantly *non-clinical* reasons.

Conclusions

The results suggest that there is significant work to do to mitigate or eliminate some of the causes of data ambiguity in clinical information systems (CIS). This work involves improving general cognition support (e.g., rendering collected data properly contextualized with other, available information), eliminating the use of secondary information (such as ICD-9 codes as proxies for problem lists) in the clinical record, and building heuristics to identify points of ambiguity for the clinician to resolve during a clinical encounter. Computational support to reduce ambiguity may help reduce the introduction and proliferation of ambiguity in the medical record, increasing the likelihood of higher inter-rater agreement among coders.

1. Introduction

Medical information is proliferating on an unprecedented scale: a simple count of new bibliographic entries indexed in the U.S. National Library of Medicine's Entrez PubMed system reveals more than a six-fold increase from 110,291 total citations added in 1960 to 688,708 in 2005.[1] Beyond this bibliographic data, the steady adoption of clinical information systems such electronic medical records (EHRs), computerized provider order entry (CPOE) systems, and specialty-specific applications as well as the unprecedented growth of health and biomedical content on the World Wide Web results in the accumulation of vast quantities of medical data on a minute-by-minute basis. Indeed, the sheer volume of clinical data now makes it virtually impossible for humans efficiently to index and retrieve information without some form of automated support. Indeed, it is the development and optimization of such automated support that drives much of the current research in the Medical Informatics.[2]

Unaided human performance in coding or identifying relevant concepts in text remains the benchmark against which indexing tasks are evaluated; results from any coding method (whether automated or not) must approach or equal human performance on the same or similar data sets to be considered successful. In addition, reliable benchmarks provide a means of comparing the different approaches to determine which performs best given a specific task. Thus, it is imperative that we develop a full understanding of how reference standards are established, what causes humans to disagree when coding clinical information, and how we can resolve areas of variability among coders. Subsequently,

there exists a substantial body of research regarding the creation and evaluation of reference standards.[3-12]

Inter-rater agreement is a measure of the level of agreement among judges performing a coding task using textual data. This measure should approach 100% in the case of a perfect reference standard. If, instead, there is a large amount of disagreement among judges, the coding task is not reliable and the quality of the resulting reference standard can be questionable.[6, 13] As a result, improving agreement through reduction of inter-rater variability is an important goal in the creation of reference standards to benchmark coding tasks. However, despite demonstrated success in studies implementing and evaluating various techniques to improve inter-rater agreement[5, 9, 14-19], there appears to be no single cognitive model developed and tested as a theoretical foundation for contextualizing the actual task of coding textual data in these studies. In addition, no taxonomy comprising types of inter-rater variability has yet emerged from this research.

This proposal presents a cognitive model of the task of identifying concepts in clinical textual data to provide a theoretical basis from which to explore inter-rater variability. The purpose of this research is to help understand and classify, from a cognitive perspective, the sources of disagreement among individuals completing specific coding tasks.

This proposal begins with a simple introduction to the field of cognitive science including an overview of common terminology used in the field. This is followed by a very basic review of neuron anatomy and physiology to provide a biological foundation for a discussion of established theories of cognition as well as the proposed cognitive model of

textual coding. Once these topics are complete, four of the major, currently accepted models of general cognition are introduced. This is followed with a focus on process models for general cognitive tasks (e.g., knowledge representation, spreading activation theory, priming effects, etc.). After this general overview, we introduce models of specific cognitive tasks more germane to this study (e.g. reading comprehension, pattern recognition, categorization, etc.) used to code medical documents. Following these foundational sections, we turn to the development of reference standards and the use and interpretation measures of inter-rater agreement. We then discuss the task of coding medical documents followed by a description of the proposed cognitive model of this process, followed by the statement of the research hypotheses derived from the model and methods that will be used to validate them.

2. Background

2.1. The Field of Cognitive Science

2.1.1. Component Disciplines

Cognitive science is the scientific study of the mind. The field seeks to understand and explain how we think, reason, feel, hope, believe, perceive, analyze, know, and learn while making sense of and interacting with our physical environments through the study of concepts such as attention, categorization, learning and expertise, reasoning and problem solving, performance, memory, mental representation, and spatial representation and imagery.[20-23] Given these broad foci, cognitive science is an interdisciplinary study, drawing on contributions from neuroscience, cognitive anthropology, linguistics, computer science, psychology, and philosophy as it seeks to understand and exploit these processes of cognition. Neuroscience explores the physiological processes that afford and mediate cognition to reveal, for example, how we are able to receive, interpret and respond to sensory information, as in how we are able to see or hear.[24] Cognitive anthropology frames cognition within social context to understand how history, culture, and society shape thought and action.[25] Linguistics focuses on how we acquire, utilize, and derive meaning from both spoken and written language.[26] Computer science (specifically the sub-discipline of artificial intelligence) models cognitive processes to leverage the speed and accuracy of computers to either fully automate processes computers can complete more efficiently than humans or provide computational support to augment human processing abilities.[24] Cognitive psychology seeks to explain how emotions, feelings, beliefs, culture, and education affect human memory, learning style,

acquisition of skill, and knowledge communication and presentation.[27] Clearly, these disciplines overlap, and, virtually all the theoretic underpinnings of the cognitive science remain under debate as both scientists and philosophers wrestle with the methodological and ontological questions various cognitive theories pose."[28] However, each of these disciplines offers a unique contextual framework from which to explore the many facets of cognition; together these fields of study offer us multiple perspectives into this complex domain.

2.1.2. Terminology Overview

Biological organisms with a central nervous system characteristically demonstrate the ability to alter their behavior based on experience. Learning is the acquisition of the skills, knowledge, wisdom, and information that makes relatively permanent changes in behavior possible. Memory is generally defined as the mechanism by which humans retain, store and recall what is experienced or learned over time. Knowledge is structure in which the information gained through the psychological processes of perception, learning, and memory recall is stored.[29, 30]

Many different categorizations of memory exist in the literature, each using a different approach in organizing the various processes and functions identified with it. For example memory can be divided into conscious and unconscious processes. Recalling where I had dinner last night and what I ate, or knowing facts like Raleigh is the capital city in the state of North Carolina represent conscious processes. Activities such as writing, on the other hand, use unconscious memory: unless one is learning to write, the

specifics of holding the pencil and using it to create marks on paper is not a conscious process (though the expressed content of those marks may be consciously driven). Memory can also be categorized by the type of memories created, stored or used; thus, some researchers categorize the concept according to whether the memory is reflexive, or is for events, facts, or procedures In addition, memory can be classified according to its reported anatomic correlates, as determined with studies revealing areas of brain activation during task performance. .[30-33]

In very general terms, working memory (also called short-term memory) is a theoretical construct referring to those processes that temporarily learn, store, and manipulate information.[30] This form of memory has a very short duration and includes what some researchers call immediate recall, a kind of parrot-like ability that allows subjects to repeat, verbatim, short series of numbers or words immediately after they are read or heard. Working memory degrades quickly if attention is not diverted by another cognitively demanding task or if the information in place is not constantly refreshed; although some people may retain information for longer periods of time, depending on the complexity of the information and their cognitive abilities.[33, 34]

Long-term memory, another general theoretical construct, lasts longer than working memory, and may persist for years, if not for life, depending on how frequently the information comprising the memory is recalled, as frequency of retrieval appears to correlate directly with retention.[31-33] In general, working memory is highly susceptible to disruption (e.g., environmental distractions, drug use, trauma, etc.) whereas long-term memory appears somewhat more refractory to such interference.[31] Both working and

long-term memory are important for learning; they are theorized to interact with one another and operate in parallel.

Working and long-term memory are theoretical constructs only. The terms are widely used, however, because they effectively describe the different cognitive processes that have been consistently observed in an extremely large body of research over the last century. They also help elaborate the biological processes associated with cognition, as evidenced in neuroscience research. I will explore more detailed definitions of working and long-term memory when introducing cognitive models of memory later in this document.

2.1.3. Biological Foundations

An engram is a postulated neurophysiological process that changes an organism's protoplasm in response to repeated stimuli and subsequently results in the persistence of a memory.[29] Engrams are often called memory traces; they are formed in working memory as sensory input is processed. As already mentioned, the duration of memories in working memory is only temporary, implying the existence of some mechanism that converts short-duration engrams into long-term memories. This process is termed consolidation, and refers to the stabilization of memory traces over time, so that experiences can become permanent in long-term memory.[30, 35]

The concept of a memory trace has a sound biological basis, described by the creation, integration, and transmission of nerve impulses along neurons, or nerve cells, the basic building blocks of the central nervous system.

In order to create memory, the brain must be able to abstract (simplify) information and encode it is some manner. It is generally accepted that the encoding and recording of events are associative by nature. That is, when we recall a memory about an event, we usually recall associated features such as where and when the event occurred.[36] This implies that the different memory processes likely operate in parallel in order to capture the relevant features associated with the remembered event (because some of the features may be visual, others may be verbal, and still others may be conceptual). And, because many events occur only once, cannot be anticipated, and may be episodic in nature, it is vital that the processes that encode and store the relevant memory traces of such events occur in real time and with great efficiency, such that event-associated information can easily be retrieved at a later time.[37] I now turn to the cognitive models based on these foundations.

2.2. Conceptual Models of Memory

2.2.1. Introduction

The conceptual boundaries between working and long-term memory are at best unclear; they are, after all, theoretical constructs. In fact, some researchers, most notably Ericsson and Kintsch,[34] argue that working memory extends into long-term memory, and postulate a third type of memory called *long-term working memory*. The functional association of certain forms of memory with neuroanatomy helps distinguish the types, at least in physiologic terms. However, the concepts of *working memory* and *long-term memory*

efficiently and coherently account for a large body of research and have been used widely since the 1960s, even though our understanding of just exactly how both types of memory function remains the focus of theoretical and philosophical debate.

To elaborate current thinking about how memory works, researchers create conceptual models of the processes involved. These models are abstractions intended to demonstrate the functional relationships among components in the memory "system" and they usually include both a verbal description as well as a schematic representation of the system's component elements and their interactions.[30] In order to create a model, the author must determine what comprises the system to be modeled, and how the component elements of the larger system are bounded.[38, 39] In the case of conceptual models of memory, the system includes everything (e.g., resources, structures, actions) that processes sensory input utilizing stored and recalled information. There appears to be considerable agreement among researchers on this general system definition. The models they propose, however, differ significantly in how the system's subcomponents (or, subsystems) are defined and operationalized. In particular, current models vary significantly in terms of where the authors establish boundaries between working and long-term memory and how the functions of these forms of memory are differentiated and what aspects of general cognition are being modeled.

I introduce several of the most influential theories here, in the order of their emergence in the literature, to provide an overview of current conceptual models of memory. For each of the models I provide a general description followed by a review the empirical evidence

supporting model formulation, so that I can refer back to this evidence as supportive of the cognitive model proposed in this document.

2.2.2. The Baddeley and Hitch Multi-Component Working Memory Model

In 1968, Atkinson and Shiffrin[40] proposed that working memory was no more than a unitary short-term storage system. Over time, this definition became inadequate to describe the non-storage activities that appeared to be managed by working memory, that is, how information was temporarily maintained and manipulated by the brain. In 1974 Baddeley and Hitch[41] created a new model of working memory to account for the information processing activities (what the authors called the "working" part of working memory) that appeared to function in concert with the storage and recall processes in long-term memory. In its original instantiation, the model comprised three components: a central executive system that managed two "slave" sub-systems: one to process verbal and acoustic data (the phonological loop), and another to store visual and spatial information (the visuospatial sketch pad). Baddeley[42] extended this model in 2000 by adding a fourth component, the episodic buffer, to account for phenomena that did not fit well to the original model, specifically how information was moved back and forth from long-term to working memory. The contemporary model is shown in Figure 1:

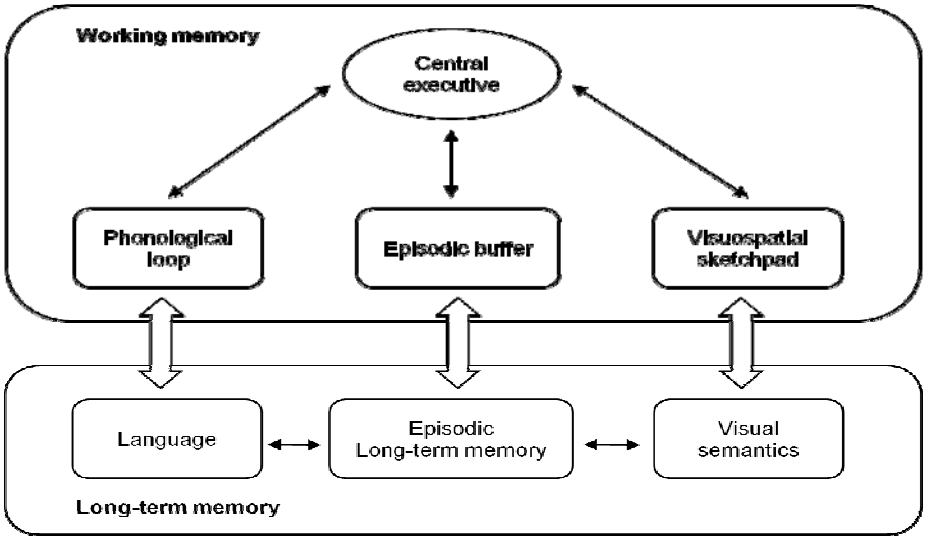


Figure 1: The Baddeley & Hitch multi-component model of working memory

In this model, a central executive manages two slave subsystems, a phonological loop that processes sound and keeps memory fresh through an articulatory rehearsal process, and a visiuospatial sketchpad that processes visual and special data. The episodic buffer moves data from long-term memory in response to inputs from the two slave sub-systems. (Adapted from Baddeley and Hitch [1974] and Baddeley [2000].)

The executive controller was originally described as limited-capacity attention manager functioning like a homunculus, "a little man who took the important decisions as to how the two slave systems should be used."(p. 6)[43] The concept was intentionally broad and served the convenient function of encapsulating the set of general processing routines that likely dealt with information management issues such as resolution of conflicts between the phonological and visuospacial subsystems, approaching novel problems, keeping attention focused on relevant information, or dividing attention when multiple tasks were taking place at the same time.[41, 43] Though the explicit functions comprised by the

executive system are not yet completely understood, the concept continues to encapsulate the necessary processes that manage the activities of working memory.

The phonological loop consists of two sub-systems. The first is a phonological store that stores the sound patterns of language, which are necessary for speech recognition. Because these patterns decay rapidly (in as little as 2 seconds), the phonological loop utilizes a secondary system: an articulatory rehearsal process that continually refreshes memory through subvocalization, the acoustic coding characteristic of silent reading that is necessary for speech production.[41-43]

The visiuospatial sketch pad subsystem is responsible for the maintenance and manipulation information needed for spatial orientation and visiuospatial problem solving. Like the phonological loop, it is comprised of two subsystems; one that processes visual input (dealing with features such as color, shape, texture, etc.) and another that manages information about spatial location, and, possibly, associated kinesthetic components. The "sketch pad" notion implies a working area where visual and spatial information interface, regardless of whether the data are accessed via sensory input or from long-term memory, so that visual information can be bound together with other sensory information to properly orient objects in space.[30, 41-43]

In the extended model of 2000, Baddeley proposed the concept of an episodic buffer to account for certain types of memory that did not appear to fit well within the original framework and to conceptualize an information integration system whose function was distinct from either the language processing of the phonological system, the visual and special data processing of the visiuospatial system, or the management activities of the

executive controller. In proposing this addition to his original model, Baddeley sought to conceptualize the process that moved information to and from long-term episodic memory, integrated this information with data from the two working memory sub-systems, and stored what can he called "crystallizations" of information in integrated, multimodal episodes or gestalts.[43, 44]

2.2.3. Ericsson and Kintsch's Model of Memory

In 1995, Ericsson and Kintsch proposed a new model of working memory to respond to their belief that other models failed to offer "…plausible accounts of the increased demand for available information required by skilled processing in…complex tasks."(p. 213)[34] In particular, the authors found the Baddeley and Hitch model of 1974 inadequate to explain observed phenomena such as why skilled subjects demonstrated greater working memory capacity than did less-skilled subjects when performing laboratory tasks, or why expert users, when interrupted during tasks that required expertise, could resume their work without apparent impact on performance. Although Ericsson and Kintsch agreed with the accepted notion that permanent storage of new memory traces could take up to 5-10 seconds,[45] they disagreed with data suggesting that access to these stored engrams involved cognitively slow recall times of around one second. For skilled performance, they argued the rapid and reliable storage and recall of information to and from long-term memory could not be explained based on a limited-capacity working memory model with constant, slow lookup and retrieval from long-term memory.

In response to these concerns, the authors suggested a mechanism whereby individuals with sufficient training and practice could leverage long-term memory as a form of

working memory to extend what is commonly considered a bounded working memory capacity. The authors called this mechanism "long-term working memory."(p. 211)[34]

In this model, individuals encode information stored in long-term memory using retrieval structures which act as cues for relevant information recall. Because skilled subjects have been shown to demonstrate not only more rapid but more extensive recall than less-skilled users, Ericsson and Kintsch believe that skilled subjects must be able to create and use more advanced retrieval structures. In general, these structures are hierarchical in nature, where retrieval cues, representing links to chunks of information, are stored in long-term working memory. The retrieval cues also include data associations – information about the coding context – that allow access to the actual long-term memory store.[34] In this way, only a small number of chunks of information need to be stored in working memory; once unpacked, these chunks function as indexes into long-term memory for locating relevant memory traces. Figure 2 demonstrates this structure:

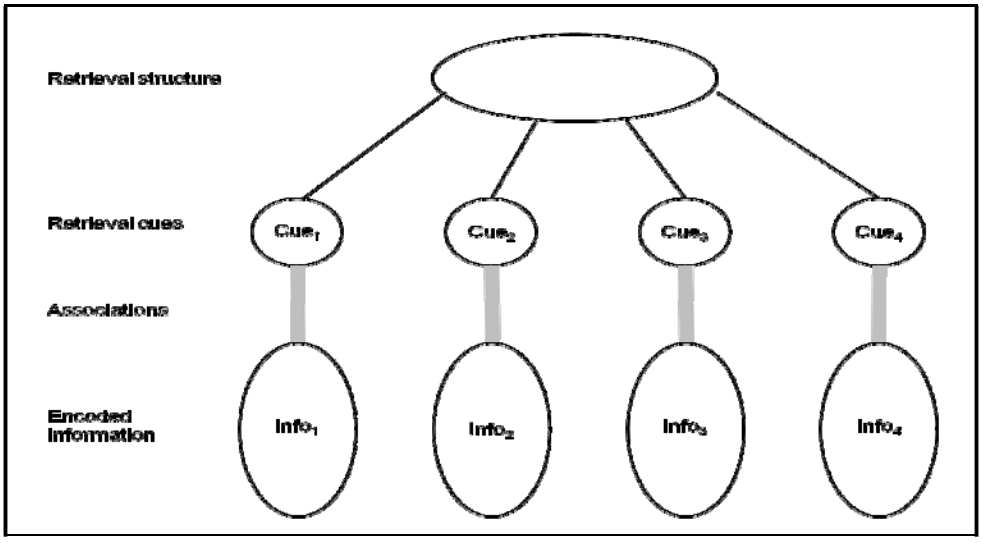


Figure 2: Ericsson and Kintsch's long-term working memory model

At the top of the model is a stable retrieval structure with its associated cues. Cues are associated with encoded information in long-term working memory and are activated by the retrieval structure to recall information as needed. (Adapted from Ericsson and Kintsch [1995].)

Ericsson and Kintsch argue that restrictions must be in place for their long-term working memory model to succeed. First, subjects must be able to store large amounts of information in long-term memory. This requires that subjects possess a large body of relevant knowledge (e.g., they have already stored a large variety of patterns of information in the domain). Second, the memory task to be performed must be familiar to the subject, because it is only under this condition that the subject can anticipate what information is likely to be necessary to accomplish the task. If these two criteria are met, then the subject can successfully store information in long-term memory. Finally, the subjects must be able to associate the encoded information stored in long-term memory with the appropriate retrieval cues. Once a set of retrieval cues stabilizes, then a retrieval structure has been formed. So, once a memory needs to be recalled, only the node indexing the retrieval structure needs to be in working memory, along with the cue indicating which type of information needs to be recalled.[34, 46]

It is important to note that Ericsson and Kintsch's model focuses on skilled behavior in the performance of memory tasks; the development of the type of efficient recall demonstrated by experts may take years to develop. However, the model supports the view that long-term memory may be more plastic than previously thought and that with substantial training and/or practice, it may be used to offset some of the limited capacity

of working memory. Finally, the model suggests that retrieval structures may enable subjects to efficiently use a simple concept to evoke many related facts or concepts in an organized fashion.[34]

2.2.4. Cowan's Memory Model

Nelson Cowan's memory model focuses on working memory capacity limitations by positing that working memory does not exist as a separate entity per se, but is instead part of long-term memory such that any short-term memory representation in working memory is just a subset of full representations in long-term memory. Working memory consists of two embedded levels: 1) activated long-term memory processes (of which there can be an unlimited number) and 2) focus of attention that can hold about four of these representations in conscious focus at a time. Activation is the process whereby an information item is "tagged" in long-term memory as relevant in the current context, making it more readily accessible to the focus of attention than other non-activated information. Activation serves to prevent conceptual associations from decaying over time due to memory decay, interference from other stimuli, or other causes of temporal decay.[47]

Cowan makes four assumptions regarding a theoretical framework for working memory capacity: 1) the only limitation in capacity appears to be in the focus of attention, 2) this limit appears to be about four chunks in humans, 3) there are no other capacity limits on other mental faculties, and 4) "…that any information that is deliberately recalled, whether from a recent stimulus or long-term memory, is restricted to this limit in the

focus of attention."(p. 91)[47] Using these assumptions, Cowan suggests that what is commonly called "conceptual" short-term memory corresponds to his activated long-term memory processes and that what is normally referred to as "visual" working memory is instead focus of attention. Figure 3 illustrates Cowan's general memory model:

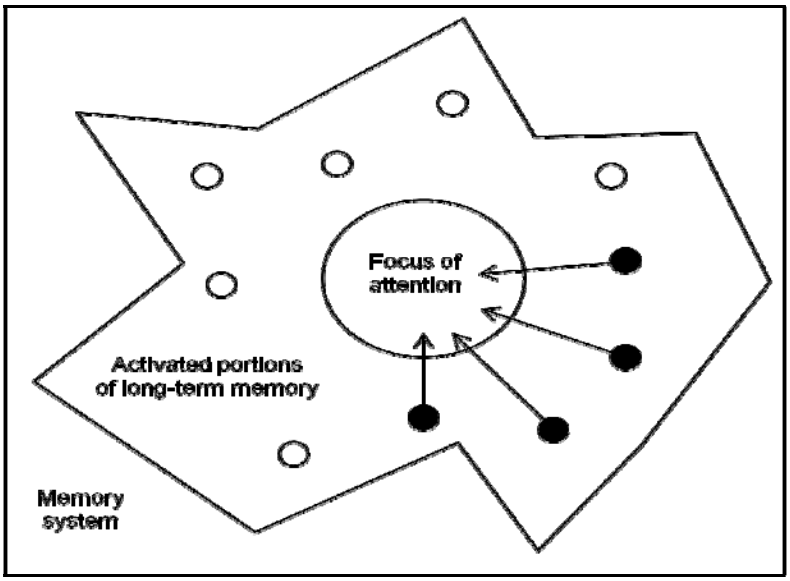


Figure 3: Illustration of Cowan's working memory processing

In this figure, the entire space conceptually demarcates memory. Activated long-term memory is represented by the jagged lines and the focus of attention is the large circle. In this example, the solid-filled circles represent four conceptual chunks of information that are currently the focus of attention. The other circles represent activated concepts in long-term memory and available to become the focus of attention. (Adapted from Cowan [2000].)

Much of Cowan's model has been explored by other researchers. Oberauer, in particular, has extended Cowan's model slightly by proposing that memory can be conceptualized as a concentric structure composed of three layers, each with different functionality. The

first layer consists of Cowan's activated long-term memory. The second structure is a region of direct and immediate access that can hold a limited amount of chunks; this corresponds with Cowan's focus of attention and holds about four chunks of information total. The third and innermost layer represents the object currently selected as the focus of the next cognitive process, and thus has a capacity limit of one chunk, because for each cognitive process, one item must be selected to the exclusion of others. Figure 4 illustrates Oberauer's modification of Cowan's original model.[48]

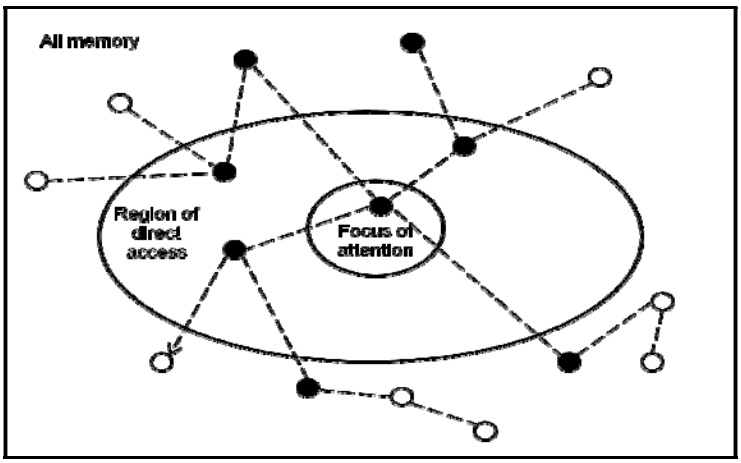


Figure 4: Oberauer's modification of Cowan's model

Long-term memory traces are represented by networked nodes and lines. Black nodes are activated. Three activated nodes are located in the region of direct access; a single activated node is located in the focus of attention, where it is selected for processing. Activated nodes outside of the region of direct access are accessible via indirect links (dotted lines). (Adapted from Oberauer [2002].)

2.3. The Computer Model of Cognition

Computational metaphors provide a helpful means for describing human cognition; indeed, humans are often likened to computers because so many features of the mind operate with similar processes. For example, humans receive cognitive inputs through the senses. We output thoughts, feelings, action, knowledge, etc. in response to the processing of inputs. As with computers, our actual processing capacity is functionally dependent on the availability of resources required for the successful performance of a task or the execution of a process. These resources can be critical or limited. Critical resources are those necessary for the successful performance of a task or the execution of a process. Thus, in cognition, just as in computer systems, limits to critical resources can lead to problems in information processing.[20] Working memory clearly represents a critical resource limitation in computer model of information processing. The fact that information placed here is transitory and that we are likely to be able to hold only "four plus or minus one" items at a time in working-memory elaborates these limitations.[47, 49] Limited resources, on the other hand, are those that may be in short supply relative to what is needed in the current context. In the computer world, such limitations include physical memory capacity, processing speed, internal I/O bus speed, etc. [20]

In 1982, as part of their work to create a scientific basis for studying human computer interaction, Card et al[50] proposed a formal model of human information processing that paralleled computer processing. The model identified a long-term memory store and a working-memory store, each with quantitative limitations and postulated computational processing speeds the authors derived from the psychological literature. Within the working memory store, the authors described a visual image store and an auditory image

store that processed incoming visual and auditory stimuli and were managed by a perceptual processor. A cognitive processor operated much like the executive controller of the Baddeley and Hitch model, managing the perceptual data, determining relevance in the current context, retrieving and incorporating long-term memory data, and storing new data or editing old data. In addition, a third motor processor managed motor functions resulting from cognitive processing.[20, 50] Figure 5 illustrates this model:

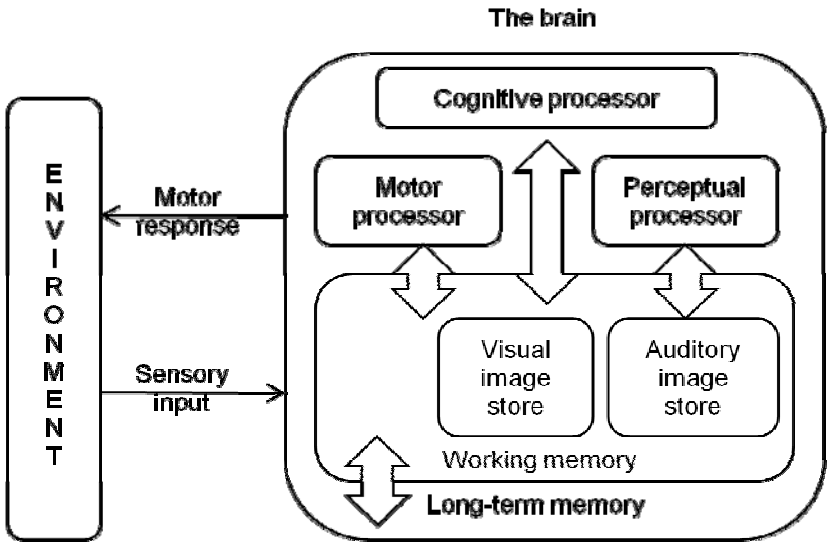


Figure 5: The human information processor

In this model, the brain responds to sensory input from the environment with the perceptual processor that stores visual and auditory data in working memory. The cognitive processor determines the relevance of this information and manages retrieval from and storage to long-term memory. The motor processor creates motor responses to the environment in association with the cognitive processor. (Adapted from Card, et al [1983])

Much like computers, humans utilize memory -- what we have called working-memory, above -- and storage (e.g., long-term memory) to store and recall information for long periods of time. However, unlike computers with set configurations (e.g., fixed size RAM or disk space, processor speed, slow or fast I/O bus for reading from and writing to disk, etc.), human memory and storage is highly flexible and is greatly affected by additional features such as intelligence, education, skill level, experience, culture, time, and other factors that can impact cognitive processes. It is at this point that explanations of the theoretical processes of cognition diverge somewhat from the computational metaphor.

When we discuss limited resources in the computer sense, we think of physical features we can enhance. We can, for example, add more processing speed or memory to improve performance on computationally intensive tasks. In human cognition, outside of training, rehearsal, or experience, which can result in improved cognitive performance on some tasks, we are not so readily able to modify such features. So, it is useful instead to consider task complexity and the demands tasks make on available cognitive resources to understand limits to human information processing. One particularly useful concept is that of cognitive load, that is, how the complexity of a task places demands on available cognitive resources. Cognitive load has been shown to correlate directly with such factors as learning time, fatigue, stress, proneness to error, and the inability to effectively manage multiple tasks simultaneously.[20]

2.4. Theories of General Cognitive Processes

This section provides foundational information about general cognitive processes. We begin with a discussion of working memory capacity and the time it takes to lay down memory traces. This is followed by an overview of theories of knowledge representation, concept activation, and priming.

2.4.1. Working Memory Capacity

As early as 1890, psychologists had already distinguished between a primary, limited-capacity memory (what we now call working memory) and a secondary, unlimited-capacity memory (long-term memory). Although the concept of a limited-capacity memory was well accepted, it was not until 1956, that George Miller famously suggested that humans can maintain "seven plus or minus two" perceptual (or, meaningful) units of information in working memory at a time,[51] and with this statement, became the first author to quantify working memory capacity. Miller referred to these perceptual units as "chunks" and described them as the mental recoding of low-information-content data into smaller units of high-information content data -- basically hierarchical abstractions of data into manageable units. Cowan more recently refined this definition somewhat, describing a chunk as "…a collection of concepts that have strong associations to one another and much weaker associations to other chunks currently in use."(p. 89)[47] Regardless of the specific definition used, since Miller's time, the "magic number" of seven chunks has been questioned, with newer theories suggesting that "four plus or minus one" chunks more precisely quantifies working-memory storage capacity.[47, 49] While the precise capacity of working memory remains under debate, there are no known

theoretical constraints on the capacity of long-term memory, though, this form of memory can be limited by a variety of factors such as intelligence, age, disease, or trauma.

2.4.2. Knowledge Representation

Much of the research in Artificial Intelligence (AI) concerns itself with modeling human cognition, specifically in the areas of knowledge representation and pattern recognition. *Knowledge representation* requires precise and consistent rules, and generally relies on *propositional representations* of objects, their current state, or their relationship to other objects.[52] Because propositional representations are discrete, atomic entities, knowledge or understanding of the *relatedness* among representations necessitates organizing methods such as *schemas* to define and demonstrate these relationships.[53]

The previous sections very generally describe the processes involved in reading comprehension. To integrate these processes within an encompassing cognitive theory requires that we build a theory of how information is stored, and in what form. For this reason, cognitive theories of reading comprehension concern themselves with how we form *mental representations* of information as we process reading material, integrate what is read with existing knowledge, draw inferences from this combination, and gain new knowledge as a result. Because readers bring several different kinds of knowledge to the reading task, depending on their experience and education, the model must account for existing knowledge, whether domain-specific or general in nature. In addition, the model must include discussion of how different forms of knowledge (e.g., syntactic and

semantic knowledge) may be stored, recalled, and brought to bear on the text being processed. As a result, models of reading comprehension rely on theories of *knowledge representation*.

In order to comprehend symbols (e.g. textual components), we must recognize and decipher the symbols, and utilize our knowledge of the world and the current context to ascribe meaning to those symbols. The computational approaches described above represent our best ability to model these tasks. In the place of neural activation we can direct pattern matching or more advanced statistical inference methods to classify text, with or without world or domain knowledge.

Propositions

It is commonly accepted that information is conceptually represented in units called *propositions*. A proposition is a symbolic structure comprising a declarative statement, or, more precisely, a structure consisting of a predicate and one or more arguments. It is the smallest unit of information or knowledge that can be true or false.[53] For example, consider the following sentence:

Sarah used her inhaler.

This simple sentence contains several propositions. The first is that Sarah is an agent performing an action. The second proposition is that an action is performed; the action is "use" (in this case, in the past tense). The fact that the action is expressed in past tense implies that the action took place before the current moment, and represents a third proposition. Finally, the fourth proposition embedded in this simple sentence explains that Sarah (the agent) used the inhaler (the object). We do not consciously identify

agents, actions, and objects within propositions when we read sentences, and these notations have no practical application in real time, nor do they have any theoretical significance in terms of reading comprehension. However, what is important is that readers do have some internal representation of information arranged in some fashion to denote concepts and their relations. When reading the sentence above, for example, it is not difficult to assume that Sarah is a person, that her inhaler is a device, and that the inhaler can be used to dispense medication. In order to make these inferences, clearly we must have some internal representation of "person" and "inhaler" and actions that can relate the two.

Simple propositions can be combined into much more complex propositions as needed to fully represent a concept. As demonstrated above, even simple sentences can contain many propositions. In general, we think of propositions as representing semantic meaning. Under this view, two sentences mean the same thing when they represent the same proposition, regardless of the words used to construct the concept. However, propositions can represent any information. For example, there is nothing that precludes propositional representation of syntactic equivalence (e.g., similar grammatical construction) among phrases – here the represented information deals more with structure than meaning.[53] To these ends, propositions can represent any concept, from the most simple to the most complex. Indeed, propositions form the basic building blocks of all schemata.

Schemas

A *schema* represents the organizational aspect of knowledge representation. In general, a schema contains slots with labels identifying the information stored at that location. In addition, the schema demonstrates the relationships among slots, as well as relationships among related schemata. For example, consider a schema of the complex medical concept, pulmonary embolism:

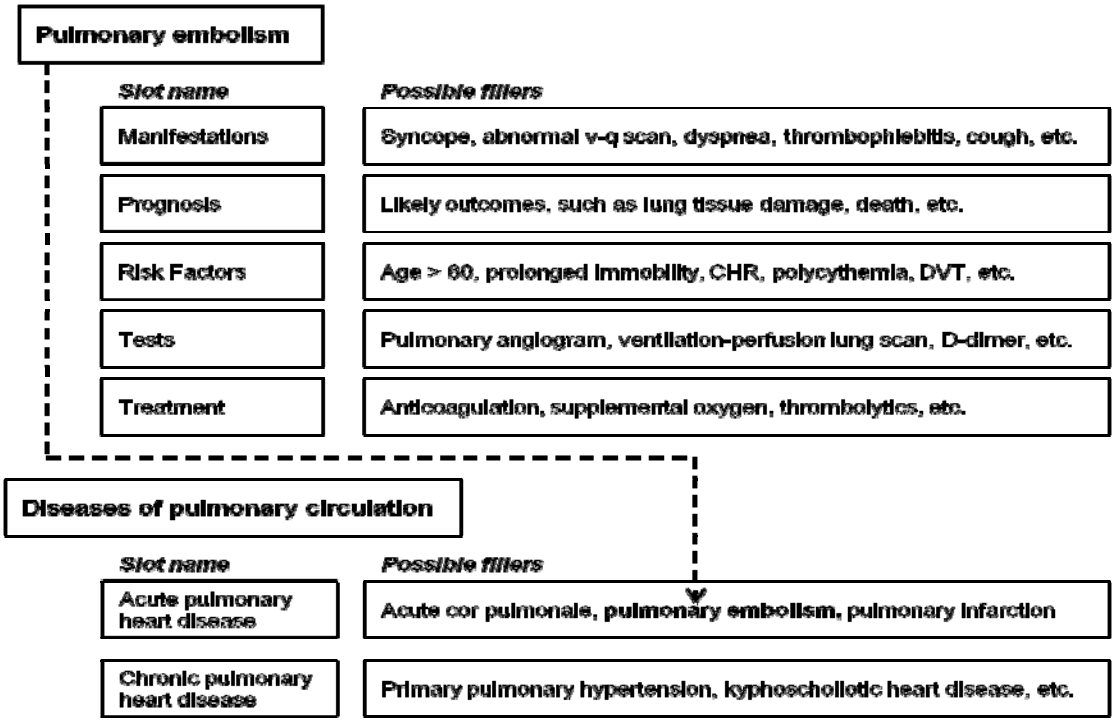


Figure 6: Example schema for pulmonary embolism

A schema for pulmonary embolism contains slots for various kinds of information, such as manifestations, prognosis, risk factors, test and treatment. Each slot has a series of possible fillers that represent default values for that slot. Note that pulmonary embolism can be a filler in the slot labeled "Acute pulmonary heart disease" in the schema for diseases of pulmonary circulation. The dashed line demonstrates a relationship between schemata. [Adapted from Turner (1992, p. 1149)[54] and Just & Carpenter (1987, p. 11)[53]]

A generic schema such as the one depicted in Figure 8 provides the basis for a more specific knowledge representation constructed by the reader when reading about a specific case of pulmonary embolism. If the text provides filler information for a certain slot (e.g., "The patient presents with an abnormal ventilation-perfusion scan.") the reader can fill the slot "manifestations" with this information. If the text does not provide information, the default fillers represent the knowledge held about pulmonary embolism which may or may not be called upon to make sense of the text being read. A schema also helps readers resolve references in the current context. For example, if the text provides information that the patient has been immobile for a prolonged period of time, has deep venous thrombosis (DVT) and is receiving anticoagulants and supplemental oxygen, the reader can likely conclude that a pulmonary embolism is being treated and therefore access the default values for prognosis.[53]

Schemas are useful for representing what a reader learns as well as what a reader already knows, because they help organize background inferences – those concepts that can be deduced without explicit description in the text. This has been demonstrated experimentally. When readers are presented with an approximate conceptual schema prior to reading difficult or complicated text, they generally demonstrate better recall and comprehension than when no preliminary schema is offered. This suggests that the readers lack the ability to formulate a useful schema when they are not familiar with the topic of the text.[55]

Semantic Networks

A semantic network is a used to represent knowledge or to support systems that reason about knowledge. The semantic network consists of a series of nodes interconnected by links or arcs. In terms of knowledge representation, each node is a single concept and the relationships among nodes in a network are demonstrated by links between the related nodes. Node links are generally labeled with information describing the type of relationship between the connected nodes, so labels include relationships such as "is a", "has", "can", etc. For example, in a purely hierarchical network, a subordinate (child) node might be related to a higher level (parent) node by the relationship *is a* meaning the child node is a type of the parent node, as in cocker spaniel *is a* type of dog. Each node can connect to multiple nodes, and links between nodes can be bidirectional.[56] Links may have what Collins and Loftus have called *criterialities*: numerical values indicating how important the link is to the concept. Higher numbers represent more critical links.[57] Paired links between any two concepts can have different criterialities, indicating general directionality in the relationship. For example, it may be critical for the concept of cocker spaniel that it is a dog (e.g., the link from cocker spaniel to dog will have a high criteriality), but less so that for the concept of dog that one type is a cocker spaniel (e.g., so the link in this direction will have a lower criteriality score). In some semantic network representations, closer relationships between concept nodes are represented by shorter lines; in others "is a" relationships may be designated with different line styles than other types of semantic relationships.[53] Regardless of the linkages among nodes in a network, the complete meaning of any concept (node) in a network is represented by all of the links to that node from elsewhere in the network.

There are many kinds of semantic networks; we are less concerned with the differentiation here, and focus instead on the general representation of information using these structures. Figure 7 depicts a portion of a *hierarchical* semantic network representation of knowledge for the concept "biological function" in the Unified Medical Language System (UMLS) semantic network from the U.S. National Library of Medicine:[58]

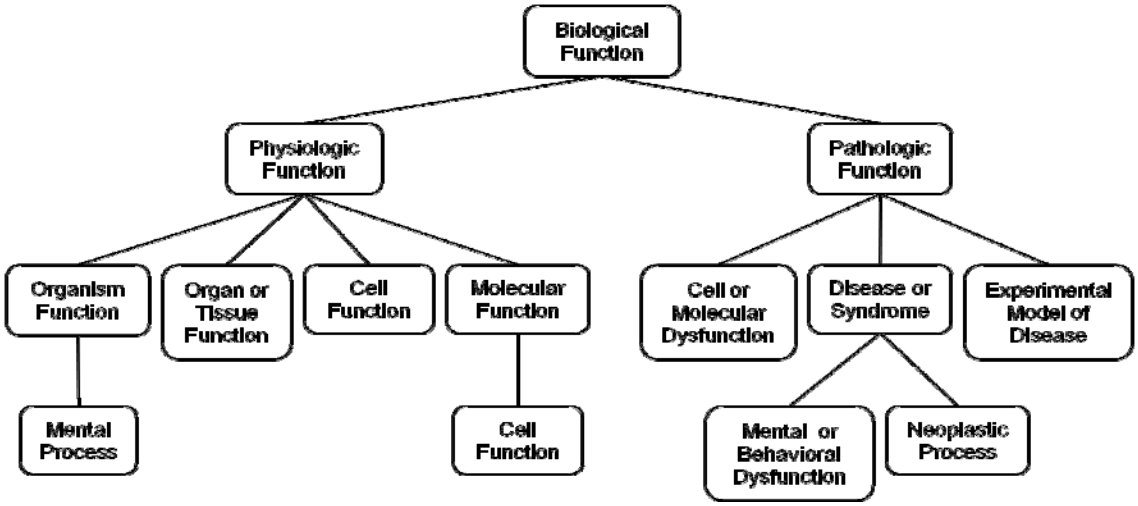


Figure 7: A hierarchical semantic network

"Biologic Function" is the semantic concept at the root of the hierarchical tree. It has two children, "Physiologic Function" and "Pathologic Function", each of which has children and grandchildren. In this representation, each child node is linked to its parent node by an "is a" link. For example, an "Organ or Tissue Function" is a "Physiologic Function", which in turn is a "Biological Function". (Copyright U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894)

In the example above, the relationships among nodes are hierarchical. Many non-hierarchical relationship types exist, and can easily be depicted in semantic networks. The UMLS, for example, identifies the following five major non-hierarchical categories or

relations for use in the semantic network: *physically related to, spatially related to, temporally related to, functionally related to,* and *conceptually related to*.[58] Other authors have identified other categorization schemes for links. Quillian, for example, proposed the following categories: *subordinate ("is a") and subordinate links, modifier links, disjunctive links, conjunctive links,* and *everything else,* but usually relationships that are best expressed as verbs, which are concepts themselves.[57] The UMLS specifics a semantic network of functional relationships used to label connecting arcs between nodes. Figure 8 gives an example of a partial hierarchy of the UMLS semantic network for the relationship "Affects":

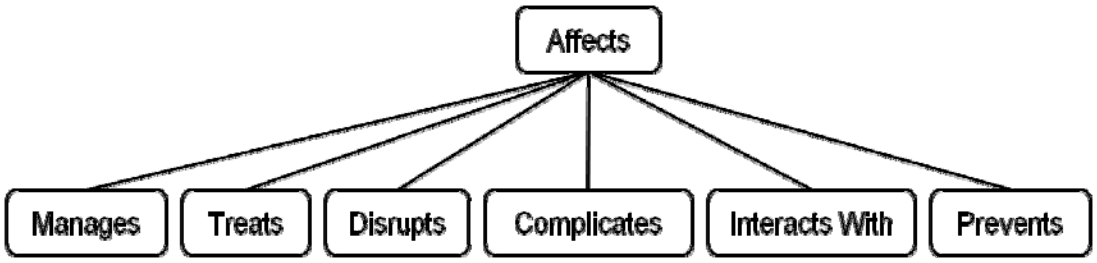


Figure 8: A hierarchy of network relationships

This figure depicts the UMLS functional relationship concept "Affects." If the relationship between two nodes is that one node "Affects" another, the relationship can be further specified by one of six types of affects. For example, anticoagulants can be used to *treat* pulmonary embolism. These functional relationship concepts are used to label links between nodes. (Copyright U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894)

Semantic networks can become very complex very quickly. Every concept can have multiple relationships with other concepts, resulting in complex graphs that are difficult

to interpret. Figure 9 shows a very small portion of the UMLS semantic network around the concept of "fully formed anatomical structure," showing some of the related concept nodes and their functional links:

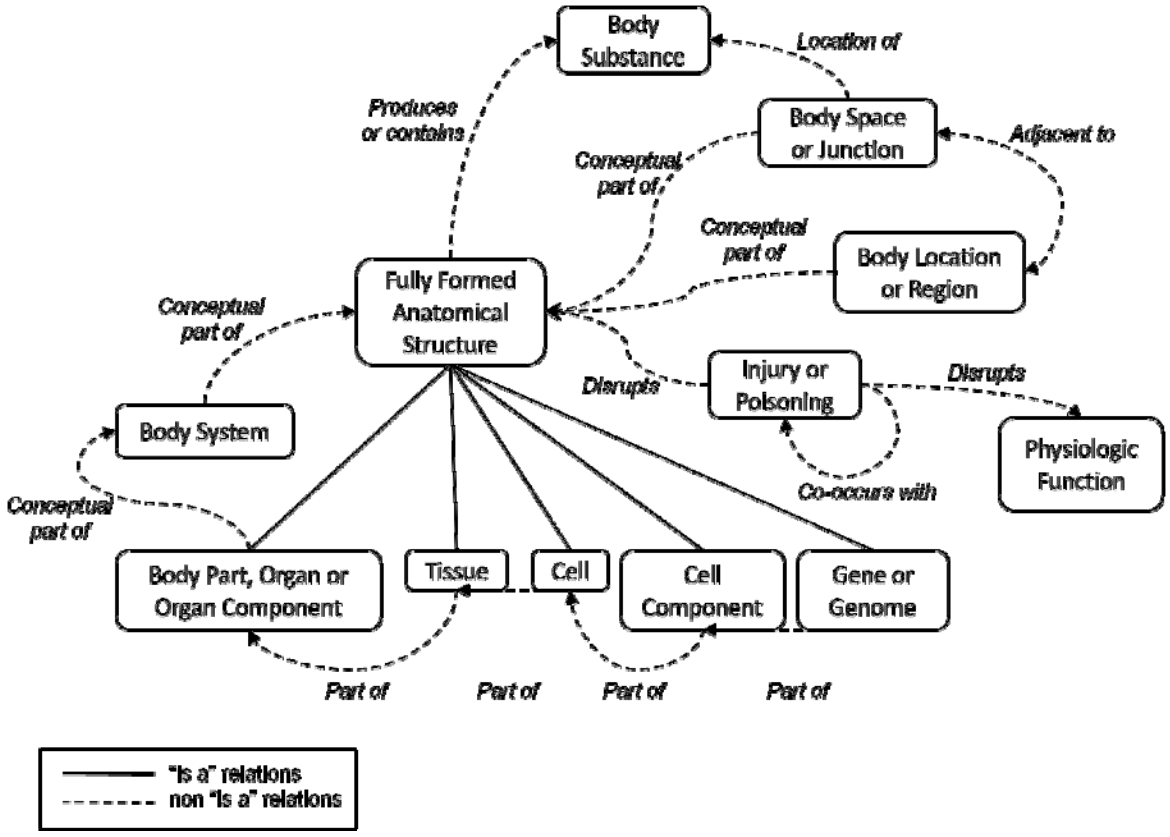


Figure 9: A semantic network of a fully formed anatomical structure

This figure demonstrates a small portion of the UMLS semantic network related to the concept of a "Fully Formed Anatomical Structure." Relationships between this concept and others in the network are represented by solid lines (for "is a" relationships) and dashed lines (for other types of functional types of relationships). For example, an injury or poisoning disrupts a fully formed anatomical structure. Injury co-occurs with

poisoning, which can disrupt physiological function. (Copyright U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894)

2.4.3. Spreading Activation Theories

Foundations

Spreading activation theory emerged in the early 1960s as researchers first began to develop computer simulations of reading comprehension. Quillian was the first to publish this theory as part of his 1966 doctoral dissertation.[59] Because the theory was developed to motivate computer algorithms, it required specification of knowledge structures amenable to digital processing. As a result, the theory relied on semantic networks, which were relatively easy to implement computationally. It is important to note that in early simulations spreading activation theory was intended to describe how a computer might "comprehend" reading material, and the theory was therefore not generalizable to the broader realm of the human cognitive processes the computer was attempting to emulate.

In very general terms, with the semantic network as a foundational structure, spreading activation theory proposes that as words are encountered in text they are mapped to concepts. Once this word-level concept mapping is complete, the mapping activates the node in the semantic network corresponding to that concept. Activation then spreads to all of the concept's linking nodes, then to nodes linked to linking nodes, and so on.[57, 59] To understand more than a single word, activated concepts for each constituent word in a phrase must somehow converge on a unified meaning. Quillian suggest that as

subsequent words are “read”, the single word encoding process repeats, resulting in *parallel activation* of the networks related to newly mapped concepts. The waves of activation cascade through the network along links until they converge on a single concept or cluster of concepts, which represents the combined meaning of the words.[57, 59] Thus, concept convergence represents the solution to a memory search for meaning.

When Quillian first introduced the model, he stated that an encoded word could be represented by one of two types of concept (or, meaning) nodes. *Type nodes* point directly to a configuration of other nodes representing the meaning of the encoded word. *Token nodes* point to type nodes via special associative links and represent indirect mappings to the concept in the type node. Token nodes make it possible to construct concepts, as in building the meaning of a word from the combined meanings of other words. This process can result in the creation of a new type node configuration, if this action is cognitively efficient. For any word, there will be one and only one type node for the concept, and any number of token nodes.[59]

The links between nodes are sophisticated as they must be able to represent any kind of relationship the nodes share. As previously discussed, Quillian identifies five kinds of links (“is a”, modifiers, disjunctions, conjunctions, and everything else) that serve to support the computational relationships among nodes he instantiated.[i] Importantly, these links are themselves concepts and can be as nested, deep, or embedded as necessary to represent a concept in a network. Thus, the full meaning of any word is represented by the network of nodes that can be reached starting with the matching type node for the

[i] Although this list is limited, it is extremely useful for computational network traversal, as it represents basic symbol manipulation logic.

concept and tracing in any direction to any level in the network.[57] This is particularly useful, because by tracing the paths to the convergence backwards to the points of initiation of the activations, the paths can be evaluated to determine if they satisfy constraints imposed by the memory request. For example, in a reading comprehension task, the constraints might be that a certain syntax or context applies to the search solution. Should one of the paths to the solution fail to meet those constraints, the path is discarded as unacceptable.[ii] Finally, when multiple paths emerge for identifying a concept, it is the sum of the criterialities of the links in the path that determines which path is better; the path with the higher sum is the best.[59]

It is worth repeating that the original theory motivated computer algorithms for modeling reading comprehension. As such, spreading activation theory, at least in its original form, was never meant to be a general theory of cognition (or even of reading comprehension). Indeed, Quillian, the originator of the theory, offered at the time that his theory might not be physiologically realistic.[57] Despite this stated concern, the theory has provided a remarkably robust foundation for the development of the theory of reading comprehension and concept identification. We introduce two of the major derivatives of the original theory here.

Extended Spreading Activation Theory

In 1975, Collins and Loftus[57] extended the original spreading activation theory to explain the results of several experiments in semantic processing. In an excellent article discussing the strengths and weaknesses of the original model proposed by Quillian, the

[ii] Despite the fact that this is a computational issue, it is likely that multiple cognitive paths undergo some similar evaluation to eliminate illogical or improbable interpretations of meaning.

authors sought to clarify the theoretical foundations of the earlier model and to better detail its assumptions about memory structure and processing and semantic matching processes. Specifically, the authors sought to adapt Qulilian's work to humans by aligning the theory with what they called "quasi-neurological terms, a la Pavlov."(p. 410)[57] Their model extension assumptions provide a powerful framework for evaluating the how concepts are identified and matched in memory, and serve to strongly support the cognitive model presented in this manuscript, so I will elaborate them here.

The first four assumptions of Collins and Loftus' extended spreading activation theory are as follows:

1. Activation spreads outward through a network in a decreasing gradient. Once a concept is encoded, the decrease in activation as it travels along network paths is inversely proportional to the strength of the nodes in the path.
2. The longer the concept is attended to (e.g., is being read, heard, rehearsed, or processed), the longer the period of time that activation is propagated from that concept's node. The rate of activation is always fixed; more attended concepts simply fire for longer periods of time. Activation can begin with a single node at a time. However, the spreading process operates in parallel through the network after the initial activation fires.
3. Activation decreases over time or if attention is interrupted or diverted. This suggests that activation levels are variable.
4. Intersections among network paths have an activation threshold that must be met before the intersection can be evaluated (e.g., to see if it complies with situational constraints or provides the best solution for a meaning search). It is assumed that

the activations from all paths leading to the intersection must sum to greater than the intersection's activation threshold for the intersection to activate.[57]

The second set of assumptions made by Collins and Loftus are concerned with the notion that semantic memory is organized by noun categories or names of things, and that it contains a lexical dictionary network that is distinct from the semantic network:

5. The semantic (conceptual) network is arranged by semantic similarity. When two concepts are highly related (e.g., they share many properties), the nodes share many links to one another via these properties.
6. A lexical dictionary stores the names of concepts. The dictionary is organized by phonemic similarity such that links among nodes in the lexical network carry the phonemic properties of the concept name as well as their relative position in the word. Nodes in this lexical network connect to nodes in the semantic network.
7. A person can control whether they access the lexical network or the semantic network. For example, one can decide to locate words that sound like "book" (e.g., accesses the lexical network), or think of concepts related to "book" (e.g., access the semantic network) or think of words that correspond to the concepts related to book (e.g., access both networks).[57]

Finally, Collins and Lofus offer 6 remaining assumptions regarding the semantic matching process; that is, how we determine if two concepts show semantic equivalence:

8. In order to determine if two concepts are semantically equivalent, there must be enough collected evidence to exceed some criterion (either negative or positive). This evidence accumulates through encounters with intersections found during

memory searches. As we have discussed, the criterialities of path links sum together, and can be thought of as the evidence present at an intersections. When an intersection is reached and there is not sufficient evidence (e.g., summed activation) to fire the activation threshold, whether positive or negative, a "don't know response" is generated.

9. Superordinate connections (whether positive or negative) between nodes can surpass other criteria for determining if concepts match or not during memory searches. For example, if there is a superordinate link between "Tucker" (my dog) and "border collie" and another superordinate link between "border collie" and "dog", this represents conclusive evidence that "Tucker" is a "dog". In the same vein, if a negative superordinate link exists between "Lump" (my cat) and "dog", this represents conclusive evidence that Lump is not a "dog."
10. Most of the criterialities of the properties of two nodes must match in order to determine that two concepts are semantically equivalent. However, as we have seen, criterialities are asymmetrical and weighted. If a highly critical property is missing in one node, this can result in a negative decision about the match, because this critical property will carry a high negative weight, which may cancel other, positive criterialities.
11. If we want to determine if two concepts are similar, the two concepts share many common properties, and one of the concepts has a superordinate, this constitutes positive evidence of a match between X and Z. By way of example, Loftus offers the problem of determining if a stagecoach is a vehicle. To make this determination, a person might compare a stagecoach to a car, which is clearly a

vehicle. Because a stagecoach and a car share many similar properties, and both are vehicles, the response is positive.

12. If two concepts share a superordinate with mutually exclusive links to the two concepts, this constitutes negative evidence. To illustrate, consider if one wants to determine if a mallard is an eagle. Mallards are ducks, which are birds. Because eagles are birds, mallards and birds share a superordinate. However ducks and birds are mutually exclusive kinds of birds, providing conclusive evidence that a mallard is not an eagle.
13. Finally, counterexamples represent important negative evidence. Consider a statement of the form "all Xs are Ys". If one can provide an example of a condition that does not fit (e.g., one example of a case where an X is not a Y), this provides conclusive evidence that X is not always a Y.[57]

These assumptions greatly elaborate what is happening at the intersections posited by Quillian in his original theory, and began to expand the theory of spreading activation to encompass human cognition in reading comprehension. Importantly these assumptions added more specifications of the various reasoning processes that take place during memory recall and integration. Loftus and Collins were investigating these issues at the same time Anderson developed Active Control of Thought Theory, which we discuss next.

Active Control of Thought (ACT) Theory

Anderson[60] introduced Active Control of Thought (ACT) theory in 1976 to apply spreading activation theory to the processes of reading comprehension outside of the

computational domain. This work was largely coincident with the work of Collins and Loftus who were also expanding the original model for its applicability to humans. To build his theory, Anderson introduced the notion of a *cognitive unit* structure, consisting of a unit node and an associated set of about five elements, as the functional unit of encoding and retrieval in reading comprehension. In a generic cognitive unit, the unit node is a proposition, and the elements are the relations and arguments of the proposition.[61] Anderson believed that the number of elements associated with a unit node was physiologically limited. As he suggested, "…it is reasonable to consider a paired associate or simple sentence to be encoded by a [single] cognitive unit, but it is not reasonable to consider a paragraph or 30-word list encoded by a single cognitive unit."(p. 262)[60] Interestingly, Anderson estimated that five elements comprised, on average, a single cognitive unit.[iii] Finally, Anderson proposed that cognitive units are organized hierarchically because propositions can serve as sub-proposition of another (just as other propositions can represent combinations of others, as when a single proposition summarizes many other concepts).[iv]

Anderson and Pirolli[62] make a strong point of differentiating between declarative and procedural knowledge in the description of the ACT theory. They emphasize that spreading activation processes apply only to the declarative knowledge encoded in the long-term memory network. Procedural knowledge, on the other hand, is represented as

[iii] This limitation of about five elements corresponds with other research into working memory capacity.
[iv] Anderson argues that there are many different types of cognitive units, with images and temporal strings of words as primary examples. These cognitive units may not contain propositions at the unit node level, though their form still consists of a generic unit node and its associated elements.

pairs of conditions and actions.[v] The condition part of procedural knowledge dictates what pattern must be matched by information that is currently active in declarative memory, in order for some action to take place. The action part of procedural knowledge is flexible enough to allow for additions to declarative memory (e.g., laying down new memory traces) and establishing new goals or external responses (e.g., establishing new actions based on existing conditions, context, and new, derived knowledge). The sequencing of condition matches and the resulting actions determines behavior.

Information encoding is an all-or-nothing process[60] and results in the creation and storage of a cognitive unit, either in response to external stimuli or as the result of internal computations, in working memory. ACT theory proposes that each of these transient instantiations in working memory carries some probability that the cognitive unit ultimately will be stored as a permanent trace in long-term memory. The probability is constant across experimental manipulations, in that it does not appear to vary with a person's motivation or intention to learn, nor does the probability appear to be affected by how long the cognitive unit is kept in working memory, though repetition of the information appears to increase recall. The probability of storage is directly related to the strength of the trace. A memory trace that has been successfully recalled from long-term memory has a strength unit of one. Every subsequent successful recall increases the strength by one unit. A stronger trace is more permanent in long-term memory; it is also more quickly recalled.[62]

[v] The authors do not discuss what structure these condition-action pairs take. One presumes they are organized in a semantic network.

Working memory must encode not only all of the incoming information available for processing, but also the traces from long-term memory that are being recalled for processing. As a result, there is overlap between working and long-term memory in terms of content, and activation of memory elements thus becomes a matter of degree in the current context; that is, some elements are relatively more active than others at any given point in time. Elements that are being processed in this moment are more active than those processed a while ago, as are elements that are more contextually relevant to the comprehension problem at hand than those elements that are not. ACT theory posits that once long-term memory traces are formed, they cannot be lost, so this apparent decrease in activation is explained as decay in the strength of the traces over time. Node strength is a function of how often it is activated, or exposed to a concept. This in turn determines the amount of activation its elements can emit throughout the rest of the related concepts in the network. So a node under attention will receive more activation, which in turn makes it stronger, so activation gathers in areas of the network where there are stronger nodes. In this manner the level of activation among network nodes reflects how closely the concepts they encode are related, because these related elements have been recalled more frequently that less related elements.[62] As soon as the currently active concept in working memory drops from attention, its activation begins to decay; this results in decaying activation of the associated network.

2.4.4. Semantic Priming Effects

Semantic priming is the effect through which word recognition is facilitated by exposure to semantically related concepts,[63]or, put another way, priming has taken place when

presentation of one concept speeds responses to related concepts.[64] This notion implies that a priming stimulus will result in automatic activation of semantically related concepts, and that this immediate activation will allow those semantically related concepts to be recalled faster than those less closely related. According to Pusen, et al, "nodes for concepts closely related to the prime are activated more quickly, more strongly, or with greater probability than are nodes for more distantly related concepts, and this activation results in more rapid or accurate responses on the main task."(p. 627)[65] This process is the same as a simple memory search; that is, to prime a concept one must first encode the word and search through memory traces to match the concept. The difference here is that priming is a facilitator to *subsequent search*, such that when a new concept is encoded, it will be matched in memory much more quickly if it has been previously primed. Thus, each encoded concept creates priming effects. Whether or not these effects facilitate recall depends on the relatedness of the encountered concepts.[57]

Consider the example of a subject reading the word *grass.* Priming effects indicate that some portions of the subject's internal semantic representation of *grass* are activated. Thus, concepts closely related to *grass*, such as *green, lawn, yard, meadow, pasture, golf course,* or even *marijuana* may be activated, depending on the subject and context. As a result, if the subject encounters related topics in subsequent test, the subject will more readily recall these associated concepts and interpret their meaning than if the subject sees less related concepts (to the subject, anyway), such as *dew, prairie, football field, lawnmower*, etc.

The priming effect is usually assumed to be automatic, though some conscious control of the process may occur when there is a long interval between exposure to the prime and the stimulus.[65] As previously mentioned (see section 4.5.2.) Collins and Loftus[57] assume a person can actually control whether they access the lexical network or the semantic network, though the authors do not discuss this as a factor of lag time between stimulus and response, nor do they explicitly relate this notion to priming effects, *per se*. This presents an interesting point. If a subject encounters a term they do not understand, it is difficult to assume that any immediate priming can occur, even with a conscious lexical lookup. Consider the surgical procedure *dacryocystorhinostomy* (a procedure to facilitate tear duct drainage in the eye). To the lay person, this term may likely prime very little. Perhaps *rhino* may prime *nose* (if the subject has some Latin education or medical abbreviation training), though it may just as easily prime *rhinoceros.* Regardless, the primed portions of the network are not likely to be useful for ascribing meaning to this particular concept.

2.4.5. Word-Frequency and Word-Length Effects

The word-frequency-effect describes the fact that are able to access more rapidly words that are more frequently used in language than words less frequently used. Similarly, word-length effects demonstrate that shorter words are recalled faster than longer words.[53] We also more rapidly identify words or phrases that make sense in the current context as opposed to less likely matches (e.g., if I say sweep, a word like broom is more likely to make sense and be expected in context, rather than, say, calendar). In addition, research suggests that we maintain some internal representation of word features that

constrain the semantics of a word in context, and that these features are arranged in semantic networks, to provide ready paths to pertinent, associated information as needed.

2.5. Theories of Specific Cognitive Tasks

2.5.1. Introduction

Research into the cognitive processes associated with cognitive tasks is wide and varied. For the purposes of this proposal, we are concerned primarily with the tasks of reading and coding clinical documentation. As a result, we focus on reading comprehension, categorization and pattern matching, and concept identification. In all of these areas we consider the impacts of domain expertise and skill on these tasks.

2.5.2. Reading Comprehension

Introduction

Reading comprehension can be understood using two distinct, though complementary approaches. The first method focuses on the processes of reading, that is, the set of operations that must be completed for reading comprehension to occur. The second approach focuses on the development and exploration of cognitive theory to explain how the brain internally represents information as it is processed.[53] This section briefly reviews each of these approaches and their contributions to our understanding of reading comprehension.

Reading as a Series of Processes

When viewed as a process, the reading comprehension task consists of a series of steps beginning with reading the written words on the page and ending with new knowledge (e.g., understanding what those words mean). Each of the processes in the series can be described by the following characteristics: 1) the information in the text at the start of the process, 2) how long it takes to read and comprehend that information, 3) what other information is used during the comprehension process, 4) sources of error during the process, and 5) what the reader has learned once the process is complete.[53] Graphically, progression through these processes generally follows these steps:

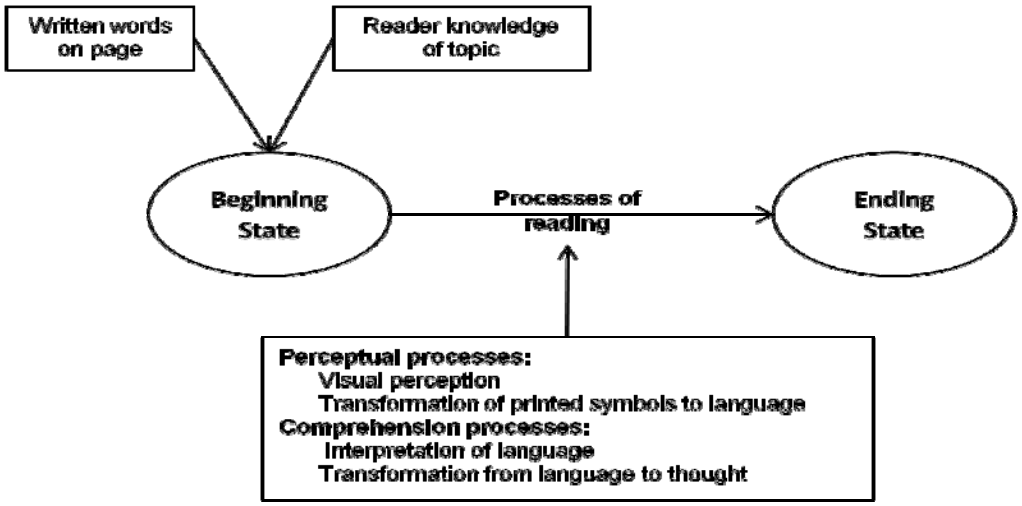


Figure 10: Steps in reading comprehension

At the start, the reader is presented with written words on a page. In addition, the reader brings his or her knowledge of the topic, as well as general world knowledge. During reading, the reader must visually perceive the text, convert the symbols on the page to their corresponding language elements, interpret the meaning of the language, and finally

transform language to its more abstract representation as thought. When reading comprehension is complete, at the ending state, the reader has acquired new knowledge.

The act of reading involves processes of perception and comprehension. For a reader to be successful, he or she must first possess the physiological ability to sense (e.g., perceive) the ink markings on a page. To comprehend what these marking mean, the reader must be able to correctly identify and aggregate collections of symbols and transform these collections to meaningful units, such as words, that represent language. The reader must then abstract meaning from these language elements. These steps involve both perception (for example, being able to identify white space to note when a word has begun or ended) as well as comprehension (being able to ascribe meaning to the perceived words). In the classic view, comprehension functions to *construct* a representation of the meaning of the linguistic input. However, comprehension can also be described in terms of *utilization*; that is, how the reader will make use of the information read (e.g., will the user store it in memory for later, believe the information is true or not, do something if the text provided instruction, answer a posed question, etc.).[66, 67] The cognitive processes of construction and utilization are not clearly separable, nor are they directly observable, so researchers evaluate naturally occurring behaviors during reading to gain insight into how the brain might work as reading takes place. One particularly useful method has been tracking eye movements while reading.

Tracking Eye Movements to Understand Reading Processes

As early as 1879, observational studies of eye movements revealed that the eye does not travel smoothly along a line of text during reading. Instead, it was noted that the eye

pauses or fixates on certain words then jumps from one location to another as the text is read. As experimental methods and technology advanced, researchers were able to precisely track and evaluate eye fixations and their durations, eye movements between fixations, and the text being fixated upon. These experiments resulted in several general discoveries about reading. First it was noted that if a reader was fluent and the text to be read was not difficult, most fixations were in the forward direction, moving from earlier words to later ones in the text. If the reader had trouble with the text (e.g., it was too difficult, or the reader was not fluent in the language) the reader often made fixations backwards in the text before moving on to subsequent text. This suggested that readers fixated longer on words they didn't know and had difficulty interpreting. As a result, eye fixation studies came to serve as proxies for identifying the psychological processes of reading.[53]

Gazes are consecutive fixations on the same word. Most college-level readers gaze on a high proportion of words in text: upwards of 80% of all nouns, adjectives, verbs and adverbs – the so-called "content words." In contrast, these same readers fixate on a smaller proportion of lower information words such as articles, conjunctions, and prepositions (e.g., *the*, *and*, *of*) – the so-called "function words." Overall, about 65% of the words in a given text are fixated. If the text is difficult for any reason, the number of fixations increases. Other text properties correlate with fixation times as well. For example, readers spend less time on topically related words when the readers are familiar with the general concept of the text. Thus, an airline pilot reading an article about a plane crash will likely fixate for less time on the word *aileron* (the moving flap on an airplane wing that controls turns) than would a person with little experience with planes or flying.

In addition, readers fixate for longer periods of times on words that are syntactically unexpected, suggesting that readers must pause to make sense of the unexpected word. For example, consider this sentence:

The defendant stood before the judge entered the courtroom to convene the trial.

Most readers commonly interpret the phrase "The defendant stood before the judge" to mean that the defendant stood in front of the judge. It therefore comes as a surprise when the reader comes to the word "entered," which does not seem to rightly follow. The reader must pause to make sense of the remaining text indicating the defendant was not standing *in front of* the judge, but stood up *before* the judge entered the courtroom. In this case, the reader will fixate longer on the word "entered" than in controlled trials where the word "entered" is used but no ambiguity exists in the sentence, suggesting that resolving the ambiguity takes imposes additional cognitive load.[53]

Language Levels and Associated Processes

Reading is the act of comprehending written language. As we have stated, to read we must first learn to identify words from the printed symbols on the page. This is the lexical level of reading, where unique combinations of letters are identified and cognitively combined as functional units. Once individual units are parsed, they must be mentally encoded and assigned meaning through lookup from a mental dictionary in a process known as lexical access. These two processes – encoding the word and accessing its meaning -- result in word recognition. Understanding the meaning of a single word obviously does not enable a reader to comprehend a complete text, nor does it allow the reader to make sense of that word in context. For this reason, the word must pass syntactic and semantic analyses. Syntactic analysis informs the reader if the word is

correctly used according to the rules of grammar; that is, whether the word is properly spelled, is of the correct tense, demonstrates proper subject/verb agreement, shows up in the proper order, etc. Semantic analysis evaluates the meaning of the word, and whether or not it makes sense in context.[53]

Historically, it was assumed that lexical parsing and interpretation preceded semantic parsing, which in turn came before integrative thought. Under this view, each process comprises a separate cognitive module, with the output of one module providing input to the next one in sequence. More recent research suggests that the combined processing is more likely interactive; that the processes are interactive and operate in parallel. Semantic processing may inform syntactic processing and vice versa.[66] It has been shown that college-educated humans read at very rapid rates of approximately 240 words per minute[53] suggesting that readers exhibit an immediacy of interpretation: that all levels of processing take place very close in time as each word is parsed. This has been borne out by recent research demonstrating that the meaning of lexical items is active by about 200 ms after signal input, as demonstrated by imaging studies that show brain activation in response to stimuli.[68] This immediacy of interpretation contrasts with the notion of reading as a "wait-and-see" task, where a section of text is read in full and then processed. As described in the section on eye movement tracking, above, the immediacy of syntactic processing can be explained by the longer duration eye fixations on unknown or infrequent words, suggesting that readers do not move on in text until the fixated word can be successfully parsed. These experiments also support the notion of immediacy in semantic processing. Sentence meaning appears to be computed on a word-by-word basis, such that each new word is integrated into a partial but flexible semantic

representation as the sentence is read. In eye movement experiments, the fact that readers fixate for longer periods of times on words that are inconsistent with the surrounding context suggests that readers are attempting to make sense of the whole sentence as it is being read.[66]

Activation

Although I have discussed spreading activation theory (see Section 4.5.), it is worth returning to this topic here to tie it to the process of reading comprehension, since the theory provides a strong foundation for models of this process. Using spreading activation theory, the mental structures involved in reading comprehension are assumed to possess activation levels that roughly correspond to how accessible or credible the structure (or concept) is at a given time. Under this view, a concept's activation level changes in response to the current knowledge state; that is, it is either active or not, and the level of activation varies along a continuum depending on if the trace is in decay. Very active concepts can arise because concept has been primed by earlier words in the text, because the concept is necessary to aggregate related concepts (as when making sense of a group of words, a sentence, or a paragraph), or the concept is the current focus of attention (as when a reader is contemplating a word). It is only when the activation level of the concept reaches a certain threshold that it is accepted as representing the text. Should two different words in the text activate the same concept, the threshold for activating that concept should be faster than if a single word preceded activation.[53] In addition, returning to the assumptions for spreading activation theory expressed by Collins and Loftus,[57] activation levels can indicate the degree of confidence a reader has

in textual inferences, and how a reader might decide they are sure, not sure, or don't know what text means.

2.5.3. Categorization and Pattern Matching

Introduction

Categorization, simply put, is the process of assigning objects to classes or categories. A *category* consists of objects which are considered equivalent in some manner. Categories usually have names (e.g., dog, tree, animal, plant, etc.) and are related to each other in *taxonomies* based on levels of class inclusion such that higher level, or more abstract categories subsume more specific categories.[69, 70] Lakoff describes categorization as fundamental to virtually all human activities:

> "There is nothing more basic that categorization to our thought, perception, action, and speech. Every time we see something as a *kind* of thing, for example, a tree, we are categorizing. Whenever we reason about *kinds* of things—chairs, nations, illnesses, emotions, any kind of thing at all—we are employing categories. Whenever we intentionally perform any *kind* of action, say something as mundane as writing with a pencil, hammering with a hammer, or ironing clothes, we are using categories…Any time we either produce or understand any utterance of any reasonable length, we are employing dozens if not hundreds of categories: categories of speech sounds, of words, of phrases and clauses, as well as conceptual categories. Without the ability to categorize, we could not function at all, either in the physical world or in our social and intellectual lives. An understanding of

> how we categorize is central to any understanding of how we think and how we function, and therefore central to an understanding of what makes us human."(p. 5-6)[71]

The Evolution of a General Theory of Categorization

The traditional view of categorization is *objectivist;* it holds that the mind functions much like a computer, manipulating abstract symbols (our internal representations of external objects) through algorithmic computation. Objects derive meaning only inasmuch as they correspond directly to things in the external, physical world. And to belong to a category an object must share features in common with all other members of the category. Categories therefore possess distinct boundaries, defined by the common properties of its members. The objectivist view assumes that categorization is a purely objective process, that there are no other factors that might affect categorization, such as our ability to form mental images, the effects of learning and recall, our capacity to perceive, human neurophysiology, and other "peculiarities" of the human mind or body. Thus, under this view, categories are independent of human characterization, and are therefore disembodied; they simply exist in the real world, and represent a single true view of the world.[71]

The objectivist view also holds that thought processes are atomistic; that is, they are composed of simple building blocks (the abstract symbols corresponding to the external world) which are aggregated as necessary into more complex arrangements. These simple building blocks (and their aggregate forms) are manipulated using the formal mathematical rules of traditional deductive logic. Reason is therefore mechanical in

nature and has no bearing on how we think; it is concerned only with the relationships, whether inferred or real, among all possible concepts in our universe. Such a view suggests that there is no difference among people in terms of the conceptual system they use. Under this view, all of us utilize the same classification systems; emotion has no role in classification because it lacks conceptual content. In addition, the objectivist paradigm suggests that the mind and the body are not only separate but independent, and the way we physically interact with the world has no bearing on how we categorize.[71]

The traditional objectivist view is supported by surprisingly little empirical evidence, and up until the mid 1950s was more or less accepted without debate. Wittgenstein was the first scientist to formally suggest that there were many categories that did not meet the objectivist definition because category members often did not actually share identical properties. He illustrated this notion with the category of "game." He noted that not all games involved winning and losing (that is, some are simply played for amusement), nor do all games require skill; many required only luck, though some games require both skill and luck. In addition, many games required additional objects (game boards, playing cards, balls, etc.) to support whereas others require none. To explain these interesting distinctions, Wittgenstein introduced the notion of family resemblances to explain how category members could actually share a wide variety of features that are similar in many different ways, but not necessarily identical. Thus category membership could be based on family resemblance as opposed to a well-defined set of common features. Furthermore, using these family resemblances, boundaries between categories could no longer be considered immutable because objects in one category could bear family resemblances to objects in other categories.[71]

Wittgenstein's view was the beginning of the notion that category membership could be based on *prototypes*: objects sharing collections of features that bore family resemblance. This observation began to explain why some members of a group could be considered "better examples" of the category than others. (For example, when thinking of the category "bird" one might be more likely to think of *robin*, *eagle*, or s*parrow*, rather than, say, *penguin*. The categorization depends in part on the observer and his or her determination regarding the shared features of the members. In the case of the "bird" category, one might think that birds are creatures that fly. Although a penguin is a bird, since it does not fly, one might not think a penguin is the most representative example of the category "bird.") In addition, the prototype view offered a plausible explanation for how categories might have gradations of membership, that is, how members might rank in comparison to other members as in categories like "people who are poor," or "dwellings that are big." In these cases, each category could be arbitrarily large with membership dependant on grade. It was precisely this notion that lead Zadeh to develop fuzzy set theory in the mid 1960s. No longer was an object either inside or outside a given category; it could exist on a continuum, without discrete boundaries, from "not in" to "in" based on its features.[71]

What emerged from this new approach to categorization was "prototype theory", the central dogma of contemporary research into categorization, or what Lakoff has called "the new experimental realism." (p. xv.)[71] This view posits that cognition is embodied, that is, it is inextricably bound to our perceptive abilities, bodily functions (movement, physical experience), character, and culture, and that we can only make sense of our

environment though these bodily experiences. In addition, cognition uses imagination in the form of metaphor, mental imagery, and metonymy (a figure of speech where a name of an item is replaced by one closely associated with it, as when "dough" is used in place of "money"), as imagination permits us to form conceptual models or categorization schemes of things that we cannot experience directly with our senses. Thought also demonstrates gestalt properties that demonstrate an overall structure much more complex that simple atomic concepts aggregated into higher order structures. Finally, cognitive processes demonstrate an ecological efficiency which cannot be fully explained in terms of algorithmic computation using abstract symbols. Because categorization is fundamental to cognition, the new view purports to explain much of cognition in terms categorization alone.[71]

There is a wealth of research into the cognitive processes of categorization. For an excellent overview of the history of categorization theory and research, the reader is referred to Lakoff (1990). For the remainder of this section, we discuss those theories that pertain most directly to the cognitive model of medical document coding that we will introduce in a later section.

Basic Level Categorization

In 1958, Roger Brown published a classic paper "How Shall a Thing Be Called," in which he first posited that humans name things "…so as to categorize them in a maximally useful way." (p. 20)[72] Brown based his observations on how adults teach children language noting that adults generally tend to teach shorter words before multisyllabic ones, and single words before phrases. In addition, adults usually call a

thing its most common name (e.g., what we tend to think of as an object's natural or real name). Thus, an adult would teach a child the word "dog" before the more specific "cocker spaniel" or the more general "quadruped." Brown considered this to represent a basic if not universal level of categorization and suggested that it had several converging properties. First, items at this level are clustered by the nonlinguistic actions that correlate with them (e.g., dimes can be used to buy things, flowers smell nice and can be picked, cats purr and can be petted). This level is the level at which we first learn and how we first name things. The names used at this level are usually short and are used more frequently. Finally, this level of categorization seems natural in our universe; it does not require effort or imagination to derive.[71, 73]

Prior to Brown's time, much of the psychological research around the topic of categorization assumed that categories were taught to children, implying that the world was "segmented" in a purely arbitrary fashion, and that categories had to be learned. This is perhaps why Brown's paper, in 1958, focused on the learning aspects of categorization. By the early 1970s, Rosch[69, 70] began to argue that categories were not entirely learned. She reasoned that because attributes of objects do not occur independently of each other, there was a natural and intrinsic separation among things. For example, she noted that animals with feathers are more likely to have wings than animals that don't, and objects that visually look like chairs are more likely to function to be sat upon than other non-chair objects such as cats. Rosch argued that there was a basic level of categorization that people naturally used and that this level was the one carrying the most information, enabling it to be most easily differentiated from categories. Rosch

suggested this basic level of categorization offered both the greatest cognitive economy and cue validity.

In Rosch's terms, cognitive economy is the result of biological efficiency and evolutionary edge. She suggests that although it might seem beneficial for an organism to be able to process every incoming stimulus to differentiate it from others and subsequently to have a wide range of very fine–grained categories, this approach might require too much cognitive overhead if it is necessary to categorize all stimuli simultaneously, especially those that were irrelevant in the given context. An ability to categorize what is "behaviorally and cognitive usable" (p. 384)[70] in the moment, Rosch reasons, is the most efficient cognitive approach, and therefore results in the greatest practical advantage for the organism.[70]

Cue validity refers to the probability that a feature is predictive of a category; a cue validity of 1 implies that a feature is 100% predictive of category membership, whereas values approaching zero imply the feature is increasingly less predictive. The sum of all of the cue validities of the features of category members determines the cue validity of the overall category. So, categories with high cue validities (e.g., approaching 1) are much more highly differentiated from other categories with lower cue validities. Rosch argues that basic level categories demonstrate the highest cue validity: She bases this on the underlying assumption that "…in the real world information-rich bundles of perceptual and functional attributes occur that form natural discontinuities, and…basic cuts in categorization are made at these discontinuities." (p. 385)[70] Categories at a higher level in a taxonomy naturally have lower cue validity than do those at lower levels

because the more abstract categories share less common features. Subordinate categories have lower cue validity because they share so many common features with other subordinate categories. Basic level categories, then, represent the most information-rich groupings of things.[69, 70, 74]

Other researchers have demonstrated this notion. Berlin,[75, 76] investigating how the Tzeltal people of Mexico expressed groupings of objects, discovered that groupings manifested as kinds in nature (e.g., plants and animals). He noted that the terms used for categories generally occurred at the level of genus. That is, when a tribe member was asked to identify a tree for the researcher, the individual would refer to it more commonly by genus name (e.g., maple tree) as opposed to any of a number of higher levels of abstraction (e.g., leaf-bearing tree, tree, or plant). Equally important, the primary identification was rarely more specific that the genus level specification (maple tree). Although the tribe member might know that the specific maple tree was a sugar maple (species), or even a Southern sugar maple (variety), these more precise identifying terms were not generally used unless the person was asked to provide more specific details. Thus, in a language where "maple tree" is a concept, adults may know many terms that can be used to identify it, but will select the one most closely aligned to the genus level to name it.[71]

Most languages have been shown to have simple names for things at the genus level; these names are the ones most frequently used, and, as a result, have the greatest cultural significance. Concepts are more easily recalled at this level. In addition, things at the genus level are perceived as gestalts or holistic perceptions; that is, we are able to form

mental representations of these things. Thus, there appears to be a general human capacity for basic level grouping, suggesting that general knowledge is organized at this level. Additional training or acculturation is necessary for subpopulations of experts to be able to conceptualize and converse at more specific levels.[71] We will return to this highly important point when describing the proposed cognitive model.

2.5.4. Pattern Recognition

As we have introduced the concept, categorization is the process whereby we identify things and how we go about grouping them. Another form of categorization, pattern recognition, refers to our intrinsic conscious and unconscious abilities to note similarities in form or structure among objects, situations, emotions, etc. and to react accordingly. There is a large field of study in cognitive psychology concerned with the concept of limbic resonance, that focuses on the psychological constructs that determine how and why we exhibit certain patterns of inter-personal behavior.[77] In its most simplistic form, limbic resonance is an unconscious process whereby we filter sensory input based on our culture, environment, beliefs, and emotions. These filters color how we interpret and interact with the world. The term "resonance" suggests that we determine meaning based on pattern recognition that matches some internal, known (though likely unconscious) set of heuristics. This implies that much of our world view is based on categorization and that the level of resulting activation of previous emotional, cultural and other experiential patterns greatly affects how we categorize and subsequently determines our functionality within and our reactions to the world

For the purposes of this proposal, we are concerned with pattern recognition in text, so the features we recognize are syntactic structures. However, recognition is not purely a matching exercise. When we think of pattern recognition, we often refer to the computational exercise of identifying patterns using algorithms that can vary from "brute force" techniques (e.g. direct word, phrase, or sentence matching) to more complex approaches (e.g., neural networks) incorporating domain and contextual knowledge as well as statistical inference. It is important to note that pattern recognition may involve pure matching (and often does), but inference based on absolute matches, the context in which the matches occur, our knowledge, and experience is a much more complex cognitive process than pure matching alone. Thus, pattern matching is necessary to determine category membership, but may not be the sole criterion. We therefore consider this approach a preliminary, though necessary step supporting other cognitive processes in categorization.

2.5.5. The Role of Domain Expertise

Expertise consists of specialized knowledge (usually in a restricted domain), and includes the skills, shortcuts, mnemonics, tricks, rules-of-thumb, and other learned heuristics that allow a person to more rapidly and effectively solve domain problems than another person lacking this knowledge or this specific set of skills. It is well known that expertise develops through hard work; few people gain expertise in any area without a great deal of study and/or practice. However, a complete understanding of the cognitive processes involved in expert knowledge acquisition remains elusive. In this section, I introduce general principles of expertise development.

A skill is any ability that is acquired through training; expertise is the mastery of one or more skills in a domain. Initially, when learning a new skill, we tend to memorize pertinent facts and rules and rehearse these as we practice to develop proficiency. This process is highly conscious, and is often referred to as the cognitive stage of skill acquisition, mostly because we must actively think about what we are doing as we do it. As we continue to practice, we begin to fine tune the skill through elimination of any errors in our understanding and/or performance. In addition, we begin to form and strengthen connections among elements needed to successfully perform the skill. These processes comprise the associative phase of skill acquisition, where portions of the skill are becoming easier, but learning is still taking place. Finally, once we have performed the skill repeatedly to an acceptable level, the skill becomes not only faster but more automatic, in that less conscious effort is required to complete it. This last phase of skill acquisition is called the autonomous stage, where mastery has emerged.[78]

It has been demonstrated that development of skills necessary to coordinate the cognitive associations to complete complex tasks improves along power law principles. What this means is that improvement in skill acquisition is exponential at the beginning, but over times reaches asymptotic levels. This is not to say that continued mastery stops—there is actually no theoretical limit on the amount of improvement that can take place—only that the rate of improvement slows markedly as one becomes more expert. This implies that that continued practice can improve skills, regardless of the level of current mastery. In addition, once mastery has been achieved, we tend to retain our skills even over long periods of time, as long as there are no obvious physiological or psychological

impediments. A small amount of practice after a delay in using a skill can bring us back to mastery rather quickly.[78]

Experts tend to rely on procedural rather than declarative knowledge. This means that as skills are acquired, experts make less use of explicit facts and events and instead begin to leverage implicit memory about how to perform the task. Using procedural knowledge is assumed to develop as experts begin to recognize gestalts or patterns among the components comprising the skill. This has been demonstrated in studies across many domains. For example, expertise in bridge appears to depend, at least in part, on the possession of a large repository of stored patterns of cards and links to the appropriate action patterns in context.[79] Similarly, expert musicians appear to code and recognize the relationships between notes rather than the specific notes themselves.[80] Chess masters also have been shown to recognize arrangements of chess pieces and to respond to these patterns, rather than individually recognizing individual pieces, computing their movements, and calculation outcomes.[81] Finally, expert computer programmers appear to demonstrate better functional organization of concepts than do novices, suggesting that better concept organization facilitates improved task performance.[82] These examples demonstrate the cognitive development that begins in the associative phase of skill mastery; it is used to explain the fact that once experts become proficient at a given task, the tasks becomes easier because the expert no longer iteratively labors through explicit rules and formulaic strategies, determining what to do next, but instead begins to directly (and often unconsciously) apply procedural knowledge to the task at hand.[78]

There are several other strategies experts utilize. In general, experts utilize strategic approaches to problem resolution; experts appear to optimally organize their internal problem solving representations to suit the domain problems they encounter. These strategic solutions begin to develop during the associative phase, and become more refined with experience. In addition, experts are able to recognize problem spaces as gestalts; that is, they are able to effectively chunk portions of the problem, and recognize where chunks repeat in similar problems, so that these chunks can be reused. This correlates well with the experimental results discussed above. Finally, experts seem to be able to more efficiently store and recall information as permanent traces in long-term memory than non-experts, improving recall speed and precision.[78]

In terms of activation theory, expertise appears to result from repeated activation of related concepts. Activation of a problem resolution pathway increases the strength of comprising concept nodes and serves to increase the likelihood that this pathway will be activated again in response to similar stimuli. Thus, an expert, who has likely experienced particular domain problem many times, will likely find it easier to recall and respond to a set of stimuli with a strong, established and easily activated response. Expertise has thus created more or less permanent memory traces, with strong concept nodes and easily activated links among them.[57]

2.6. Reference Standards

2.6.1. Introduction

A reference standard is an established norm against which values can be compared.[3-12] Common examples of reference standards include the gold standard (used to value paper currency), the atomic clock (representing the best known approximation of time), and the standard meter (the distance light travels in an absolute vacuum in 1/299,792,458 second[83]). For the purposes of this study, the reference standard will be the consensus opinion of a minimum of six expert raters as to the presence or absence of information in a clinical document. In all classification experiments, the reliability of the gold standard is expressed as a combination of inter-rater agreement and reliability.[13] This section discusses this concept in detail.

2.6.2. Inter-rater Agreement and Reliability

Introduction

Judges are commonly used to rate information when the information cannot be objectively scored in a purely quantitative way, that is, as right or wrong, correct or incorrect, positive or negative, etc. This is especially true in assessing clinical documentation where the provided information often represents a subset of what is known about the patient's actual state or what went into formulating the treatment plan. In such cases, judges must evaluate the degree to which the information matches the concept or construct of interest (e.g., the patient has well controlled asthma, the patient received proper treatment for hypertension, etc.). Clearly, each judge will make determinations based on his or her knowledge, beliefs, and experience regarding the topic

under consideration. As a result, even judges deemed experts are liable to disagree on certain information details, meaning there will be variation in the information scoring task. Thus, inter-rater agreement and inter-rater reliability are rarely perfect.[8, 10, 13]

Both inter-rater agreement and reliability should approach 100% in the case of a perfect reference standard for a document coding task. If instead there is a large amount of disagreement among judges, the coding task may not be reliable and the quality of the resulting reference standard can be questionable.[6, 13] Therefore, improving inter-rater agreement and reliability (e.g., reducing inter-rater variability) is paramount to the development of reliable reference standards against which other coding methods can be evaluated.

Definitions

Tinsley and Weiss[84] define inter-rater agreement as the extent that different judges rate subjects or situations in exactly the same way. If, for example, a 5-point Likert scale were used as the rating instrument, exact inter-rater agreement would result when all judges scored all subjects using exactly the same values (e.g., all judges rank subject #4 with a 3). This differs, according to the authors, from inter-rater reliability, which "…represents the degree to which the ratings of different judges are proportional when expressed as deviations from their means." (p. 359)[84] So, high inter-rater reliability means that judges rate subjects in the same relative way, though the absolute scorings used to rank the individuals are different. For example, consider a 7-point Likert scale used to evaluate 3 subjects. If all judges ranked the subjects in order by subject number

(e.g., subject #2 is ranked highest, then subject #1, then subject #3), there is high inter-rater reliability, regardless of the absolute rankings the judges assigned to the subjects.

Stemler[10] makes a similar argument, based on his observation that measures of agreement among raters are often treated as unitary concepts and can lead to misleading and imprecise reporting. Instead of differentiating between inter-rater agreement and reliability, he groups measures under the general umbrella term of inter-rater agreement, but identifies three broad sub-classes of agreement, based on review of studies in the literature: 1) consensus measures, 2) consistency measures estimates, or 3) measurement estimates. Each of these measures varies in terms of underlying assumptions and how the resulting values are interpreted. In addition, each manifests certain advantages and disadvantages. For this dissertation, I will use the broad term inter-rater agreement, and take care to report each statistical measure, as suggested by Stemler, to summarize my data.

Consensus Measures

Consensus measures assume that "reasonable" observers should be able to agree completely on how to score information,[vi] and that if the judges agree exactly, then they share both a common understanding of the construct and a similar enough internal representation to support that understanding. Consensus measures are therefore most useful when the rated data are nominal and fall into discrete categories based on easy to distinguish qualitative categories. However, these measures also work well when a linear

[vi] In general, these measures are not always scores of information. For example, raters may be asked to judge behavior such as sports performance. Because this study investigates coding of medical documents, I restrict this discuss to information scoring, though the measures apply across the broad domain of tasks that can be scored.

continuum (e.g., a Likert scale) is used to rate a construct, and the judges can agree on the exact quantitative boundaries among levels.[10]

Consensus measures are most commonly calculated as percent-agreement amounts. This estimate adds up all cases for which the judges agree and divides this by the total number of cases all of the judges rated. In general, reliable consensus measures should approach 70% or greater.[6] This measure offers two distinct advantages: high intuitive appeal and ease of calculation. However, the results can be misleading. If the information of interest occurs very infrequently (e.g., there is a small number of true positive cases) then high percent-agreement may just represent the fact that the small number of true positive results fell at a particular location on the rating scale, and that these examples were easy to rate by all judges. In addition, the consensus process can be difficult to achieve because it assumes that judges can be trained sufficiently to demonstrate exact agreement, which may not be true, even after intensive training.[6, 10]

Percent-agreement measures of consensus can be modified slightly to work around the requirement of exact agreement. For example, in experiments utilizing Likert scales with 7 or more points on the continuum, raters may be assumed to agree if they rate information within one point of other raters. This allows the measure to approximate general agreement. However, this approach can be problematic on rating scales with very few discrete points (say, 4 or fewer) as most of the points are adjacent to one another and thus may spuriously inflate inter-rater agreement values.[10]

Finally, Cohen's kappa statistic (κ) can be used to estimate consensus among raters in reliability studies. The statistic is designed specifically to eliminate the possibility of chance agreement among raters by normalizing the rating scale such that chance agreement reduces to zero. Kappa is defined as the difference in the probability of observed agreement and the probability of agreement due to chance divided by the probability of agreement. It is important to note that a κ of zero does not represent lack of agreement among judges; it simply means there is no difference between the levels of agreement among the judges than that could be predicted by chance. Values of κ approaching zero can also easily occur among judges rating very unbalanced samples (data sets where the prevalence of true positives is either very high or very low), even when the judges are highly reliable. This also implies that κ can be less than zero, if judges agree less often than would have been predicted by chance.[10] In general, a $\kappa > .75$ equates to excellent agreement,[6, 85] though some authors consider a $\kappa > .60$ sufficient to indicate substantial agreement.[10] A $\kappa < .4$ denotes marginal agreement, and a κ falling between the two cutoffs represents acceptable agreement.[85]

Consensus measures offer several advantages, including their ease of calculation and good applicability for nominal data that cleanly map to rating scales. In addition, these measures can help identify areas where judges misunderstand how to apply a rating scale, thus motivating focus for training. Finally, high levels of consensus imply that judges are effectively providing the same information. Thus, if two judges agree to the point of consensus—that is, they know how to correctly apply the rating scale—the judges may then be treated as equivalent raters such that they can split the data and independently

work on subsequent rating examples (for the same rating exercise) without having to replicate each other's work.[10]

In terms of disadvantages, I have already alluded to the potential difficulties of interpreting the results, especially when they are mid-range (e.g., .40 <= κ <= .60) , as well as the problems of training judges to improve their agreement on use of rating systems. Aside from the time and expense of training judges to the point of consensus, forced consensus may interfere with the statistical independence of the judges' ratings which can negatively impact the validity of the resulting scores. Also, consensus estimates must be calculated between pairs of judges for all judges evaluating data; large numbers of judges can make it impossible to reach consensus. Finally, Stemler[10] suggests that consensus estimates can result in overly conservative estimates of agreement between two judges who demonstrate systematic differences in using a scoring system and cannot be trained to reach consensus. In such cases, a low consensus measure of agreement may appear to contradict a high consistency measure (see the next section), and result in confusion.

Consistency Measures

Consistency measures of inter-rater reliability, unlike consensus measures, do not assume that the judges share a mutual understanding of the construct to be measured *per se*, but instead, that the judges are consistent in their individual classifications of the data being evaluated. These measures are most useful for continuous data, although applications exist for certain types of categorical data.[10] Several statistics can be used to calculating

consistency measures, including the Pearson's correlation, Spearman's rank, and Cronbach's alpha coefficients.[10]

The Pearson's correlation coefficient (ρ) is perhaps the most popular statistic of the group, primarily because it is relatively easy to calculate and allows for evaluation of continuous scores (that is, it can deal with intermediate values on Likert scales, such as a 3.5 on a five-point scale of 1-5). This statistic does rely on normally distributed data. Values for this score fall between -1 and +1. If resulting value is greater than zero, this indicates positive correlation. Negative values denote negative correlation and zero values indicate no correlation. The major disadvantage of this statistic is the requirement that the data are normally distributed, thus, if the data are skewed, the measure of correlation is suspect, and other, non-parametric statistics are indicated.[10] As with consensus measures, correlations must be calculated between two judges at a time for the same task.

Spearman's rank coefficient is a non-parametric statistic useful for approximating a Pearson's correlation coefficient when the data are not normally distributed. Usually, when this statistic is used, judges have made ordered ranking of items, as in rating from best to worst, most to least complete, etc. To compute Spearman's rank coefficient (r_S), the raw scores from each judge are converted to ranks, and the differences between the ranks of each observation by the two judges are computed.[85]

Finally, Cronbach's alpha is used when the researcher desires to find a single consistency measure across multiple judges rating in a study. The major problem with this statistic is

that all judges must rate all data or the statistic must be computed on only the subset of data that was rated by all. Cronbach's alpha typically increases when the correlations between the judges increase, which is why it is a good measure in internal consistency.[10]

Consistency measures of reliability offer three major advantages. First they eliminate the need to train judges up to a level of consensus, and require only that judges are consistent in using the rating scales by his or her definition of the scale. In addition, these measures can produce a single measure of consistency across multiple judges, as in the case with Cronbach's alpha. Finally, these measures work well with continuous data. Consistency measures also afford some disadvantages, the most important of which is that it may not always be desirable to allow judges to "agree to disagree" particularly when exact agreement is indicated. Also, since judges can differ in more ways than just in the scores they assign, such that these statistics will not elaborate the differences. Finally, many consistency measures are sensitive to data distributions. For example, if rankings or scores fall into only a few categories, then the correlation may be different than expected due to low variability in the data. For these reasons, researchers are generally cautioned to report consistency measures only in concert with other measures to provide a more complete picture of overall inter-rater reliability.[10]

Measurement Estimates

In general, measurement estimates help to reduce the number of variables for analysis while elaborating the relationships between variables in an unbiased way.[86] These estimates make use of all of the scorings (including variations) from all judges under the assumption that using all available data leads to the most reliable and informative scores.

Under this approach, judges don't have to reach consensus on how to use scoring systems, because the methods allow for estimation and explanation of how judges determined their scores. In addition, these estimates allow evaluation of each judge's rating on any single underlying factor in the dataset.[10]

Principal components analysis (factor analysis) is one of the most popular methods of computing measurement estimates. Under this analysis, eigenvectors calculated from a covariance matrix on mean-normalized data are ordered from highest to lowest to determine the principal components in the data; the eigenvector with the highest value is the principal component of the data.[87] Analysis of the resulting components yields the percentage of variance that can be explained by each principal component in the data. If judges are in agreement, it is expected that the shared variance on the first principal component will be greater than 60%, giving some indication that the judges share common understanding of the construct under evaluation. This analysis has the advantage of creating of a single summary score for each judge based on the strongest dimension found in the data, which allows judges to be compared along the dimension with the highest relevance. The approach does have the disadvantage that it assumes judges don't make errors when rating data.[10, 87]

There are several other methods of creating measurement estimates, including generalizability theory and many-facets Rasch models which are outside of the scope of this manuscript. The reader is referred to Linacre[88] for a detailed explanation of these models.

Generally, measurement estimates offer the advantage that they provide a single statistic enabling direct comparison of the severity of all judges on all items, even if the judges rated different items. This eliminates the need to perform pair wise comparisons for all judges on all items. In addition, these statistics offer an empirical estimate of how consistently judges apply ratings across evaluation instances. Measurement estimates offer some disadvantages; they can be difficult to calculate without software, require careful data set up, and cannot deal with nominal level data.[10]

Inter-rater Agreement Measures in Medical Informatics

Medical informatics studies relying on reference standards often involve data classification. In document classification studies, in particular, reference standards are difficult to create without expert judges, because negative case counts are either poorly defined or uncountable. For example, when judges are asked to identify text phrases that elaborate the concept to be identified (e.g., positive instances or cases), there is no way to determine the number negative instances, because the non-relevant phrases are often poorly defined, may overlap, and vary in length. This precludes the use of general statistical measures such as Cohen's Kappa (κ), which requires the negative case count. As a result, these studies rely on the metrics commonly used in IR studies: precision and recall.[13]

Precision is the proportion of cases that a judge determines is positive, that were determined to be positive in the reference standard. Precision is the same as positive predictive value. Recall is the proportion of positive cases in the reference standard that were rated positive by the judge; it is the same as sensitivity. The harmonic means of

precision and recall are often combined into the *F-measure,* the weighted harmonic mean of precision and recall (also called F_1 when precision and recall are evenly weighted), used to make pair-wise comparisons between judges. Hripsack[13] has demonstrated that the average pair-wise F-measure among experts approaches the average positive specific agreement among experts, and this in turn closely approximates the κ statistic for the data set, making such an approach an effective proxy for the precise measure, even when a negative case count in not computable.

3. A Cognitive Model of Coding Medical Documents

3.1. Introduction

Cognitive models can be as simple or as detailed as necessary to explain the theorized mechanism underlying any observed process. Much of what comprises the cognitive model, therefore, depends on a careful specification of the cognitive activity being modeled. For this dissertation, my focus is on the cognitive process of concept identification in free text. More specifically, I care about identifying concepts in outpatient ambulatory care documents compiled via electronic health record software. Specifying a cognitive model of this process relies heavily on much of the literature reviewed to this point in this proposal. However, it is worth noting that despite a large amount of literature proposing general models of cognition, the research on human concept identification in text is at best limited; in general, the vast majority of literature I am able to find attempts to model general reading comprehension, the effects of priming, and the applicability of spreading activation theory to written language understanding in general. As a result, there is hardly any current literature related to concept identification in medical text, and substantially less in concept identification over an entire document (whether medical or not). I do take care to note that much work has been done in the areas of natural language processing, neural network development, Bayesian decision networks, etc. in the computational exploration and implementation of concept identification strategies. This is not my focus here. This leaves a great deal of room for creativity in model development, at least at first glance. However, on review of the presented literature, concept identification must leverage many of the principles that researchers much more experienced than myself have painstakingly elaborated, so I don't

take an enormous amount of creative license. The model I propose borrows heavily from the experts.

3.2. The Coding Task

3.2.1. Overview

First, we must begin by identifying the elements comprising the coding task itself, so that the cognitive model can elaborate and test these activities. I begin with the activity of coding then discuss the other variables around the task.

What these experiments ask the subject to do is to read an ambulatory care document and to answer a question about the document. Very generally, then, the subject must read and comprehend the document, assess the content for the presence or absence of the concept represented by the posed question, highlight the text that supports the decision, and annotate that text to explain the reasoning process. The most general process makes no requirements on the order of the process; that is, the subject may be given the concept to match before reading the document or afterwards. Explanatory information about the concept may be made available, and if so, the timing of its presentation may vary (that is, it can be shown before the subject reads or afterwards, and may or may not be available continuously during reading, mark-up, and/or annotation). I address these timing issues in the experimental design.

The general coding process can be graphically represented by adapting the model of reading comprehension provided in Figure 10 (page 27):

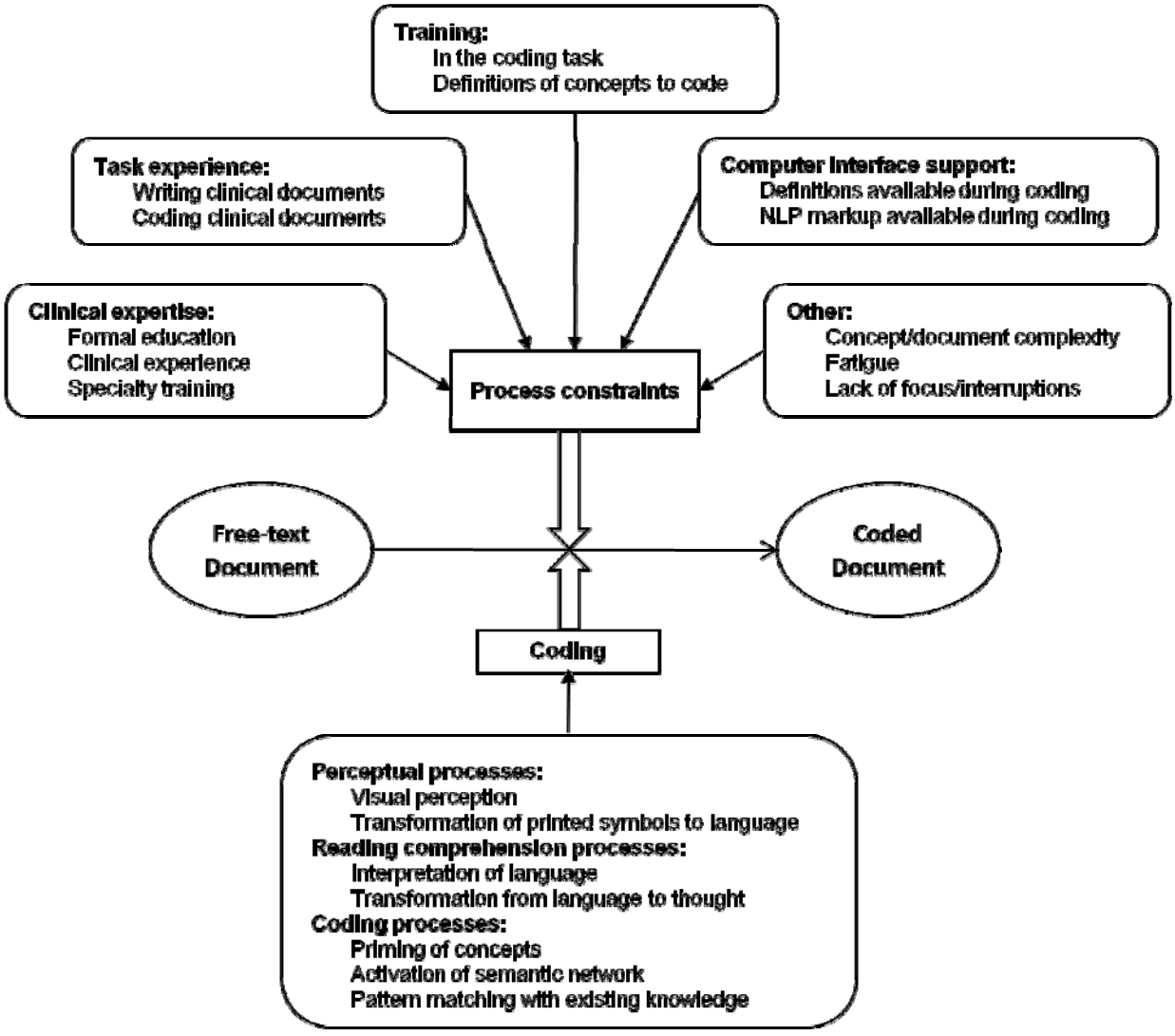


Figure 11: A model of the task of coding clinical encounter notes

This model demonstrates the various process constraints surrounding the clinical tasks, as well as the processes involved in the cognitive activity of coding. In order to move from a free-text document to a coded one, the judge must first perform the perceptual and reading comprehension processes that allow him to read and comprehend the text on the page. However, to actually code that text, the understood concepts must prime the judge's semantic network, and from this step, pattern recognition occurs to identify the concept. (These activities are explained in detail in the cognitive model description.) The figure

identifies several categories of constraints to this process. These constraints are user-specific.

This model of coding elaborates the process of reading comprehension by adding the cognitive processes involved in semantic network priming, activation of the entire network, and pattern matching to achieve concept recognition. These activities comprise the cognitive aspects of the task, and are elaborated in more detail in the model description that follows. In addition, the model shows many of the user-specific issues that ultimately impose on the cognitive processes involved in coding, acting as constraints on the model.

When a subject is asked to code a document, several initial issues must be addressed. First, the subject must be able to read and comprehend what is read. In addition, the subject must be able to read the information via an electronic display (for the purposes of this study). The cognitive model proposed in this study does not attempt to account for the perceptual and cognitive skills of reading comprehension, but instead focuses on the cognitive processes of concept identification. As a result, this portion of the task relating to concept identification is described in the presentation of the model itself. However, it is important to discuss the constraints on processing presented in Figure 11, to provide a foundation for the experiments that will be conducted on the hypotheses related derived from the model.

Concept identification in coding clinical documents can be constrained by many factors. Most of these constraints are intrinsically independent variables related to the subject performing the coding; others have to do with the documents that are being coded. First

and foremost, we consider subject background knowledge or clinical expertise. Intuitively, it makes sense to assume that an experienced clinician is likely to possess more clinical knowledge than a lay person. Thus, the clinician will recognize patterns of care and treatment, understand the relationships among physiological systems, recognize the effects of disease on these systems, have a general sense of prognosis and outcomes from the illness, etc. In addition, it is logical to assume that a clinician will be able to more readily infer information from the text based on his or her personal experience. This assumption implies that not only is a formally trained clinician more able to identify concepts than a lay person, but that a clinician with significantly more experience or specialty education will more readily identify concepts than a less experienced clinician, particularly if the concepts are highly complex, required inference to identify, or are highly specific for clinically rare diseases in a subspecialty area.

Clinical experience represents only one form of the background skill a subject may bring to the concept identification task. Experience reading, writing, and evaluating medical documentation should provide at least an initial advantage in concept recognition for subjects performing this task. Medical documentation consists of highly abbreviated and variable descriptions, clinical abbreviations (whether standardized or user-specific), and highly specific medical vocabulary. In addition, encounter notes usually follow a loose but consistent structure starting with a chief complaint (e.g., reason for today's visit), and objective examination of the patient, an assessment of the chief complaint in the context of the examination, and a plan for treatment. Although these elements should be complete in the record, often they are not. All of these issues can make clinical documents only partially informative, at best, to the lay reader, as not only will the reader

be unfamiliar with the format, but may stumble over unfamiliar vocabulary and abbreviations and the apparent terseness of clinical documentation. A clinician may encounter these same problems when reading documentation compiled by a specialist who will naturally rely on specialist vocabularies and abbreviations to record care. Finally, whether or not the subject has actually performed a coding exercise has bearing on the task at hand. Medical billers have experience coding, but may be biased towards identifying only billable detail as opposed to clinical concept detail. Clinicians often have to provide billing codes at the time of documentation, and may therefore possess these billing-type skills, but regardless, they can be assumed to be likely to be able to identify concepts in text if for no other reason that they must routinely review clinical records to abstract salient points in order to provide care. Although clinicians may lack expertise in concept identification in specialty records, this should occur with much less frequency than for non-clinicians. Lay persons will likely have little experience here, though they may have performed similar tasks during reading comprehension tests.

Training, when well designed and evaluated, can improve performance on most any skill; this is certainly true in terms of concept identification. Several studies indicate that training human coders to use a schema for coding improves agreement among those performing the coding.[4, 5, 14, 89, 90] Indeed, such approaches have great intuitive appeal if for no other reason that they bring coders to a similar set of standards or definitions against which to rate text. Chapman, et al,[5] in particular, have demonstrated that lay persons can perform almost as well as physicians in locating clinical conditions in emergency room reports, despite the fact that lay people demonstrate a slower learning curve. Furthermore, carefully constructed annotation schemata also improve performance

when these guides are constructed by experts from real-world examples.[4] Such results indicate that training is very much worth the time, energy, and expense.

There are several aspects of training to consider. The first of these, how to use a computer with a mouse, is not considered here as I assume the coders have these skills already. The second issue relates to understanding the coding task. For these studies, the user must be able to highlight and annotate text that supports or refutes the concept to be identified. Furthermore, the subject must be able to annotate that text with the reasoning behind the decision to select the text. As with skills related to computer use, I assume study subjects have or will be able to gain these skills readily.

Finally, coding can be constrained by personal or subject-level factors. Motivation comes to mind in this regard, and I must make the assumption that subjects participating in this research are on task and are not attempting to undermine the results. However, even the highly motivated subject can fall prey to fatigue, preoccupation, and interruption, which can diffuse focus and undermine task-specific thought processes, resulting in errors, omissions and assumptions. These particular constraints are difficult to control here; I will ask that users attempt to limit interruptions and work on the task when they can set aside time, but there is no guarantee that this is possible always.

3.2.2. Types and Timing of Questions

How able a subject is to answer these questions depends on all of the constraints discussed in the last section, however, the task of coding a document varies somewhat depending on the ordering of the components of the coding task. If the task requires the

coder to read an entire document to decide which concept(s) it contains, this activity differs significantly from one in which the coder is asked, before reading the document, if the text contains one or more concepts. Answering questions about concepts after a document is read is a generalized comprehension task evaluating how well the coder encoded the information as it was read, and how efficiently the coder is able to recall that information when prompted with a question. When the coder is given the concept to identify *a priori,* or knows what questions will be asked regarding the document, the reading task becomes more specific; in this case, although reading comprehension plays a significant role, concept classification becomes more important as the reader has a specific focus before the reading task commences.

3.2.3. Simple vs. Complex Coding Tasks

Interestingly, there is very little *specific* literature establishing definitions of simple versus complex coding tasks, either with regard to concept structure in text or concept structure in medicine. Vergis, et al,[18] in a study involving the development and use of a structured assessment format for general surgery operative reports, suggested that simple items are those that are "…either correctly described or not (e.g., 'date of surgery')", whereas complex items comprise "…a spectrum of possible descriptions (e.g., 'technique of reconstruction'." (p.25) The study graded operative reports on overall completeness and clarity. Simple items were graded using a binary scale (e.g., either 0 or 1, representing *no* and *yes*, respectively). Complex items were graded using a 5-point ordinal Likert scale, ostensibly to allow for variation in interpretation and the difficulty in assigning exact values to more abstract concepts.

In a study designed to generate a reference standard set of representative cases from syndromic case definitions, Chapman, et al[91] discuss the difficulty in differentiating among simple and more complex case definitions, and how this directly impacts the identification of concepts in text. The authors suggest that difficulties arise from non-existent or incomplete case definitions for this specific purpose, the lack of a standard set of syndromes to monitor, and the fact that the actual syndromes occur very rarely in patient populations. Locating evidence of syndromic outbreaks thus becomes something of a guessing game, as experts attempt to constrain the specific concepts associated with general disease syndromes while lacking sufficient data to fully characterize all aspects of a given syndrome.

In a later study concerned with inductive creation of annotation schemas for marking up emergency department reports to identify clinical conditions, Chapman and Dowling[91] state that "…there are no standard guidelines for determining which words from a textual document to include in manual annotations, and the vague task can result in substantial variation among manual indexers."(p. 196) In addition, the authors suggest that simple annotation of textual material involves classification into a predefined set of terms, whereas more complex tasks involve not only this classification step but also additional encoding of the many potentially different kinds of relationships among concepts. The additional encoding can involve statements of opinion, probabilistic or causal relationships among concepts, presence or absence of conditions and how they change over time—essentially any other information that provides supporting evidence for concept classification.

In a report on detecting clinical events in medical records using the Mediclass System, Hazlehurst, et al[92] offer that structured data entry, although it can improve overall process control and promote the reliable and accurate capture of some phenomena (such as diagnostic billing codes), captures only a small proportion of the information generated at a clinical encounter. Because important clinical information can be lost when it does not readily translate to a standardized coding scheme, data accuracy improvement derived from use of these standards can come at the expense of other data more specific to the diagnosis and treatment of a patient. Using standard coding schemes for clinical documentation can thus simplify the concept identification task, but may provide an incomplete picture. This suggests that identifying concepts in free text is the more complex task, particularly in identifying negation in medical language, isolating and evaluating concept modifiers that indicate quantity (e.g., the value of a laboratory test) or quality (e.g., disease severity), and placing knowledge representation, derived from concept identification, appropriately in context.

So what then, are the criteria that differentiate simple from complex coding tasks? The question is particularly difficult to answer. The limited amount of research above and the lack of clear definitions suggest that all coding identification tasks are complex, and results from the need to process natural language. Even natural language constrained to specific medical domains consists of "…highly truncated and poorly formed language constructs" (p. 519)[92] that require translation, interpretation, and inference. In the absence of specifically stated evidence to support or negate the presence of a concept in text, inferences from other text areas can cast doubt on the absolute existence of any

concept. So, if we are to separate simple from complex coding tasks, we must set some arbitrary bounds, and while there are many factors involved in this determination, we can identify a few general rules that guide us. We do not use the requirement of concept identification in free-text as a measure of complexity, as all tasks will be performed on free text documents. So, to distinguish between a simple and a more-complex task, we rely on preliminary definitions used by I2B2.[93] A simple concept is one that is stated in the text (I2B2 refers to these as "textual" concepts). In addition, for this study, these "simple" concepts use standard (e.g., non-clinical) English. A complex concept is one that requires some level of inference or judgment by the coder, and contains a strong clinical component.

Limiting simple coding tasks to identification of the presence or absence of text in a document reduces a certain amount of cognitive overhead, as the task of concept identification becomes one of visual pattern matching and simple reading comprehension. This represents a "less is easier" philosophy. The need to define simpler tasks in terms of frequently occurring words helps to level the playing field between experts and lay persons by eliminating the need to memorize unknown terms. More complex tasks, where the coder must infer a judgment require additional cognitive work to relate the current information to experiential history, recall learned information, or weigh and evaluate imprecise information. Finally, intuitively we think of a task that does not require domain expertise as simpler than a task that requires this additional knowledge. In addition, it follows that such tasks are ones where coders are more likely to agree. These intuitive leanings remain to be proved.

3.3. The Proposed Model

3.3.1. Introduction

This proposal hypothesizes that the cognitive task of coding free-text ambulatory care documents is one of classification in context. This means that a subject must read and comprehend the contents of a document, then, in response to a question, determine if sufficient evidence exists (either directly stated in the document or inferred from document contents) to answer the question. Important in the consideration of this activity is whether or not the question is posed before the coder reads the document or afterwards. In addition to the timing of the question, we must also consider what information regarding the definition of the concept is provided to the coder, and when. I will first broadly discuss the cognitive activities associated with concept identification and return to these temporal issues later.

3.3.2. The Representation of Knowledge

I offer that concept identification is inextricably bound the particular mental abstraction used by the reader to represent knowledge. That is, the reader must possess the capability to abstract what is read and subsequently store that information in a manner that makes its recall more or less immediate. The reader must also be able to "hold" information about what is already known or learned (experiential, educational, cultural knowledge) in a manner that makes it easy to bring to bear on the information that is read. Finally, the reader must be able to integrate both sets of knowledge to form cohesive thought and make determinations regarding the degree of match between concept and text. To these

ends, this model of concept identification in text requires that we identify how read or previously known information is stored structurally.

As discussed in previous sections, semantic networks and schemas represent methods of knowledge organization. Before the reader even approaches a text comprehension/concept identification exercise, she has almost immediate access to an enormous amount of declarative and procedural knowledge. *This model proposes that this knowledge is organized in networks and that although the network contents may vary, the fundamental structure of the information is the same and closely resembles the semantic network, where nodes contain propositions, and links contain the information that relates those nodes.* Such a structure places no constraints on the type of relational information between nodes; the network can therefore represent any relationship among propositions, including procedures that act upon them.

3.3.3. Using Asthma as an Example

The construct of *asthma* comprises many concepts as well as an inordinate amount of qualifying detail (relationships among component concepts). No single clinical expert likely retains every permutation of knowledge related to asthma, and the knowledge related to this complex clinical condition changes over time. As a result, the clinician must maintain a working knowledge of asthma that she can call upon when determining if, when reading a medical text, a patient has asthma or not. In the case of the non-clinician, it is likely that this person also possesses working knowledge of asthma ranging from recognizing the word and ascribing the most general meaning to it (e.g., “a patient with asthma has trouble breathing”) to more sophisticated knowledge up to the clinical

expert level (e.g., as when the patient has a close relative with asthma and is intimately involved in the relative's care). Regardless of the amount of knowledge a person possesses about a medical concept, this knowledge is arranged in a network of concepts with relationships among the concepts.

No two knowledge representations are the same, regardless of the level of expertise, because each individual will have a knowledge representation of asthma built over years of experience in combination with cultural norms, personal beliefs, etc. So, for example, where a pulmonary specialist might have a highly detailed internal semantic representation of asthma that includes categories such as: 1) signs and symptoms of asthma, 2) diagnosis of asthma, 3) tests and studies used to diagnose asthma, 4) treatments for asthma, 5) when to use which treatments, 6) patients not likely to respond to treatment, 7) recent cases in which I managed an asthmatic patient, 8) case studies I have heard about but have not seen, etc. The lay person, on the other hand, with little direct experience of asthma may have only the following general concepts included in his internal representation: 1) difficulty breathing, 2) my cousin uses an inhaler when he starts wheezing, 3) grass allergies cause my cousin's asthma to get worse, 4) one can die from an asthma attack, etc.

Clinicians may likely cluster similar data similarly, as may lay people. That is, people will develop general defining categories at high levels, and will filter to more specific knowledge at lower levels. *I propose that clinicians and lay people may have different basic level categorization determinations, and that these are close, but that the clinician will have greater specification at all levels than will lay people.*

3.3.4. Activation of Knowledge

The act of reading comprehension is believed to be the activation of stored concepts. Activation can be lexical or semantic; both types of activation launch a memory search for matching patterns (one for word matches, the other for concept matches). I assume that these features or current activation theory operate in this manner and I do not worry if the original activation point is single or results from multiple activation cues, or if the activation is a single or multi-step process (both of these issues are under strong theoretical debate). *I suggest that the expert will identify and match clinical asthma (or other medical) concepts more readily because the expert has a richer network of asthma-related concepts, so will identify more things that relate to asthma that the lay person. In addition, the expert will be able to infer more from the text than will a lay person, only because the network of concepts related to asthma contains this knowledge or contains patterns similar to this knowledge.*

Explicit positives and explicit negatives will be matched by both experts and lay people because no detailed conceptual representation of the concept of asthma is required; the person only needs to comprehend the meaning of the text. However, in the absence of explicitly stated negatives and positives, this model assumes that some level of inference is required, and suggests that the depth of the semantic network determines the level of the inference that is possible. An expert may have a deep enough network—one that contains explicit examples or matches for even rarely occurring text patterns—that no subsequent inference is required, though there is no guarantee that this is the case. I postulate that the proposed model explains why experts will therefore be able to draw inferences that lay persons cannot. In addition, experts may have general knowledge

(about how documents are structured, and where information about asthma is located) that lay people may not have, that will further offer an advantage. For example, experts may have an experiential sense of where, in a clinical record, particular information has a very high probability of occurring (e.g., in the history, or review of systems, etc.). It remains to be seen if this information can be learned by lay readers during the experiments, as the documents will contain fairly consistent formatting.

I hypothesize that an *a priori* concept description activates the individual's internal cognitive representation of both the concept to find as well as closely related concepts as determined by an individual's internal heuristics, so that the individual is "primed" to locate commonly expressed patterns relating to the actual concept descriptions in the text. Much like a physiological (e.g. neuronal) response to a stimulus, asking the coder to determine if a concept is present or absent in the code elicits a physiological response that lowers the activation threshold of text patterns that potentially match and simultaneously raises or blocks the activation thresholds of probable non-matches. The breadth and depth of both the general concept formulations as well as related concepts in the individual's mental model are a function of the coder's experience and knowledge in the domain, as well as the coder's relevance judgments about the information selected. This hypothesis is supported by research by Chapman et al, who show that coding performance increases, and inter-rater variability decreases when coders are trained using schema for concept definition.[4, 5] The cognitive model suggested here explains these performance improvements because *training provides the mental priming effects (through detailed explanation of the concepts) as well as a pre-defined categorization method (the schema) to support coder cognition*.

3.3.5. Training Effects

The proposed model, being solidly grounded in spreading activation theory, with the theory's related concept of priming, suggests that *training will improve performance at any level, from lay person to expert*. In addition, this suggests that the greatest improvement in concept identification will be noted in lay users, who have the most to gain from training. Finally, because activation theory proposes that trace activation decays over time with loss of focus or trace use, it is reasonable to assume that the knowledge gained during training may be lost over time by both groups. However, experts with high strength concept nodes and links are likely to demonstrate as great a loss of knowledge as lay persons. Because training has been demonstrated to improve task performance, this study will not test or evaluate these effects.

3.3.6. Iteration and Context

The task of coding is iterative; each iteration results in a new context which, combined with continued pattern matching, results in a new activation potential or priming of the individual's internal conceptual model, as well as redefinition of potential categorization schemes, to determine whether or not the concept is present in the text. This suggests that a simple, direct, and/or common concept should be easily identified by all coders and should demonstrate less inter-rater variability, and that a more complex one will result in greater variation among coders depending upon their experience and expertise. More specifically, this suggests that *explicitly stated (or negated) concepts will more likely show less variation among coders than those that require inference based on context. In addition, inter-rater agreement should correlate in the specific text patterns identified by*

coders, such that simpler concepts are likely to be identified by highly similar text fragments, and more complex concepts are more likely to show greater variation in the text selected to support the coder's determination.

3.3.7. Assumptions

Consistent across the many scientific disciplines contributing to our understanding of cognition is the notion that our massively parallel-processing brains can filter enormous amounts of sensory input to generate symbolic representations (or, abstractions) which can be manipulated by mental processes (e.g. actions or operations that accomplish a goal) so that we can make sense of and interact with our physical environments to understand, problem-solve, and learn. Importantly, the processing of this enormous amount of sensory input relies on the ability to recall stored knowledge and to interpret the incoming stimuli using that knowledge.

With this in mind, there is a large set of potentially confounding variables related to the acquisition and storage of information which should be controlled to make the experiments tenable. For example, to the study will not assess the working memory capacity of each subject in this study, and therefore will not be able to account for differences in cognitive processing speed or ability among individuals. Additionally, to mitigate potential effects of non-native language interpretation or translation on the processing of clinical documents only fluent English speakers may participate in this study. To control for reading ability in general, all subjects must be college educated. A complete list of assumptions about the *subject* is presented here:

- Subjects have no neurological or perceptual deficits (inability to see text, read words, etc.) that prevent them from reading and understanding information displayed on a computer monitor.
- Subjects are comfortable using a mouse, are familiar with a computer, and require no basic computer training to participate in the study (beyond training in how to use the data-collection application).Subjects are assumed to understand the coding task required of them, and can demonstrate that they are able to perform the task satisfactorily.
- Subjects are fluent in English.
- Subjects are college-educated and are assumed to possess college-level reading skills.
- Clinical domain experts have graduated from medical, nursing, pharmacy, osteopathy, or dental school. (Note, unless an informatics expert is formally trained in one of these clinical domains, the subject is considered a lay person.)
- Non-clinical study participants may be students in any of the clinical schools mentioned above, but have not yet completed their core education.
- Subjects aren't trying to cheat: the study assumes subjects are on task and aren't intentionally attempting to undermine the research.
- Experiments will ask subjects to complete concept identification in a single document in a single session. This is to prevent interruptions which may upset experimental design. Users may certainly view and mark-up more than one document in a sitting; they must complete the mark-up/concept identification task for a single document without interruption.

As with subjects, there are assumptions that must be stated prior to elaborating the cognitive model. Assumptions about the *model* include:

- The model is not a model of reading comprehension; it is a model of concept identification. As a result, it is assumed that the reader can comprehend the text being read.
- This study is not concerned with whether or not concept processing takes place in working memory, extended long term memory, or long term memory. The study only assumes that the coding process can be defined in terms of cognitive representations and how these representations are leveraged.
- The model is not concerned with cognitive load or other factors that can inhibit optimal cognitive functioning. To this end, the model does not account for subject fatigue, interruptions, or other factors that may impede concept identification. (In fact, part of the experimental design is intended to obviate these issues.)

3.4. Research Hypotheses

3.4.1. Introduction

This study is based on two hypotheses, discussed in detail below. The hypotheses represent very basic statements about the expected differences in agreement among expert (clinician) and non-expert (lay persons) coding clinical encounter notes. Each hypothesis is explained in the context of the background theory supporting it as well as the proposed model.

3.4.2. Hypothesis #1

Experts will demonstrate higher inter-rater agreement among themselves than will lay coders.

Based on the proposed cognitive model, this happens because, regardless of the network used for knowledge representation, experts possess a deeper and more richly detailed network than that lay persons. This is especially true when lay persons are not trained (e.g., do not have definitions provided to them). In addition, the knowledge representations used by experts more likely contain similar concept information than do those used by lay persons, as well as multiple, overlapping pathways to get to similar concepts from differing clinical information. Despite the fact that the representations vary in layout, for clinicians, the overall knowledge they contain will be fairly equal.

Although this hypothesis supports a belief that experts will vary in their coding less within their group that within the lay-person group, this hypothesis does not make any statements about how much the variability will differ among groups (only, hopefully, that it is statistically significant). Clinicians can easily disagree. What is particularly difficult here is identifying what the basic level understanding is for any single clinician, and how that maps to the basic level understanding of another. As with genus level classifications, this hypothesis carries a tacit assumption that these basic level representations are fairly close and can be accessed cognitively in a multitude of ways (which vary due to clinical experience, training, etc. and how these affect the development of memory traces). Testing the "location" (e.g., within the semantic network) of the basic level, however, is difficult. It is possible that clinical guidelines can be used to assert what a basic level

understanding *should be*, and to evaluate clinicians' responses in that light, noting where they fail to identify concepts elaborated in these practice guides.

3.4.3. Hypothesis #2

When coding free-text documents, inter-rater agreement for both experts and lay people is higher when the task is simple and lower when the task is more complex. Inter-rater agreement will improve among both experts and lay persons when the concepts provided in the text are simple, and are explicitly stated. This suggests also that agreement should improve within groups. These results are based on the assumption that lay persons will not need to rely exclusively on their internal knowledge representations to make inferences, which dilutes the experts' domain knowledge advantage. Conversely, inter-rater reliability will fall off rapidly for lay users when concepts are complex and are not explicitly stated. The performance degradation will be less for experts, who are presumed to have greater ability to infer specific detail because of their more elaborate knowledge networks, and because they have stronger memory traces based on experience.

It is possible to assume that concept identification is no more than simple pattern matching (in most cases). For example, if asked "Does this patient smoke?" one would likely look for patterns like "The patient denies smoking," or "The patient smoked for several years. Quit since 1983" or similar patterns. In such cases no extensive internal knowledge about smoking, its effects on the respiratory and/or cardiac systems, implications in the development of cancer, motivation to quit, addictive behavior, etc. needs to be accessed for me to make a simple and quick decision. In addition, and

importantly, the knowledge needed to answer this question is not medically specific, but highly general: aside from performing the tasks of reading comprehension and determining meaning from the words, the only inference a subject must make in the second case (e.g., the example where the patient reports they no longer smoke), occurs during the interpretation of the temporal significance of the statement, "Quit since 1983." Thus, the assumption that concept identification is simple pattern matching is only partially true: even a "simple" question can require significant experiential inference to resolve. Although the hypothesis speaks only to the level of agreement, it will be worth noting if simple pattern matching emerges commonly in concept identification.

This hypothesis suggests another point of analysis. Here, the issue is how does a simple or complex task determine the level of internal processing a user must perform? In terms of a highly clinical question, this depends on both the question and what is actually in the document. A simple question may require simple, straight-forward information to answer. However, if this information is not in the document, then the clinician may deeply search his semantic network for relationships that might allow him to infer an answer based on the limited information that is provided. A more complex question may produce the same results. However, in the presence of careful and comprehensive clinical documentation, the level of complexity of the question may have no bearing on the level of inference, and hence, require only shallow processing of a detailed network to derive the answer.

3.5. Summary

The proposed cognitive model is extraordinarily simple. It states that internal knowledge representations are encoded in semantic networks, and that experts have richer (e.g., more detailed, more highly leveled, more tightly interconnected) semantic networks in their domains of expertise than do lay people. Clearly stated concepts in text, especially explicitly stated negatives and positives will not require use of internal semantic networks and will therefore show little difference between expert and lay coders. Concepts requiring inference will demonstrate greater variability among experts and lay coders because. Concepts provided with annotation guidelines (e.g., concept definitions) will narrow the difference in performance between experts and lay persons because they reduce reliance on internal semantic networks by externalizing and standardizing the knowledge representation.

4. Methods

4.1. Experimental Design

This study compared the similarities and differences between non-clinicians and clinicians when performing cognitive coding tasks requiring answers to questions about clinical documents. Two groups of eight clinicians and eight non-clinicians reviewed a series of 60 clinical ambulatory encounter notes, thirty pertaining to smoking and 30 referring to asthma. For each document, each subject answered two questions. For each question, the subject was asked to highlight the text she used to determine the answer, and, additionally, to annotate the selected text (or, "snips") as desired with comments describing why the particular text was chosen or how it helped the subject answer the question. Data were collected via a password-protected web site and stored in a secure database.

4.2. Subject Selection

4.2.1. Inclusion Criteria

All subjects enrolled in the study met the inclusion criteria detailed in section 3.3.7., above.

4.2.2. Sample Size

The hypotheses presented above compare the levels of agreement between non-clinicians and clinicians. Thus, the sample size is the number of pair-wise comparisons necessary to achieve statistical power. Lacking a preliminary test data set to use to estimate average agreement or variance, several sample size calculations were run over hypothetical, best-guess data. The following table demonstrates the several combinations of assumptions used in the calculations; these are explained in detail below the table:

	Clinicians % agreement on questions			**Non-clinicians % agreement on questions**		
Run	Simple	Complex	Absolute Difference	Simple	Complex	Absolute Difference
1	99-100	90-100	1-10	95-100	80-100	5-20
2	95-100	80-100	5-20	90-100	70-100	10-30
3	90-100	70-100	10-30	80-100	60-100	20-40
4	85-100	65-100	20-35	70-100	50-100	30-50

Table 1: Hypothetical values for sample size estimation

This table contains hypothetical data used to estimate the sample size needed in this study to determine statistical significance.

The hypothetical data were used to generate numbers ranging from tight agreement among subjects, and where it had much greater variance. In each case, an average value for the provided range was calculated. The estimated standard deviation (σ) was determined by dividing the range in agreement by four. For example, for run #2, above, the range of percent agreement by clinicians on a simple question is 1 (99-100 percent agreement). This makes the rough estimate of the standard deviation ¼ or .25. In the same run, the range on the complex question is 10, providing a rough estimate of σ of

10/4 or 2.5. All of these calculations were performed for all questions and the absolute range difference between clinicians and non-clinicians for simple and complex questions. These values were entered into STPLAN (a command-line based study sample calculator from BAM Software at http://biostatistics.mdanderson.org).

All calculations were performed to produce a statistical significance of .05 and a power of .80 for normally distributed data. Estimated standard deviations were never the same for two-sample tests, so two-sample t-tests for samples of unequal variances (Welch approximations) were used instead. The largest sample size calculated for the above data, regardless of whether the t-test was a one- or two-sample test was 8.766 or 9. This sample size was determined for both runs #2 and #3; sample size estimates fell lower for all remaining calculations, so even if the percentage of agreement within groups was overestimated in the hypothetical data, the calculated sample size of 9 pair-wise comparisons held.

This means that 9 pair-wise comparisons must be performed to meet the desired criteria for statistical significance. Four clinicians (or non-clinicians) would result in 6 total pair-wise comparisons within each group ((n(n-1))/2 = 4(3)/2 = 6). Five clinicians (or non-clinicians) would result in 10 pair-wise comparisons within each group (5(4)/2 = 10). So, 5 clinicians and 5 non-clinicians as subjects provide a sufficient number of comparisons to power this experiment to achieve statistical results based on the range of variances in the hypothetical data.

4.3. Clinical Document Selection

4.3.1. Inclusion Criteria

Three hundred clinical encounter notes dealing with smoking and/or asthma (as identified by ICD-9 codes) were requested for the study. The notes were completed by residents and faculty using Epic (http://www.epic.com) following ambulatory visits to the OHSU outpatient general medicine and pulmonary clinics. The requested records included consecutive records meeting ICD-9 criteria. Of this set, the first 30 full notes related to smoking and the first 30 related to asthma comprised the study documents. (The larger set included a high proportion of notes documenting faculty review of residents' work, and thus did not include full patient evaluations, so these documents were skipped.)

4.3.2. Sample Size for Documents

The number of documents included in this study was based on the practical consideration of what was a reasonable set of documents to ask reviewers to code. A very simple pilot study with anesthesia preoperative evaluation documents showed that 3 non-clinicians could highlight text, annotate it and answer the posed questions *manually* in about 90 seconds per document. This suggested that 60 documents might take a maximum of about 2 to 4 hours total) for a test subject to annotate, which seemed to be a reasonable amount of time to ask of busy volunteer subjects.

The number of documents selected did not impact the sample size for the statistical analysis to be performed. Because this study qualitatively analyzed annotation data, a desirable amount of qualitative data is necessary. There is no published documentation

indicating what a minimum document sample size should be for annotation studies; 50-60 documents are commonly used, however none of the studies provides supporting evidence for why that number of documents selected. Thus, the selection of 60 total documents for this study was somewhat arbitrary. It is worth noting that this number may had an impact on the variance estimates used for subject group size calculations in that a greater number of documents might have revealed greater variance among subjects and between groups.

4.4. Concept Questions

The intent was to ask a simple question and a more complex one in each domain. This experiment tested the following two domain questions:

In the domain of asthma treatment/care:

1. Does this document clearly state that the patient's asthma is well-controlled? (simple)
2. In your opinion, is this patient's asthma well controlled? (complex)

In the domain of smoking cessation:

1. Does this document clearly state that this patient smoke? (simple)
2. In your opinion, does this patient smoke? (complex)

The intention was to demonstrate that the hypotheses held across concept identification dealing with various facets of patient care. The "easy" or "simple" questions asked subjects to identify clear (to the subjects) statements in the text that answered the posed question. The second ("complex") questions required subjects to read the text and infer - based on the subject's knowledge and/or experience - what the response to the question

should be. The smoking questions were considered simpler overall when compared to the asthma questions because the language of these questions is less likely to be medicine-specific, that is, smoking data rarely consist of what non-clinicians may consider impenetrable clinical jargon.

4.5. The Annotation Tool

4.5.1. Introduction

Although many freely-available annotation tools exist, the elaborate functionality of these tools, coupled with their enormous and often confusing panoply of feature sets, suggested that a simple solution tailored to this study would reduce programming overhead and subject training. For these reasons, we developed a simple annotation tool tailored specifically for this study. The guiding principles for tool development included 1) platform independent architecture, 2) potential for web deployment, 3) compatibility with the Postgres open-source database engine, and 4) development on a UNIX server. As always, these principles needed to fully support a clean interface with straight-forward usability for subjects.

The annotation application was written in the Ruby (http://www.ruby-lang.org) programming language on the Rails (http://rubonrails.org) web-based development framework, running on a Postgres back-end database. The web application displayed contents and all collected data were stored in the Postgres (http://www.postgresql.org) database, meaning that documents and questions could be changed without altering the programming interface. Ruby was chosen for the ease one can build a functional web-

based application on the flexible Rails framework. In this case, the web interface acts as a container to display data held in a database. Any set of documents could be displayed, just as any two questions could be displayed, as these data are database elements. Currently the application allows for two questions per document, though this could be changed. Importantly, the application itself is independent of the data it displays and collects, making it useful for other, relatively simple annotation tasks.

4.5.2. Document and Question Display Interface

The web-interface displays all documents in the following fashion. Subjects may review and annotate a single document at a time. Once the document is complete, subjects may advance to the next document. Subjects are not allowed to return to a previous document once it has been reviewed.

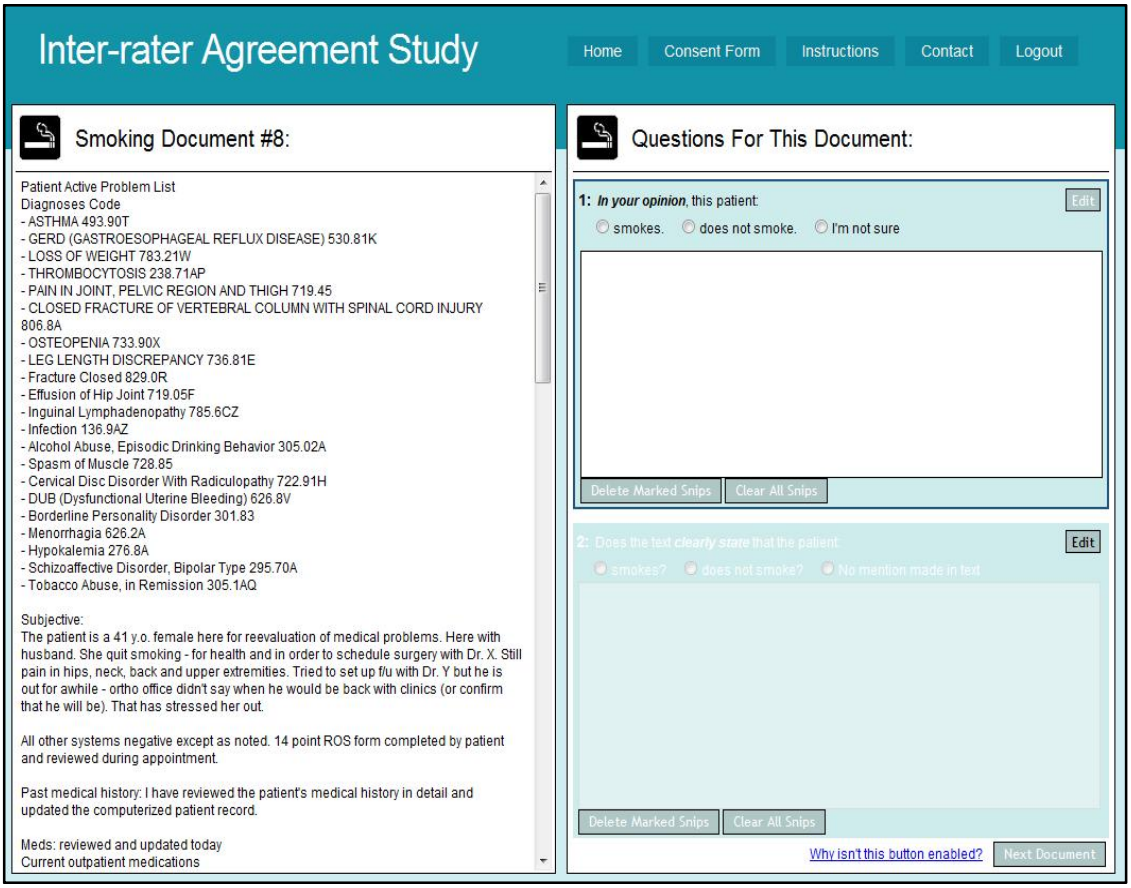


Figure 12: The data collection application interface

The document to annotate is displayed on the left. The document display area has a scroll bar to the right to allow the user to view complete document contents. The two questions to be answered for the document are displayed on the right, one above the other. When the page is opened, the first question is active; the user can toggle between the questions as desired. The user select text "snips" from the document by highlighting them with the mouse. Highlighted snips write to the enabled questions box where the user may annotate the snips with comments as desired. Once both questions are answered, and text snips and/or comments are entered the user selects the "Next Document" button to advance to the next document.

The following screen shot demonstrates the appearance of the interface when the user has selected text snips and commented on them for question #1:

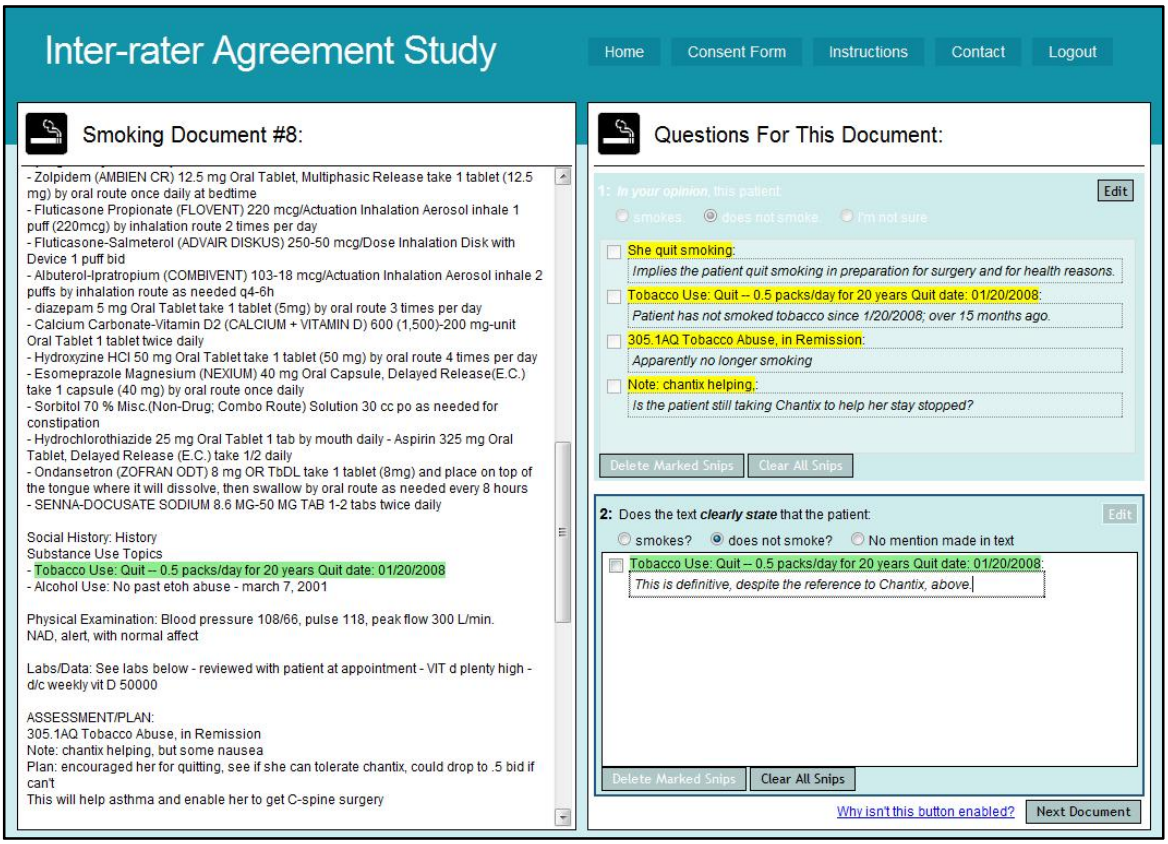


Figure 13: Data collection application with user-entered data

In this screen shot, the user has answered both questions. Question #2 is currently enabled, and the single selected document text snip has been entered in the question area. The user has entered a comment ("This is definitive despite the reference to Chantix, above.") associated with the selected text snip. Note that snips selected for question #2 are highlighted in green, both in the question answer box and in the document itself; when the user toggles back to question #1, the snips selected for that question are highlighted in yellow, both in the document text and in the question window. This document meets the criteria for completion here (e.g., both questions are answered, and

both have supplied snips and/or comments to support the answer), so the "Next Document" button is enabled, and the user may move to the next document.

4.5.3. General Application Functionality

The application operates in the following manner:

1. Subjects highlight text snips (using the mouse) that represent information related to the concept of interest. Subjects are asked to highlight any text that helped them make the determination about how to answer the questions posed on the right side of the screen.
2. Once text snips are selected, the snips write to the question snips area for the current question. A comments box is appended to each snip, where the subject may enter any explanatory comments desired (e.g., why this snip was selected, how it supported the answer, why the snip was confusion, etc.).
3. Snips and their associated comments may be deleted en masse, by selecting the "Delete all Snips" button, or individually, by checking the checkbox next to each snip/comment combo and selecting the "Delete Marked Snips" button. There is no theoretical limit on the number of text snips a subject may select.
4. The subject must answer the question (by selecting one of the provided radio buttons) before the question is considered complete.
5. The questions and radio responses for documents are as follows:

   In your opinion, this patient

   a) smokes.

   b) does not smoke.

c) I'm not sure.

Does the text *clearly state* that the patient

a) smokes.

b) does not smoke.

c) No mention is made in the text.

In your opinion, is this patient's asthma

a) well controlled.

b) not well controlled.

c) I'm not sure.

Does the text clearly state that the patient's asthma is

a) well controlled.

b) not well controlled.

c) I'm not sure.

This response is the coder's best judgment, based on the presence of sufficient information to answer the question. For example, in response to the smoking question "In your opinion, this patient…" suppose a coder answers "smokes." If the coder highlights text such as: "The patient denies smoking" in one part of the document, and "The patient lives at home and his father smokes 3 packs a day" elsewhere in the document, the second statement implies the patient is exposed to large amounts of second-hand smoke, which the rater may feel provides sufficient evidence that the patient smokes. This response is a judgment call by the coder.

All question answers, along with their associated text snips and comments are saved in the Postgres database. Snip offsets in the documents are calculated during post-processing of the data once all subjects have completed all questions.

4.5.4 Data Collection

All data are written to a relational Postgres database, and are collected via the web tool. All data elements are time-stamped with the time of creation.

4.6. Experimental Design

4.6.1. Consent

All subjects were formally consented according to OHSU Institutional Review Board guidelines (OHSU IRB #4943).

4.6.2. Training

Each subject was trained in how to use the web data collection tool during an approximately one-half hour training session. During this time, I demonstrated the use of the web annotation tool and remained with the subject while they annotated documents specifically selected as test documents for learning the application. The training session provided time for subjects to ask questions about the annotation tool, what was expected of them, and how to contact me.

Each subject was provided with three training documents (two smoking and one asthma document) for learning the interface. The training documents were selected specifically to demonstrate the interface cues (e.g., identifying icons, titles) designed to alert the user to

the document type. In addition, training documents were clearly watermarked to indicate that they did not count as study documents. All data collected during annotation of the test documents were discarded.

4.6.3. Document Presentation

Each subject reviewed 58 total documents.[vii] This review constituted a single experiment through which data to test both hypotheses were collected. The subjects were asked a simple and a complex question on each document and were instructed to annotate their answers as described in 4.5.3., above. All subjects were asked the same questions for each document.

Subjects were asked to complete their annotations over a two-week period, in any number of independent sessions convenient for them. In addition, subjects were instructed to complete the current document annotation before terminating a session (that is, subjects were notified that ending a session without completing the annotation on a document would result in the loss of the data for that document[viii]).

All coders reviewed the documents in the same order, and answered the same questions on all documents. This was intended to mitigate learning effects due to document processing order. Since all raters saw all documents in the same order, everyone had the same opportunity to learn from the previous documents. In addition, the limit of two weeks for completion of this task was intended help non-clinicians retain learning

[vii] A counter error resulted in only 28 of 30 asthma documents being evaluated in this study. This is discussed in the results section, below.

[viii] In fact, the tool is designed such that exiting without completing both questions on the document caused the data for that document to be lost. In this case, the user was required to complete that document as the first document of the next session.

accumulated through document markup and thus help minimize the within- and between-group effects due to learning.

Finally, to prevent confusion as to whether or not a subject was answering questions related to asthma or smoking, all smoking documents were displayed in order first followed by all asthma documents. (Simple pilot testing of the interface demonstrated that despite visual cues distinguishing types of documents, users often highlighted text supporting the wrong concept (e.g., the subject highlighted smoking information for an asthma document), because the documents were similar in form and style, and the subjects appeared to anticipate (incorrectly) the question to be answered.

4.6.4. Study Completion

Once a subject had completed annotating all 58 documents, the subject had completed their participation in the study, and his/her login was automatically disabled.

4.7. Hypothesis Testing

4.7.1. Introduction

The data collected for this experiment provided two primary levels of analysis: 1) document-level analysis, and 2) phrase-level analysis, which are discussed in detail below In addition, the analysis involved both quantitative and qualitative methods. The following data were collected for each question, in each document:

1. The answer to the question (there were two questions per document).

2. One or more subject-selected text segments from the document in support of the given answer.
3. Zero or more subject-entered comments explaining why the text segments were selected.

Of these, items #1 and 2 were analyzed using quantitative methods. The comments data were evaluated qualitatively. A final summary review synthesized the results for discussion.

4.7.2. Question-Level Analysis

All hypotheses were first analyzed at the question level. *Two raters agreed if they give the same answer for the same document question* (e.g., if, when asked if this patient smokes, both coders answer "I'm not sure/No mention is made in the text"). Because this study sought to explore differences between clinicians and non-clinicians, pair-wise comparisons within the clinician and non-clinician groups were performed separately.

Subject to subject mean agreement for a single question across a document set was calculated as:

$$\frac{answers\ pairs\ where\ subjects\ agree}{total\ answers\ pairs}$$

Twenty-eight pair-wise comparisons were possible with 8 (or, *n*) subjects:

$$(n/(n-1))/2 \ = \ (8*7)/2 \ = \ 28$$

For smoking documents, this resulted in 840 total answer pairs for a single question:

$$30\ documents * 28\ comparisons/question = 840\ comparison/question$$

Subject-to-subject mean agreement was therefore the percentage of time among 840 answer pairs where both subjects selected the same answer. For asthma documents, only 28 documents were analyzed by the subjects, resulting in 784 comparisons per question:

$$28\ documents * 28\ comparisons/question = 784\ comparisons/question$$

Hypothesis #1: Clinicians will demonstrate higher inter-rater agreement among themselves than will lay coders.

For this hypothesis, pair-wise comparisons were performed among members of each subject group. Mean agreement among all subjects in a study group was calculated, then compared against the mean measures of agreement *between* study groups. The following table illustrates hypothetical comparisons among *clinicians* for a set of smoking documents (note: the table is abbreviated to demonstrate only 5 total subjects):

Comparison*	% Agreement for the Simple Question (Q1)	% Agreement for the Complex Question (Q2)
S 1 - S 2	97	90
S 1 - S 3	97	80
S 1 - S 4	98	86
S 1 - S 5	99	81
S 2 - S 3	100	89
S 2 - S 4	100	85
S 2 - S 5	99	86
S 3 - S 4	97	83
S 3 - S 5	100	85
S 4 - S 5	98	89
S 1 –S 3	97	90

* S = Subject

Table 2: Hypothetical data for pair-wise comparisons among clinicians

This table demonstrates pair-wise comparisons between subjects across question answers. Each comparison represents a between-subject average agreement over all smoking

documents. These pair-wise comparisons were performed for both questions in all documents for each study group.

For this *document level* experiment several statistical analyses were possible. Normal distributions were assumed for all statistical analyses.

1. To compare clinician and non-clinician agreement on simple questions, the following hypothesis was tested:

   $H_0: \mu_1 = \mu_2$

   $H_a: \mu_1 \neq \mu_2$

   Here, μ_1 is the average (mean) agreement among clinicians, and μ_2 is the mean agreement among non-clinicians. The null hypothesis states there is no difference in agreement between the two study groups. The alternative hypothesis is that there is a difference in agreement among clinicians than among non-clinicians

   This hypothesis was tested with a 2-sample, 2-sided t-tests of independent samples. The specific t-test was be selected based on the standard deviations of the samples. If the variances were equal, the Student t-test for equal variances was be used. Conversely, when the variances differed, t-tests for samples of unequal variances (Welch's approximation of the Student t-test) were used.

   The choice of a two-sided t-test may seem odd in light of how the hypothesis is stated, e.g., that clinicians will agree *more* than non-clinicians. For this study, if one-sided tests were performed under an alternative hypothesis of:

   $H_a: \mu_1 > \mu_2$

it could become difficult to interpret an instance of u1 < u2, should it occur. For this reason, the two-sided test using the simpler alternative hypothesis imposed a careful assessment of the *direction* of significant results to explore and explain using the proposed cognitive model.

2. To compare clinician and non-clinician agreement on complex questions:

   The same null and alternative hypotheses as used for simple questions, above, were used for comparison of agreement among clinician and non-clinicians on complex questions.

Hypothesis #2: When coding free-text documents, inter-rater agreement for both clinicians and non-clinicians is higher when the task is simple and lower when the task is more complex.

Testing this hypothesis required three tests. The first two evaluated the statistical significance of agreement *within* the study groups. The third evaluated the difference in agreement *between* groups.

The comparison of clinicians and non-clinicians for this analysis required first that the absolute difference in agreement on simple and complex questions for both study groups was evaluated. Referring back to Tables 1 and 2, these absolute differences were demonstrated in column 4. To perform this analysis, the following hypotheses were chosen:

$H_0: \mu_1 = \mu_2$

$H_a: \mu_1 \neq \mu_2$

As discussed previously, μ_1 is the average (mean) difference in agreement between simple and complex questions for clinicians, and μ_2 represents this same value for non-clinicians. The null hypothesis in this case implies there is no difference in these means. The alternative hypothesis states that the average difference in agreement for clinicians between simple and complex questions is *different from* that for non-clinicians.

This hypothesis was tested with a 2-sample, 2-sided t-tests of independent samples. (Two sample t-tests were employed for the same reasons discussed under hypothesis #1.) The specific t-test was selected based on the standard deviations of the samples, as discussed under Hypothesis #1.

4.7.3. Phrase-Level Analysis

Preliminary Considerations

To gain insight into how to analyze phrases selected by codes, a simple pilot study was performed using three volunteer coders (two informatics students as lay coders and one expert (this author)). Each subject was presented with the following definition of well-controlled asthma:

In general, well-controlled asthma would include:

- *Regular use of controller medication.*
- *Little or no use of rescue medication (less than once a day, except for patients with exercise-induced asthma who might use a rescue medication*

prophylactically prior to exercise, or less than half of one MDI[ix] (<100 puffs) a month of rescue medication.

- *No interference due to asthma with normal activities of daily living (e.g., going to work or school, activities at home, etc.).*
- *No shortness of breath in the last 4 weeks.*
- *Asthma symptoms (wheezing, coughing, shortness of breath, chest tightness or pain) waking the patient up at night or earlier than usual in the morning less than once a week in the past 4 weeks.*

All coders saw the following text chunk from a preoperative anesthesia evaluation note "review of systems" section:

Pulmonary: asthma, presented to ER multiple times this year for RAD, never intubated, records indicate anxiety likely cause of breathing difficulties on Nov ER visit, chronic dry cough for 1 month with sweats.

The document also contained a list of medications[x]:

tradazone, effexor, clonazepan, combivent, prednisone taper [for chest wall inflammation], levoxyl, tegretol, topamax[xi]

Finally, the document included an indication that the lung exam was normal:

Lung Exam: WNL

[ix] MDI is the acronym for "Metered Dose Inhaler". One-half MDI would represent use of ½ of the medication in a dispensed inhaler, which approximates about 100 puffs (doses) of inhaler medication.

[x] The list contains spelling errors, and a combination of trade and generic names

[xi] **Tradozone** is used for the treatment of depression, panic attacks, and other behavioral symptoms.
Effexor is used to treat major depressive disorder, anxiety and panic disorder.
Clonazepam is used to treat panic disorders and convulsive disorders such as epilepsy.
Combivent is an inhaler used for COPD, bronchospasm and asthma.
Prednisone is an oral steroid used for mitigation of immune responses.
Levoxyl is a thyroid hormone used to prevent or treat goiter.
Tegretol is used to treat seizures, nerve pain and bipolar disorder.
Topamax is an anticonvulsant.

The three coders highlighted the following text from the pulmonary passage and provided the following answer to the question "Is this patient's asthma well-controlled?":

Coder	Highlighted text	Asthma Controlled?
Student #1	asthma, presented to ER multiple times this year for RAD, never intubated, records indicate anxiety likely cause of breathing difficulties on Nov ER visit, chronic dry cough for 1 month with sweats	Yes
Student #2	asthma, presented to ER multiple times this year never intubated indicate anxiety likely cause of breathing dry cough for 1 month	No
Expert #1	asthma, presented to ER multiple times this year for RAD Nov ER visit	No

Table 3: Coding of pulmonary notes in a pilot study

The three coders highlighted the following segments of the medication list:

Coder	Highlighted text	Asthma Controlled?
Student #1	tradazone, effexor, clonazepan, combivent, prednisone taper [for chest wall inflammation], levoxyl, tegretol, topamax	Yes
Student #2	prednisone	No
Expert #1	combivent, prednisone	No

Table 4: Coding of medications in a pilot study

Finally, the three coders highlighted the lung exam section in this manner:

Coder	Highlighted text	Asthma Controlled?
Student #1	Lung Exam: WNL	Yes
Student #2	<not selected>	No
Expert #1	<not selected>	No

Table 5: Coding lung exams in a pilot study

Following this exercise, the three coders discussed their markup. Interestingly, and not unexpectedly, all had different reasons for coding what they did. There is a great deal to discuss qualitatively about these results; however, several potential confounders were noted, and are discussed here.

The author was the only one of the three coders who looked at the date of the preoperative visit (which is stamped on the form). Since it was in December, and the note indicated that the patient had been in the ER in November for asthma, I concluded the asthma was poorly controlled based on that single fact, with the compelling information that the patient had multiple ER admissions over the last year. I considered the medication list informative for the presence of an inhaler, and prednisone, which, although it was being used for a chest wall inflammation, might have controlling effects for current asthma. In the absence of dose or frequency information on the medications, I considered the list otherwise uninformative. Finally, a normal lung exam during the visit only indicated to me that the patient was not currently having symptomatic asthma. I did consider highlighting "*chronic dry cough for 1 month with sweats"* but in the absence of information indicating whether or not this caused the patient to wake at night (which was required to meet the provided definition of well-controlled asthma), I elected to omit that phrase.

Student #1 highlighted the most text. He elaborated his reasoning as to why he thought the patient's asthma was well controlled based two findings: 1) a normal lung exam during the visit, and 2) because the asthma episodes were anxiety-related, the student felt that the medication list including at least 3 medications for anxiety, demonstrated that the

patient's anxiety must be well-controlled, meaning the asthma must be well-controlled as well. I found this particularly interesting in that neither of these two factors was described in the definition of well-controlled asthma that I gave him (and he had access to during the markup session).

Finally, student #2, herself an asthma sufferer, thought that whether or not the patient had been intubated indicated a measure of control for asthma. She did not consider a normal lung exam on the day of the preoperative visit suggestive either for or against current asthma control. During our discussion, the student suggested that she might have added *SPO2: 98% on Room Air* (a value entered under the "Physical Exam" section of the document) as an indicator that the patient was somewhat debilitated in moving oxygen (e.g., the student expected "normal" to be 99-100%). Again, this particular feature was not among the items in the definition. Finally, due to her personal experience with asthma, she highlighted "*dry cough*" because this had been a symptom she noted prior to her own asthma attacks. This coder selected *prednisone* from the medication list, even though the list indicated it was not currently being used to treat asthma, but missed *combivent* due to lack of familiarity with the medication.

From this simple pilot exercise, it became clear that:

- Peoples' reasoning about facts can cause them to ignore the instructions
- Peoples' personal experience with a disease can cause them to ignore the instructions
- Some people, in an effort to be complete, may highlight more text than is necessary to comply with the instructions

- Some people, in an effort to be succinct, may highlight only that text that definitively, in their opinion, answers the posed question.

Evaluating Text Overlap

Observed agreement between a pair of subjects, for this study, was calculated as:

$$\frac{Total\ overlapped\ characters - Total\ characters\ not\ selected\ by\ either\ subject}{Total\ characters\ in\ document}$$

Expected agreement between a pair of subjects was calculated using a standard r by c (2 by 2) contingency table where data were arranged as follows:

		Subject 2		
		Selected	**Not Selected**	**Column Totals**
Subject 1	**Selected**	Selected by both (a)	Selected by 1 and not 2 (c)	a + b
	Not selected	Selected by 2 and not 1 (b)	Not selected by both (d)	c + d
	Row Totals	a + c	b + d	a + b + c + d

Table 6: Demonstration table for calculation of expected agreement

The contents of this table are discussed below.

Expected agreement, calculated from this table follows these steps:

Expected agreement for cell a: $\frac{(a+b)(a+c)}{(a+b+c+d)}$

Expected agreement for cell d: $\frac{(c+d)(b+d)}{(a+b+c+d)}$

Overall expected agreement = Expected agreement cell a + Expected agreement cell d

Ogren, et al[94] have utilized an annotation technique that differs from that proposed for this study. For the experiment, the authors had coders identify text, assign a SNOMED concept code to the text, assign a context code (one of *current, history of,* or *family history of),* as well as a status code (one of *confirmed, negated,* or *possible).* The analysis utilized these match criteria to generate K-statistics only if there was some textual overlap:

- Spans of text are identical
- Spans of text overlap
- Spans of test overlap and the concepts (SNOMED) match
- Spans of text overlap and the assigned contexts match
- Spans of text overlap and the assigned status codes match
- Spans of text overlap and the SNOMED concept codes, contexts, and statuses match.

These metrics are appealing, and are used to guide the quantitative measures of text agreement for this study. For the two hypotheses, the null and alterative hypotheses do not differ from the previous, document-level analyses, so they are not restated here. And, based on the preceding discussion, Kappa scores can be calculated on the following span data for each of the hypotheses:

1. Spans of text are identical
2. Spans of text overlap
3. Spans of text overlap and the answers to the simple and complex questions match.
4. Spans of text overlap and the answers to the simple and complex questions do not match.

These statistics were used to estimate agreement. However, more in-depth analysis of phrase level data suggested that each hypothesis had one or more corollaries, and that these corollaries could be tested with Kappa statistic comparisons, and simple counts.

Hypothesis #1: Clinicians will demonstrate higher inter-rater agreement among themselves than will non-clinicians.

- Corollary #1: Clinicians will select more similar (e.g., overlapping) spans of codes than non-clinicians (e.g., clinicians will demonstrate higher instances of #1 and #2 in the list of Kappa calculations, above).
- Corollary #2: Clinicians will show more instances of overlapping code spans with agreeing answers on simple and complex questions than will non-clinicians (e.g., clinicians will demonstrate more instances of #3, #4, and #5 in the Kappa calculation list, above).

Hypothesis #2: When coding free-text documents, inter-rater agreement for both clinicians and non-clinicians is higher when the task is simple and lower when the task is more complex.

- Corollary #1: Clinicians and non-clinicians will select more similar (e.g., overlapping) spans of text when the question is simple than when it is complex (e.g., both subject groups will demonstrate higher instances of #1 and #2 in the list of Kappa calculations, above).

4.8. The Study

4.8.1. Training Questions

Several common subject questions arose during the training sessions. Fortunately, most of these were anticipated as a result of the brief pilot study completed prior to this formal work. These questions generally concerned definitions of terms, and more specifically what was meant by "clearly state" (when asking if the document clearly stated the patient smoked or had well controlled asthma), "in your opinion", or "well controlled" (when asking both asthma questions). Because this study was intended to explore the cognitive differences between non-clinicians and clinicians, great care was taken to *avoid* carefully defining these terms because providing specific definitions might cause both groups to converge on the definition(s) chosen by this researcher and not the subjects' current knowledge or "sense" of the concepts. For example, "smoke" can mean "to inhale" burning materials such as tobacco, marijuana, methamphetamines, etc., or "prepare" meats for eating, "pitch a fastball" in baseball, or "best" someone in a competition (as in "I smoked 'em!"). Clearly context dictates a more clinically-relevant interpretation here, but still, substantial room was *intentionally* left for subject interpretation of terms. As a result, several answers were prepared by this researcher ahead of time to questions about definitions. The following table summarizes the most common questions received, and the answers provided to subjects:

Question	The Researcher's General Answer
What does "clearly state" mean? (Asked by both non-clinicians and clinicians.)	It means that you can locate text in the document that answers the question. How much, and what text to include is up to you.
What does "well-controlled" mean? (Asked by non-clinicians, for the most part.)	If a person with chronic pain has pain that is "well-controlled" what does this mean to you? Use that definition to answer the question here.
Are you asking for my personal or my medical opinion? (Commonly asked by clinicians.)	No medical decisions will be made based on the answers you give. So, it is up to you which you use. The researcher asks only that the subject be consistent in the type of opinion expressed.

Table 7: Prepared answers to anticipated questions

This table demonstrates the general format of the prepared answer to anticipated questions.

4.8.2. Data Collection

Training began on March 2, 2009 and completed on March 11, 2009. The last subject completed entering data on Wednesday, March 25, 2009, at which time the web server daemon granting access to the web site was halted, so that no further access of the study web site was allowed.

4.8.3. Data Preparation

Introduction

In all cases of data collection, some data review and correction are necessary to accommodate subsequent statistical analysis. Before this review commenced, however, the entire database was written *without modification* to a file capable of restoring the complete database to its original form. The file was verified to assure it would produce an exact copy of the original database.

Following this, simple data cleaning involved removing all data (answers to questions, user-specific documents with mark-up, text-snip selections and associated comments, and login information) for all non-study subjects including members of my dissertation committee, two student testers, all data entered by me during the design, testing and debugging of the web site, and multiple logins created while testing the login scripting routines.

Verifying Counts

To verify the completeness of the data, simple statistics were run to evaluate counts. For example, the study design included three training documents (two smoking and one asthma) followed by 60 documents (30 smoking and 30 asthma). As discussed, each document asked the subject to answer two questions, and to identify the text in the document that helped them arrive at their answer. Simple calculations revealed that for 17 subjects (9 clinicians and 8 non-clinicians), each reviewing 60 documents with 2 questions, should yield 17 * 60 * 2 or, 2040 answers. Instead, there existed a total count of 1,854. In addition, because of the design of the interface, each time a user marked up a document to select text snips for an answer, a copy of the document was saved for that answer. (This enabled the researcher to distinguish between similar or identical text snips selected for both questions.) These documents were stored as user documents along with their associated identifiers in a separate table. For each subject, then, there should have been two documents across the entire set of 60 documents or 17 *2*60 or 2040 total documents. This disparity required investigation.

The first important discovery was that, for one clinician subject, answers to questions existed for only about half of the documents in the study. After a follow-up discussion with this subject it was determined that the best explanation for the data loss was that he used a beta version of the Safari browser for his Macintosh computer from home. Somehow, the web site displayed only every other document to this subject. The decision was made to remove the data entered by this clinician from the study, as it was incomplete However, this failed to fully resolve the observed document and answer counts with what was expected. There were both too few documents, and too few answers remaining.

Further investigation demonstrated that despite entering the documents into the database using an automated sequential indexer, I had inadvertently skipped a number when I assigned document order identifiers to the documents. As a result, asthma documents, which had ids of 0 to 30 had document orders of 0 to 31. I had skipped from document order 23 to document order 25. Because the interface was driven by this document order, and there was no document 24, the interface automatically incremented a counter and displayed document 25, correctly. This caused the study to also end one additional document early, because the study was designed to show document ordered 30 as the last document, not document number 31. As a result no subject answered all 30 asthma documents; each subject answered 28 instead. Once this discovery was made, the answer and document counts were verifiable as correct. At this point, a second, restorable copy of the database was created and saved.

Normalizing Answers

The original interface design stored question answers, verbatim, in the database. As a result, answers required normalization to support comparison of like items to like items. The following answer transformations were implemented for this process:

Document	Question	Original Answer	Standardized Answer
Smoking Q1	Does the text clearly state that the patient smokes?	smokes	yes
		does not smoke	no
		not sure if smokes	not sure
Smoking Q2	In your opinion does the patient smoke?	yes	yes
		no	no
		not sure	not sure
Asthma Q1	Does the text clearly state that the patient's asthma is well controlled?	controlled	yes
		not controlled	no
		not sure asthma	not sure
Asthma Q2	In your opinion is the patient's asthma well controlled?	well controlled	yes
		not well controlled	no
		not sure	not sure

Table 8: Normalization of question answers

Each subject had 116 answers (58 total documents reviewed, with two answers per document) in the database after erroneous records were removed, as described in the previous section. All normalized answers were saved to a file, of which a backup copy was made. All subsequent answer count analysis was based on this normalized file.

Cleaning and Mapping Snip Selections

For usability and to help differentiate between the first or second question for each document text snip selections both in the viewed document and the current question box were highlighted with yellow (for question 1) and green (for question 2). In addition, a copy of each document was saved for each question, containing the markup of the text selected by the user Each of these saved documents required post-processing to determine

the actual character offsets of the selected text against a standard (e.g., not marked up) version of the document. Each selected text snip had an order number associated with it; allowing processing algorithms to determine the correct match in the document in the event a subject selected repeated text (which was common). Once these offsets were calculated, the HTML markup was removed, and the offsets were re-tuned to account for the removal of the markup and to assure their correct match to the *original, not marked-up version of the document*. This effectively standardized all user responses for a given document to the original document displayed to the user.

Because this standardization was performed computationally, it also included several text processing steps to remove extraneous detail. The following modifications were made:

- Blank snips were removed
- Snips consisting of only HTML markup were removed
- Leading and trailing spaces were trimmed
- Leading and trailing punctuation were trimmed
- Leading and trailing HTML markup was removed
- Snips consisting of a single letter were removed
- Whole words were reconstructed at the beginning and ends of phrase selections where subjects had begun or ended their phrase markup in the middle of a word.[xii]
- In those cases where a subject selected multiple instances of the same text (such as when "Asthma Exacerbation" existed in the Problem List as well as in the

[xii] The algorithm to do this was very simple: If the phrase does not start on a word break, back up until a break is found. If the phrase does not end on a word break, move forward until a word break is found. Then recalculate the true snip offsets.

Assessment or Plan), offsets were recalculated to capture the correct instance of the repeated text in the original document.

Obviously, each of these processing steps recalculated the correct offsets of the remaining text in the original document.

There were 4,270 snips in the original dataset; the modifications undertaken above resulted in 3,878 total "cleaned" text snips. These snips along with all of their identifiers (document id, user id, question id, snip order id, comment id) were saved to a file used as the baseline for snip analysis. The file was manually reviewed by the author on line-by-line to verify the results. Four subsequent instances of extraneous HTML markup were noted and removed, along with manual correction of the associated document offsets, resulting in 3,874 total text snips. The normalized data were saved to a restorable database file. All subsequent snip analysis was run against this file.

Processing Comments Data

Comments contained no markup and required no offset calculations. As a result, 4,270 total comments were easily saved in a file with their associated identifying information (document id, question id, user id, associated snip id, and comment order). Because the analysis of comments was primarily qualitative in nature, the file was imported into an Excel spreadsheet, with all comments sorted by user type, document id, question id and snip order.

A preliminary review of the actual data demonstrated the presence of many blank comments. Users were not required to enter comments when selecting text snips; indeed,

many subjects simply selected a question answer and then one or more text snips to support that answer without elaborating with comments.[xiii] In addition, imprecise use of the mouse by subjects, and programming imperfections lead too occasional empty selections consisting only of HTML markup - most notably: line breaks (
) and hard spaces ( ) which did not serve as useful comments. Removing blank comments and extraneous markup as just described resulted in 1901 remaining comments for the 3,874 final text snips selected. This final set of comments was saved as a backup.

Methods for Qualitative Analysis

A pure grounded theory approach to this analysis would require that all comments be read independent of their association with a question, answer, or document type. This approach proved uninformative in the absence of the contextual information surrounding the comment, precluding a purely grounded approach. As a result, comments were organized with their associated contextual data (e.g., document id, subject type, document type, question number, answer, snip order, snip, and comment) as follows:

```
66|E|Asthma|2|yes|0|SNIP: [At last visit, she presented with an
URI with asthma exacerbation. These symptoms have fully
resolved. The patient presents with no new complaints.]
66|E|Asthma|2|yes|0|COMM: []
66|E|Asthma|2|yes|1|SNIP: [Note: Lungs clear today. URI symptoms
resolved]
66|E|Asthma|2|yes|1|COMM: [It appears that asthma is not a major
problem with this patient and there was simply a recent
(although perhaps not first) episode.]
66|L|Asthma|2|not sure|0|SNIP: [I'm not sure]
66|L|Asthma|2|not sure|0|COMM: [The text mentions that she was
last in due to an asthma exacerbation. But without further
information, that is insufficient to judge whether it was an
isolated incident or a repeated incident, either of which would
indicate whether her asthma is well-controlled or not.]
```

[xiii] In retrospect, this might have been a design flaw in a study of cognition. Comments proved invaluable in the qualitative analysis of data, and there were many instances where the absence of comments made understanding an answer selection, in view of the selected snips, somewhat confusing.

Figure 14: Data formatted for qualitative analysis

Data are organized in vertical-bar-delimited rows containing a document identifier, the subject type ('E' = expert, 'L' = lay subject), the document type, the question identifier, the original question answer, the order of the snip selected to support the answer, a string identifier indicating if the row item represented a snip (e.g., "SNIP" or comment (e.g., "COMM"). Rows were paired to display a snip and it associated comment in the order entered by the user. Blank comments are included in this display. Square brackets around snips and comments served as visual indicators to verify that all leading and trailing spaces had been removed (to assure accurate character counts) and served no other purpose here. (Answer numbers and answers have been normalized for readability.)

This document was imported from UNIX into Microsoft Word®, where it consisted of over 17,200 lines and 96,491 words. A backup copy of this file was made. All subsequent qualitative analysis began with this file.

4.8.4. Data Analysis

All analyses involving interaction with the database for this study took place using the Perl Programming Language (version 5.8.3 for sun4-solaris) running on a UNIX server (5.9 Generic_118558-38 sun4u sparc SUNW,Sun-Fire-V240 operating system). Clean, de-identified data were imported into Microsoft Excel® for quantitative analysis. Most statistical calculations were performed with built-in Excel functions. Perl program results (specifically displaying document and question information, along with selected snips and comments) were saved in Microsoft Word® and manually reviewed for qualitative

analysis. Qualitative work was performed with pen and paper for preliminary content analysis and later theme categorization.

5. Results

5.1. Summary Statistics

5.1.1. Subjects

The study enrolled nine clinicians and eight non-clinicians, exceeding the estimated power to determine statistical significance for the results. The nine clinicians included two registered nurses (one research nurse and a CRNA) and seven Anesthesiologists (one senior resident and 6 staff). Although the study was not powered for eliciting gender differences (and, indeed, the sample size, even with the additional subjects, was too small to detect these), it is notable that the expert group consisted of five males and four females. The lay group consisted of two males and six females. The study did not collect other demographic data.

5.1.2. Data Count Summary

The following table shows the counts of the collected data for the study:

Category	Non-clinicians (n = 8)			Clinicians (n = 8)		
Document type	**Smoking**	**Asthma**	**Total**	**Smoking**	**Asthma**	**Total**
Documents	30	28	58	30	28	58
Questions	60	56	116	60	56	116
Text Snips	1066	1239	2305	900	1044	1944
Comments	576	659	1235	338	330	668

Table 9: Collected data summary totals

A simple χ^2comparing the count of text snips selected by non-clinicians and clinicians reveals no significant difference ($\chi^2(1,N=16)=0.102E-02, p=0.975$). Similarly, there is

no significant difference in the count of comments entered by the groups ($\chi^2(1, N=16 = 2.72, p=.099$).

5.2. Question Analysis

5.2.1. Preliminary Analysis

The analysis begins with a summary of the question answers followed by a comparison of the distribution of answers between study groups for each of the questions in the document sets. For these results, "Question 1" refers to the question asking the user if the document *clearly states* that the patient smokes or has well controlled asthma. "Question 2" refers to the question asking the *user's opinion* as to whether the patient smokes or has well-controlled asthma. Question 1 is considered the "simpler" (e.g., less cognitively complex) of the two questions, and smoking documents are considered the "simpler" of the two document sets.

Smoking Document Question Answers				
Subject Group	**Answers**			
Answer	**Yes**	**No**	**Not Sure**	**Row Totals**
Non-Clinician	287	146	47	480
Clinician	272	148	60	480
Column Totals	559	294	107	960

Table 10: Comparison of answer distribution in smoking documents

A comparison between non-clinician and clinician total answers reveals no significant difference in the distribution of answers among the questions ($\chi^2(2, N=16) = 2.00, p = 0.369$) combined.

Separating out the questions yielded the following results. For smoking question #1:

Subject Group	Smoking Question #1 Answers			
Answer	**Yes**	**No**	**Not Sure**	**Row Totals**
Non-Clinician	141	70	29	240
Clinician	132	69	39	240
Column Totals	273	139	68	480

Table 11: Comparison of answer distribution for smoking question #1

Smoking question #1 is the "simple" smoking question: "Does the text clearly state that the patient smokes?" A χ^2 test comparing non-clinicians and clinicians for question #1 in smoking documents reveals no difference between the groups in the distribution of selected answers ($\chi^2(2,N=16) = 1.77, p = 0.412$).

This comparison was repeated for smoking question #2:

Subject Group	Smoking Question #2 Answers			
Answer	**Yes**	**No**	**Not Sure**	**Row Totals**
Non-Clinician	146	76	18	240
Clinician	140	79	21	240
Column Totals	286	155	39	480

Table 12: Comparison of answer distribution for smoking question #2

Smoking question #2 is the "complex" smoking question: "In your opinion, does this patient smoke?" A χ^2 test comparing answers to question #2 in smoking documents yielded no statistical difference between the subject groups with regard to the distribution of answers ($\chi^2(2,N=16) = 0.415, p = 0.813$).

The same comparisons are made for the answers to questions on the asthma documents:

Asthma Document Question Answers				
Subject Group	**Answers**			
Answer	**Yes**	**No**	**Not Sure**	**Row Totals**
Non-Clinician	187	107	154	448
Clinician	160	107	181	448
Column Totals	347	214	335	896

Table 13: Comparison of answer distribution in asthma documents

A comparison between lay subject and expert total answers reveals no significant difference in the distribution of answers among the questions ($\chi^2(2,N=16) = 4.28, p = 0.118$).

As with smoking documents, each of the answers was evaluated separately. The results for asthma question #1 are below:

Subject Group	Asthma Question #1 Answers			
Answer	**Yes**	**No**	**Not Sure**	**Row Totals**
Non-Clinician	86	40	98	224
Clinician	64	41	119	224
Column Totals	150	81	217	448

Table 14: Comparison of answer distribution for asthma question #1

Asthma question #1 is the "simple" asthma question: "Does the text clearly state that the patient's asthma is well controlled?" There is no statistical difference ($\chi^2(2,N=16) = 5.27, p = 0.072$) between subject groups in answer distribution for asthma question #1.

And for asthma question #2:

Subject Group	Asthma Question #2 Answers			
Answer	**Yes**	**No**	**Not Sure**	**Row Totals**
Non-Clinician	101	67	56	224
Clinician	96	66	62	224
Column Totals	197	133	118	448

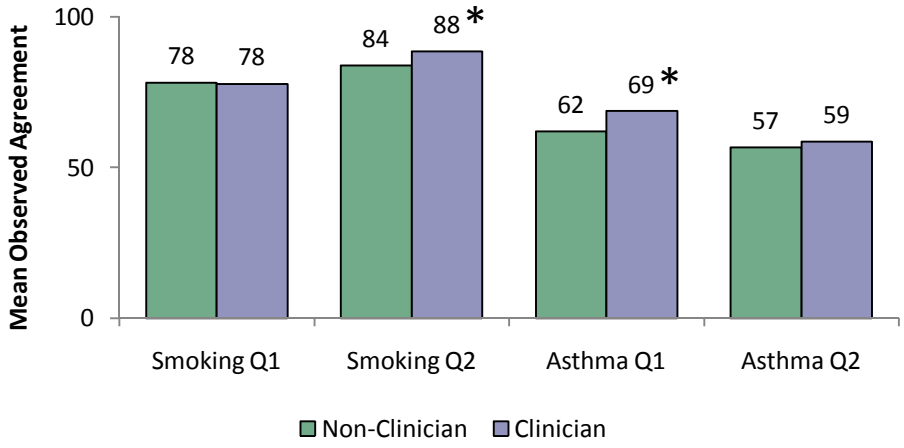
Table 15: Comparison of answer distribution for asthma question #2

Asthma question #2 is the "complex" asthma question: "In your opinion is the patient's asthma is well controlled?" There was no significant difference ($\chi^2(2,N=16) = 0.440, p = 0.803$) between subject groups in answer distribution for asthma question #2.

These preliminary statistics do not reveal much of use with regard to either of the proposed hypotheses: a), that clinicians will demonstrate higher inter-rater agreement among themselves than will non-clinicians, or b) that each group will show greater agreement when the questions are simpler than when they are more complex. Instead, at this point, there is no variation in the overall distribution of answers among questions between groups. We now turn to the evaluation of the level of agreement among study subjects.

5.2.2. Hypothesis #1

To gain better insight into the level of inter-rater agreement *within* groups at the level of question answers, pair-wise comparisons were run first among the members of one subject group then among subjects of the other group and compare the percentages of agreement. Since the first hypothesis asserts that clinicians will agree more often than non-clinicians for answers to all questions in all documents, we expect to see a higher mean rate of agreement among clinician comparisons than among non-clinician comparisons for each of the four questions posed within the two document types:

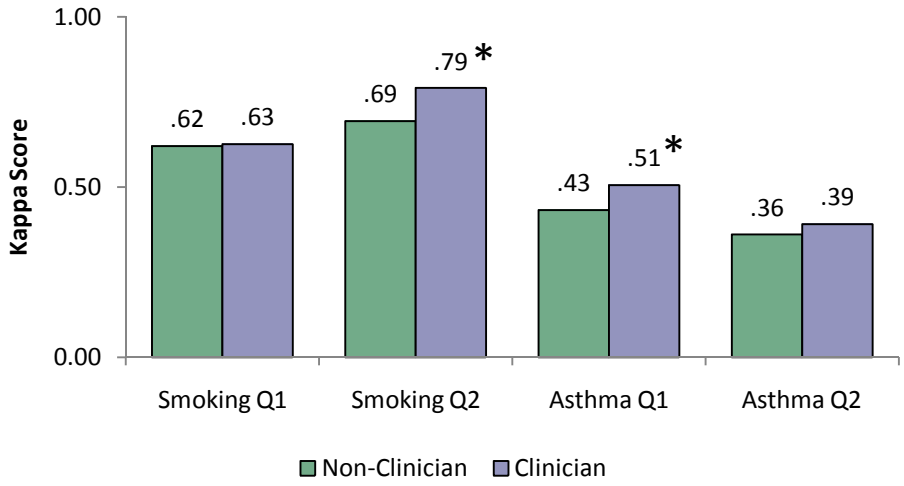


* $p < 0.05$

Figure 15: H_1 comparison of percent observed agreement on question answers

Clinicians show a significantly greater percent observed agreement than non-clinicians for smoking question #2 ($t(54) = 3.477$, $p = .0010$) and asthma question #1 ($t(54) = 2.409$, $p = .0194$). There is no significant difference between the study groups for the remaining two questions.

These results demonstrate that clinicians agree at least equally to or significantly more than non-clinicians. To further analyze the level of agreement, Cohen's Kappa was used. The following figure demonstrates the results of this analysis:



* $p < 0.05$

Figure 16: H_1 comparison of mean Kappa on question answers

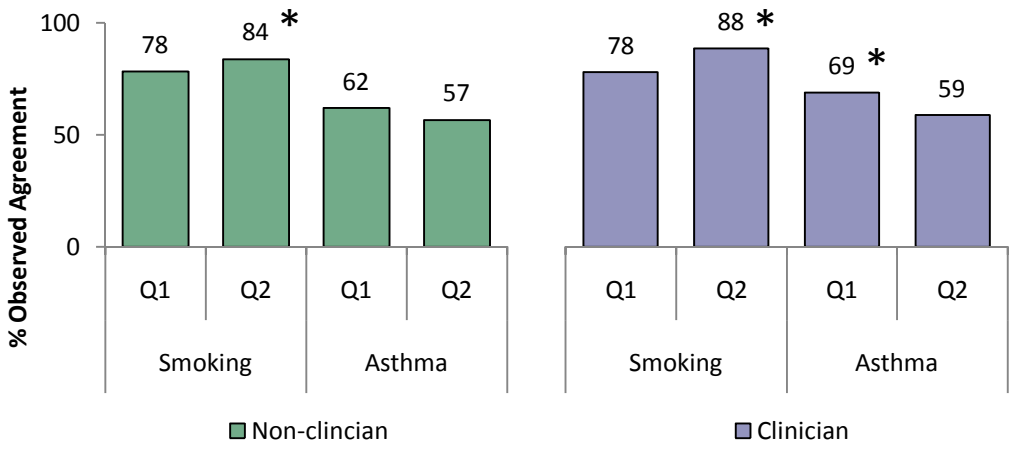
Two of these comparisons yield significant results: smoking question #2 ($t(54) = 4.204$, $p < .0001$) and asthma question #1 ($t(54) = 2.076$, $p = .0426$) and in both cases, clinicians show higher answer agreement than non-clinicians. The results for the remaining two questions reveal no significant differences in group means.

In summary, there is strong agreement between the percent observed agreement and Kappa analyses of the agreement between clinicians and non-clinicians in their answers to questions for both document sets.. Both measures of agreement reveal the same results. Clinicians agree with each other *at least equally or significantly more* than non-clinicians for all four questions. These results will be analyzed in great depth in the discussion section.

5.2.3 Hypothesis #2

Hypothesis #2 posits that regardless of subject group, the rate of agreement will be higher for simpler questions than for complex questions. Thus, under this hypothesis, non-clinicians will agree with each other more on the simple question on a document type than on the complex question on that same document. The hypothesis offers no insight into how clinicians and non-clinicians might agree *across groups,* instead, the hypothesis posits only that *within groups,* subjects will agree more on simpler questions.

To set up this analysis all of the pair-wise comparison percent agreements were reviewed for the simple smoking question ("Does the text clearly state that the patient smokes?"). The comparisons were then matched to their equivalents for the complex smoking questions ("In your opinion, does the patient smoke?"). And, as with the above analysis, this was repeated for both questions in each document type and for each subject group. The goal was to detect a difference in the levels of agreement between the simple and the complex question (e.g. Q1 and Q2) for each document type (e.g., smoking or asthma), for each individual study group (e.g., non-clinician or clinician). The first comparison, as for Hypothesis #1, was for the difference in percent observed agreement:

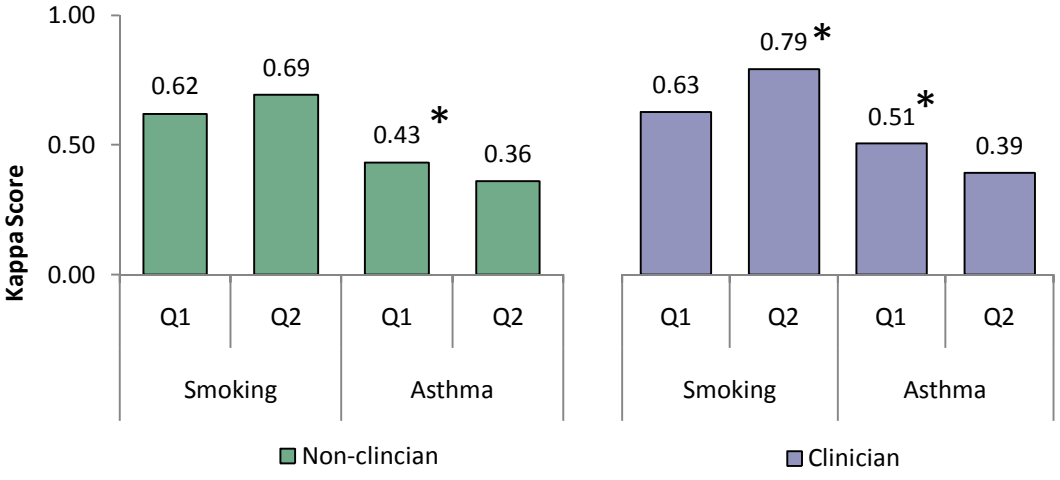


* $p < 0.05$

Figure 17: H_2 comparisons of % observed agreement on questions

Non-clinicians agree more on *complex* smoking questions (t(54) = 2.307, p = .0269) but demonstrate no significant difference in agreement between simple and complex asthma questions. Clinicians agree statistically more on *complex* smoking questions (t(54) = 5.893, p < .0001) and *simpler* asthma questions (t(54) = 4,135, p < .0001).

As with Hypothesis #1, Cohen's Kappa was used as a second measure of agreement:



$p < 0.05$

Figure 18: H_2 comparisons of mean Kappa on questions

Non-clinicians show no significant difference in levels of agreement between simple and complex questions for smoking documents. However, this group agrees significantly more on the simpler question t(54) = 2.114, *p = 0.0319*) for asthma documents. Clinicians demonstrate significantly higher agreement on the *complex* smoking question ($t(54) = 5.694$, $p < .001$) and the *simpler* asthma question ($t(54) = 5.328$, $p = .0018$).

The results were mixed for Hypothesis #2, which posited that all subjects should agree more on simpler questions than more complex ones within a document type. Percent observed agreement revealed the most inconsistent pattern of results. Only one of three significant differences (clinicians, asthma document questions) supported the hypothesis. Two of the three significant differences were significant in the *wrong* direction; that is, the agreement was higher on the *complex* question, not on the simpler one, as expected. Notably, this occurred on the smoking documents for both clinicians and non-clinicians.

In terms of Kappa for Hypothesis #2, the results revealed two significant differences between simple and complex questions: Clinicians agreed significantly more on the *complex* smoking question and the *simpler* asthma question. Non-clinicians showed no significant differences in their responses, though the results revealed a visual trend towards greater agreement for the more *complex* smoking question, and again on the *simple* asthma question, similar to the results for clinicians.

5.2.4. Summary of Question Results

This section has presented the quantitative results of the analysis of the answers to questions in both document sets for each study group. In summary, there is a significant difference between non-clinicians and clinicians in the distribution of answers across questions across documents. Drilling down into this difference reveals that much of the variance involves the far greater number of times clinicians answer "Not Sure" to questions than non-clinicians. The implications of this variance are unclear with regard to the hypotheses (at least at this point).

5.3. Phrase-Level Analysis

5.3.1. Introduction

Please recall that a text "snip" is a string or phrase of text selected by a subject from a document. Each snip indicates information a subject felt was useful to support their reasoning behind selecting a given answer. Each individual snip is identifiable by document type, user id, question id, and order id. The order id represents the order in which the user selected the text. All snips have associated comments, which may or may not have been completed by subjects; that is, adding comments was not required.

For this phase of the analysis, two subjects demonstrated some level of agreement when the text of any of their selected snips sets for a given question overlapped.

To assure broad and deep coverage in the analysis, textual comparisons were evaluated in the context of whether or not the subjects agreed on their answers. That is, the levels of agreement among text snips were evaluated not only when the question *answers* agreed, but also when they didn't agree, between subject pairs. Please recall that *the answers provided by two subjects agree if they are identical, that is, both subjects answer one of "yes", "no", or "not sure" to the question being evaluated.* Recall that all answers are normalized to both simplify and clarify this comparison. A "yes" can be clearly understood as meaning "smokes" in response to the question "Does the text clearly state that the patient smokes?", just as it obviously means "well-controlled" in response to the question "In your opinion, is this patient's asthma well-controlled?" This normalization did not result in any loss of precision.

5.3.2. Document-Level Summary Statistics

Descriptive statistics about text snips for smoking documents are as follows:

	Smoking Documents							
	Non-clinicians				**Clinicians**			
Quest	Total Text Snips	Total Comments	% Snips with Comments	Average Snips/Doc	Total Text Snips	Total Comments	% Snips with Comments	Average Snips/Doc
1	406	192	47.29	6.77	348	129	37.07	5.80
2	660	384	58.18	11.00	552	209	37.86	9.20
Total	1066	576	54.03	17.77	900	338	37.56	15.00

Table 16: Descriptive statistics for smoking document text snips and comments

We have already shown that there is no significant difference in count of snips or comments between clinicians and non-clinicians for smoking documents (see Table 9, section 5.1.2). This table, however, reveals deeper information about the snips and

comments: Comparison between non-clinicians and clinicians reveals a significant difference ($\chi^2(1,N=16) = 37.80, p < .001$), with non-clinicians entering a greater number of comments per snip. There is, however, no significant difference ($\chi^2(1,N=16) = 0.11, p = .74$) in the average number of snips per document between groups.

Similarly, for asthma documents:

	Asthma Documents							
	Non-clinicians				Clinicians			
Ques	Total Text Snips	Total Comments	% Snips with Comments	Average Snips/Doc	Total Text Snips	Total Comments	% Snips with Comments	Average Snips/Doc
1	332	196	59.04	5.53	286	95	33.22	4.93
2	907	463	51.05	15.12	552	209	37.86	9.52
Total	1044	330	31.61	18.00	758	235	31.00	13.07

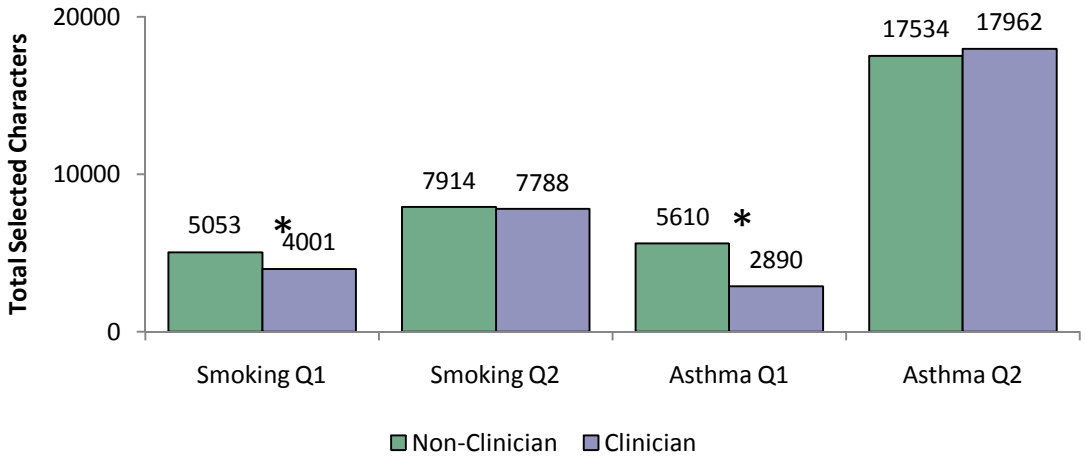
Table 17: Descriptive statistics for asthma document text snips and comments

Comparison between non-clinicians and clinicians reveals a significant difference ($\chi^2(1,N=16) = 82.10, p < .001$), with non-clinicians entering a greater number of comments per snip. Unlike smoking documents, however, there is a significant difference in the average number of snips per document between groups ($\chi^2(1,N=16) = 21.9, p < .001$), with non-clinicians selecting more snips.

5.3.3. Snip-Level Summary Statistics

Total Selected Characters

Although not directly related to the hypotheses, which deal with levels of agreement, it is interesting to consider additional text-level statistics. For example, the following table compares clinicians to non-clinicians by total character count for selected snips:

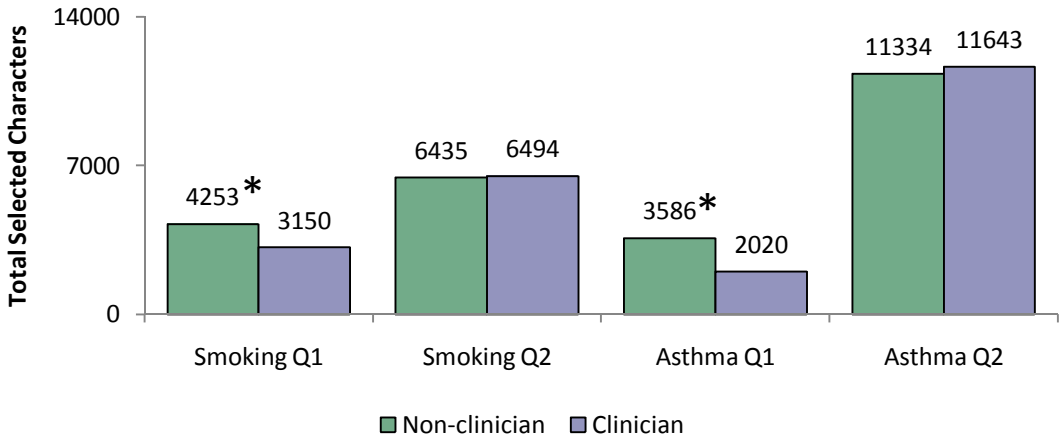


* $p < 0.05$

Figure 19: Comparison of total selected characters between study groups

Comparison between non-clinicians and clinicians reveals a significant difference ($t(54) = 2.78$, $p = .008$), with non-clinicians selecting more characters overall for question #1 in smoking documents as well as significantly more characters ($t(54) = 4.84$, $p < .001$) than clinicians for asthma question #1. There is no significant difference between the subject groups for the remaining questions.

It is interesting to note if the amount of selected text varies depending on whether or not subjects agree on question answers. This is an answer-level analysis, not a document one. That is, subjects do not have to agree on both answers on a document, only one. Here we consider the mean total selected characters among subject pairs showing agreement on single question answers.

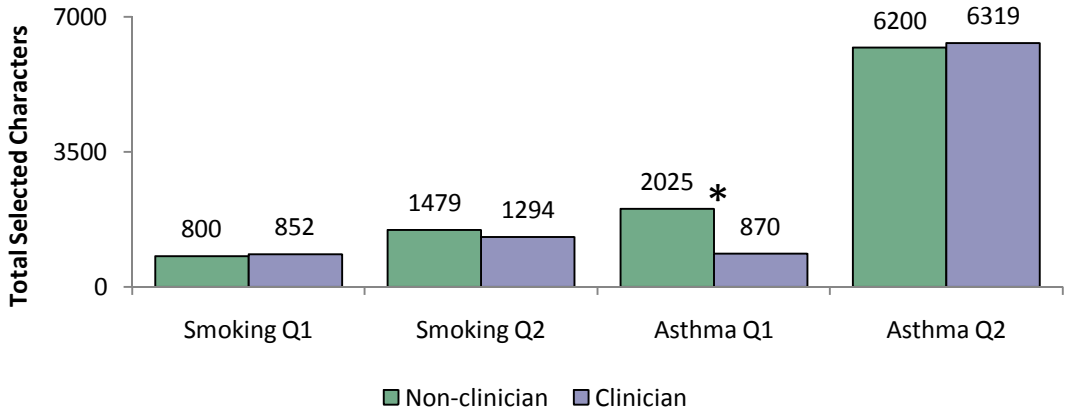


* $p < 0.05$

Figure 20: Total selected characters when subjects agree

This table demonstrates that non-clinicians select significantly more text on smoking question #1 ($t(54) = 3.1252$, $p = .0029$) and asthma question #1 ($t(54) = 4.2879$, $p < .0001$) than clinicians. The remaining results show no significant differences among groups (smoking question # 2: $t(54) = 0.1096$, $p = .9131$, asthma question # 2: $t(54) = 0.2182$, $p = .8281$).

For completeness, both subject groups text selections were compared in the context of disagreement on questions answers:



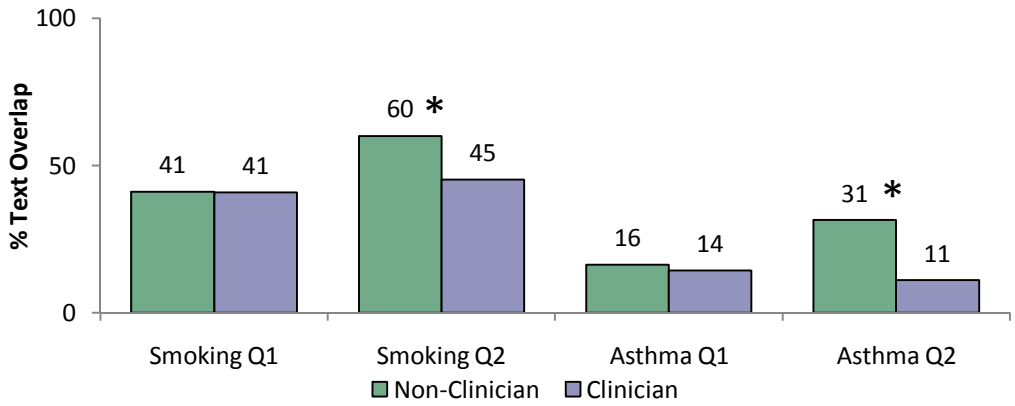
* $p < 0.05$

Figure 21: Total selected characters when subjects disagree

This table demonstrates the differences in mean selected text character count when subjects disagree on their question answers. Only one result is significant, and that is for asthma question #1 where non-clinicians select significantly more characters than non-clinicians ($t(54) = 3.9205$, $p = .0003$), when the subject group pairs disagree on their answers. The remaining differences are insignificant: smoking question # 1: ($t(54) = 0.5578$, $p = .5793$); smoking question # 2: ($t(54) = 0.7794$, $p = .4391$); asthma question # 2: ($t(54) = 0.1562$, $p = .8765$).

Text Overlap

In addition, we evaluate percent selected text overlap for each group:

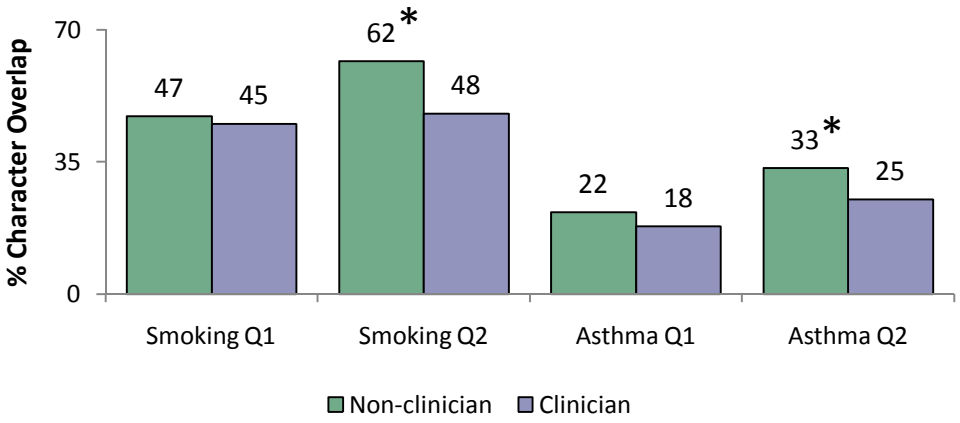


* $p < 0.05$

Figure 22: Comparison of total overlapped characters

Comparison between non-clinicians and clinicians reveals a significant difference ($t(54) = 6.04$, $p < .001$), with non-clinicians overlapping more characters for smoking question #2, and again for asthma question #2 ($t(54) = 3.29$, $p = .003$). There is no significant difference in text overlap for the remaining questions.

As for selected characters, these results can be further differentiated according to how the subject pairs answered the document questions. First, we examine how percent overlap changes when subjects agree on their answers:

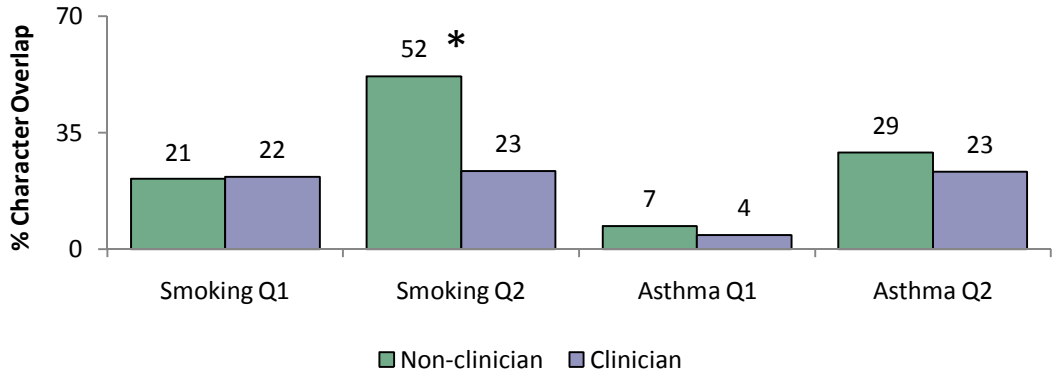


* $p < 0.05$

Figure 23: Percent overlapped characters when subjects agree

This table visually reveals that non-clinicians tend to overlap more characters with one another in all cases, however only two of these results demonstrate significance: smoking question #2: ($t(54) = 5.3385$, $p < .0001$) and asthma question #2 ($t(54) = 2.9677$, $p = .0045$). The remaining differences are insignificant: smoking question #1: ($t(54) = 0.7062$, $p = .4830$); and asthma question #1: ($t(54) = 1.9190$, $p = .0603$).

And, for completeness we look at text overlap when subjects disagree:



* $p < 0.05$

Figure 24: Percent overlapped characters when subjects disagree

Visually, non-clinicians appear to select more overlapping text when they disagree on answers in pair-wise comparisons. Despite these visual cues, only one of these differences is significant, where in smoking question #2, non-clinicians demonstrate significantly greater mean percentage character overlap ($t(54) = 7.0635$, $p < .0001$), than non-clinicians. The remaining results are insignificant: smoking question #1: ($t(54) = 0.1719$, $p = .8642$); asthma question #1: ($t(54) = 1.1029$, $p = .2750$); asthma question #2: ($t(54) = 1.6180$, $p = .1115$).

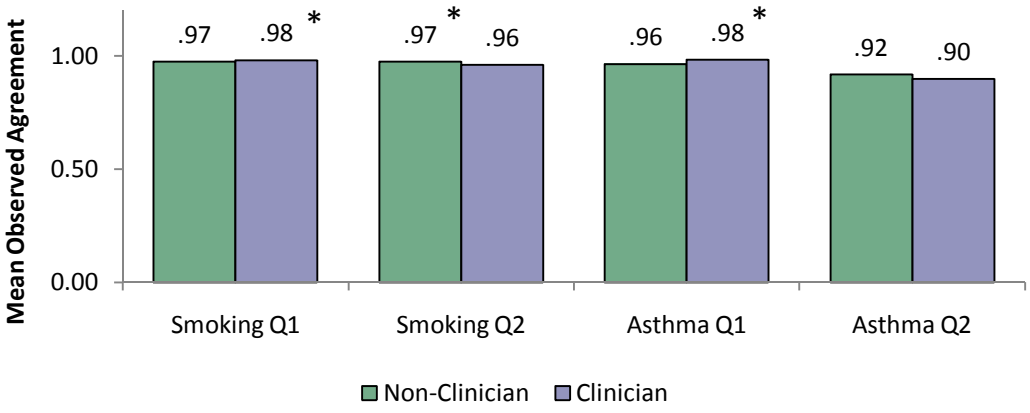
These preliminary statistics are somewhat peripheral to the hypotheses tested in this study, but yield interesting results worth noting in interpreting the study results overall. In terms of overall text selection, non-clinicians select significantly more than non-clinicians. In addition, non-clinicians enter significantly more comments per selected snip. When the selected characters are examined based on whether the subject pairs agreed or disagreed on their question answers, the results are somewhat inconsistent. When subjects agree on their answers in pair-wise comparisons within groups, non-clinicians select significantly more overall characters for only two of the four total questions (smoking question #1 and asthma question #1) than clinicians; there is no significant difference in the groups for the remaining questions. When subjects disagree, non-clinicians select significantly more overall characters than clinicians on only one question, asthma question #1. There is no difference between groups for the remaining comparisons. Thus, in general, non-clinicians select more characters than clinicians. Both groups select more text when they agree than when they disagree.

When these analyses are repeated for the percent of overlapped characters within groups for questions in terms of mean percent of overlapped characters, non-clinicians actually overlap significantly more than clinicians for two out of the four questions (smoking question #2 and asthma question #1). When this analysis is rerun only for the situation in which the pairs agreed, non-clinicians overlap significantly more characters than clinicians, but for only two of the four questions (smoking question #2 and asthma question #2). Finally, when the subjects disagree, non-clinicians demonstrate a significantly higher mean percent overlap than clinicians on a single question: smoking question #2. The remaining results are insignificant. In general, it appears that if there are going to be significant differences, non-clinicians overlap more characters than clinicians.

5.3.4. Hypothesis #1

Please recall that for this phase of the analysis, two subjects agree when the text snips they have selected for a question overlap. Thus, if clinicians are going to agree more often than non-clinicians, they should demonstrate a higher level of text overlap among their selected snips. Summary data (e.g., all snip data included without regard to whether

or not the pairs agreed on the question answer) is presented first:

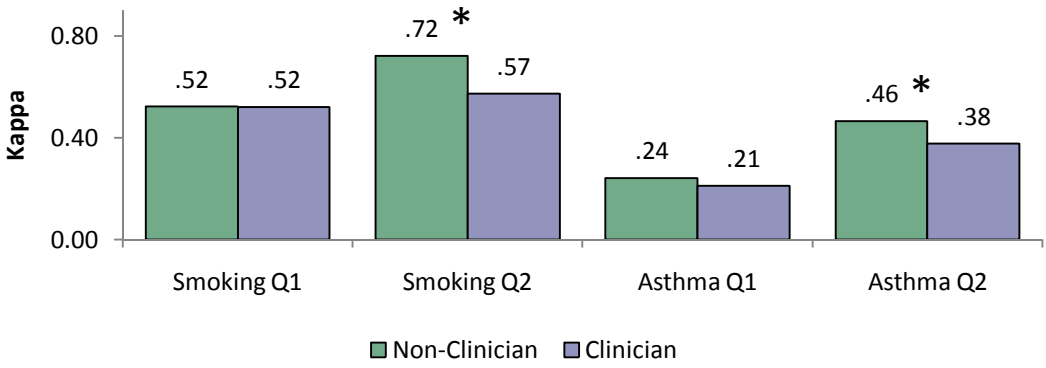


* $p < 0.05$

Figure 25: H_1 comparison of mean observed agreement at the snip level

This table compares the study groups on percent observed agreement within each group for each question on each document. There are significant differences in agreement here, though with no consistent pattern. Clinicians agree significantly more on smoking question #1 ($t(54) = 2.79$, $p = .008$) and asthma question #1 ($t(54) = 5.11$, $p < .001$). However, non-clinicians agree significantly more on smoking question #2 ($t(54) = 2.43$, $p = .018$). Finally, though it appears visually that non-clinicians appear more there is no significant difference between groups.

These same comparisons are repeated using mean Kappa scores within groups:



* $p < 0.05$

Figure 26: H_1 comparison of mean Kappa at the snip level

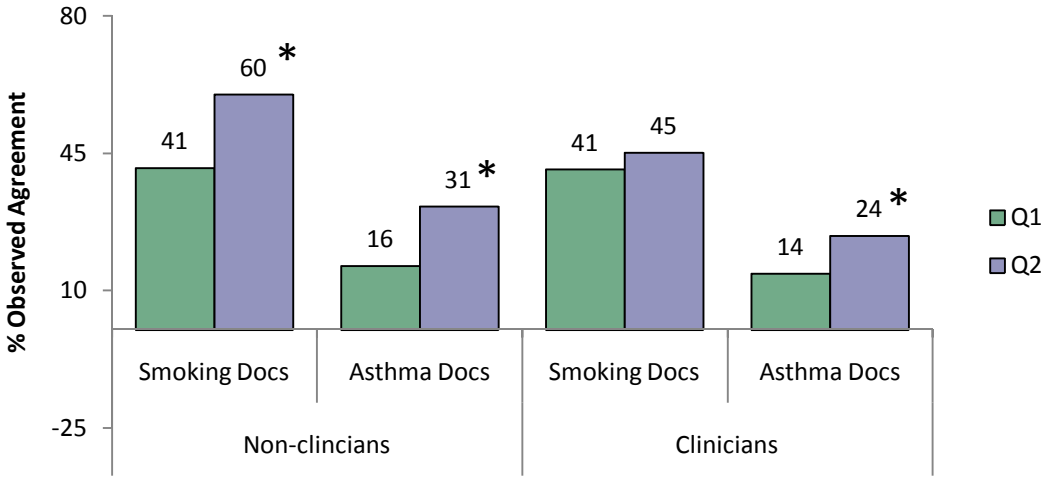
This table compares the study groups using Kappa as the measure of agreement within each group for each question on each document. Non-clinicians agree significantly more than clinicians in terms of Kappa for smoking question #2 ($t(54) = 6.2945$, $p < .0001$) and asthma question #2 ($t(54) = 2.5419$, $p = .0139$). There is no statistical difference in the remaining questions: Smoking question #1 ($t(54) = 0.0449$, $p = .9644$) and asthma question #1 ($t(54) = 1.4308$, $p = .1583$).

The results in support of Hypothesis #1 – that clinicians will agree more often than non-clinician – are mixed. When using mean percent observed agreement as the agreement measure, clinicians agree significantly more on two of 4 questions (smoking question #1 and asthma question #1), non-clinicians agree more on one question (smoking question #2) and there is no difference in agreement for the final question. In terms of agreement as measured by Kappa, non-clinicians agree significantly more than clinicians for two questions (smoking question #2 and asthma question #2); there is no significant

difference noted for the remaining questions. The lack of correlation between mean percent observed agreement an Kappa is surprising, and no consistent pattern of agreement emerges from the analysis.

5.3.5. Hypothesis #2

Hypothesis #2 requires a rearrangement of data to examine within-group agreement between questions on a given type of document. The simple question in a document set should result in higher agreement than the complex question in the document set for each subject group if Hypothesis #2 is supported. Thus, for this analysis, non-clinicians and clinicians are evaluated separately. As with previous analyses, we begin with percent observed agreement:

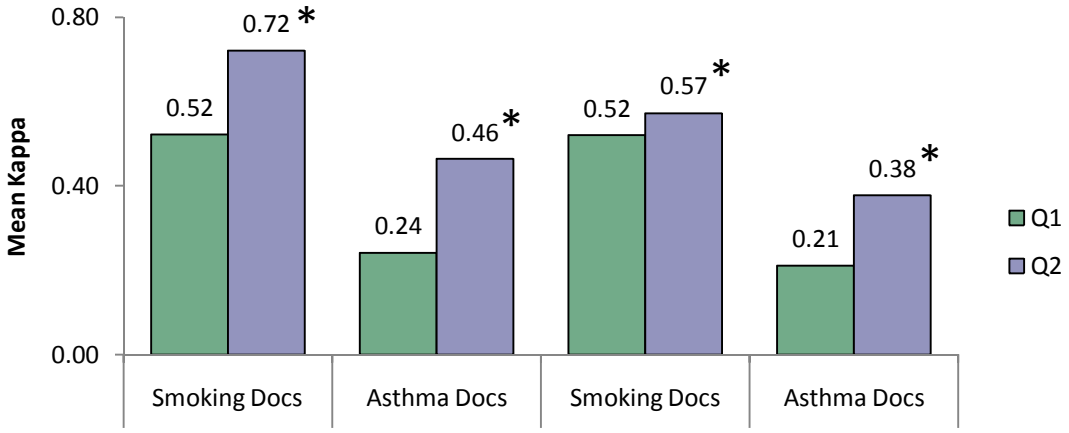


* $p < 0.05$

Figure 27: H_2 mean % observed agreement at the snip level for both study groups

Non-clinicians agree significantly more on the complex smoking question (Q2) than on the simpler question (Q2): ($t(54) = 5.7270$, $p < .0001$). Similarly, non-clinicians agree significantly more ($t(54) = 6.669$, $p < .0001$) on the complex asthma question. Clinicians demonstrate a visual trend towards higher agreement on the complex smoking question, though the result is not statistically significant ($t(54) = 1.9266$, $p = .059$). Finally, clinicians show a significantly higher percent observed agreement on the complex asthma question ($t(54) = 4.0394$, $p = .0002$).

Evaluation of the mean Kappa scores for each group closely parallels the observed levels of agreement observed when we measure % observed agreement.



* $p < 0.05$

Figure 28: H_2 Kappa at the snip level for both study groups

When comparing mean Kappa scores between question answers for non-clinicians, agreement is significantly higher for the more complex (e.g., Q2) in both documents. For smoking documents: $t(54) = 6.4189$, $p < .0001$, and for asthma documents: $t(54) = 8.2533$, $p < .0001$). In addition, under Kappa, clinicians agree significantly more on the

complex smoking questions ($t(54) = 2.1115$, $p = .0394$ and complex question for asthma documents ($t(54) = 5.5537$, $p < .0001$).

The snip-level results regarding measures of agreement under Hypothesis #2 counter the proposed hypothesis by demonstrating statistically higher agreement for complex questions than for smoking questions across two measures of agreement and 7 out of 8 comparisons. In the single instance where results are not significant (for percent observed agreement on the smoking questions for clinicians) the visual trend is consistent with the remaining measures. As with all of these results, this invites further investigation.

First we return to an evaluation of the total selected characters by subjects in groups. Referring back to Figure 19 **(**Comparison of total selected characters between study groups), recall that for the two simple questions (smoking question #1 and asthma question #1) non-clinicians selected significantly more individual characters than clinicians. For the complex questions there was no statistical difference between groups. This suggests that the larger amount of text selected likely captures all of the information related to smoking or asthma necessary to form an opinion, and that because the groups selected such large and similar amounts of text; there may be a larger amount of overlap for the complex questions. Such a suggestion is speculative at best, but has some foundation. Questions asking for an opinion may allow users to cast a wider net in terms of collecting evidence for an answer. This differs markedly from the precision required to indicate where text is explicitly written. This, combined with the fact that clinical documents have an inherent structure with a fairly consistent manner of data presentation

(at least for the document sets used in this study), may confound the results by the fact that most subjects selected most of the text in the document that had any relation to the topic for the complex questions. Again, these results may suggest more that subjects are able to identify *relevant* text, and do so consistently. The motivations for text selection have been discussed as potential confounders for Hypothesis #1; these same considerations apply here. Once again, the study cannot determine, even with the presence of comments, the precise amount of time people selected text because it was *indeed relevant* or because they thought it *might be*. What we can say for sure is that questions asking for a subject's opinion in this study always result in a large amount of text being selected, and that there is a significant amount of overlap, representing agreement in those bodies of text. Whether or not this reflects true agreement or is simply a matter of all subjects selecting text that appears topically appropriate to answer the related question remains unclear.

5.4. Comment Analysis

5.4.1. Introduction

Data Summary

The following table displays overall questions counts and baseline statistics.

Total Comments Entered by Study Groups By Question				
	Non-clinicians	**Clinicians**	**Row Totals**	**Totals**
Smoking Q1	191	129	320	914
Smoking Q2	382	212	594	
Asthma Q1	196	95	291	987
Asthma Q2	462	234	696	
Column Totals	1231	670	1901	1901

Table 18: Comment summary statistics

Preliminary analysis reveals that there is no significant difference between non-clinicians and clinicians in the total number of comments entered for smoking questions ($\chi^2(1,N=16) = 1.90, p = 0.168$). Similarly, no significant difference emerges in the number of comments entered for asthma questions between the two subject groups (($\chi^2(1,N=16) = 0.0877, p = 0.767$).

These results may initially appear to contradict earlier results in section 5.3.2. (Table 17: Descriptive Statistics for Smoking Document Text Snips and Comments). In that table, a significant difference in the amount of comments entered per snip is noted, with non-clinicians entering significantly more comments per snip than clinicians. Here, the table demonstrates only the total number of comments entered, unrelated to the text snips selected.

Patterns of Answers as an Organizing Framework

The lack of consistent statistical support for the hypotheses resulted in a careful analysis of patterns of question answers, to understand why clinicians and non-clinicians appeared to agree, regardless of the simplicity or complexity of the questions. As a result, the answer totals, by subject group, were examined on a document-by-document basis. This

led to the discovery, that when non-clinicians and clinicians disagreed, the most common pattern of disagreement appeared to occur in instances where a majority selected one of "yes" or "no" and the remaining subjects responded "not sure". The following table provides a side-by-side comparison of the answers provided by both subject groups for smoking question #2 ("In your opinion…."):

Non-Clinician Answers			
Doc ID	yes	no	not sure
36	8		
37		8	
39	8		
40	7		1
44	3	4	1
45	6		2
49		8	
50	2	5	1
54	8		
55		8	
56	5	2	1
59	8		
60	8		
61		8	
63	4		4
64		4	4
65	8		
69	8		
70		8	
71	8		
74	8		
78	8		
79	1	6	1
84	8		
86		8	
89	8		
90	6		2
92	8		
94		7	1
95	8		

Clinician Answers			
Doc ID	yes	no	not sure
36	8		
37		8	
39	8		
40	7		1
44		4	4
45	5		3
49		7	1
50		8	
54	8		
55		8	
56	6	1	1
59	8		
60	8		
61		8	
63	4	2	2
64		1	7
65	8		
69	8		
70		8	
71	8		
74	8		
78	8		
79		8	
84	8		
86		8	
89	8		
90	6		2
92	8		
94		8	
95	8		

Table 19: Comparison of answers for Smoking Question #2

The side-by-side comparison of the non-clinician and clinician answers for Smoking Question #2 demonstrates generally strong agreement among subjects in both groups. The table is color-coded as follows: Cells in pale green demonstrate instances of 100% agreement. Cells in gold represent instances of general agreement. Cells in pink represent

cases of general disagreement where subjects answer “yes” and “no” (as well as “not sure”) in response to questions.

A visual inspection of answers for this smoking question revealed a very interesting trend. Independent of the document, question or user type, the level of agreement on the answer was extraordinarily *high*. For this question alone, there were 20 instances of complete agreement for non-clinicians vs. 22 instances of complete agreement for clinicians. The second most commonly occurring pattern was that of “general” agreement (those cells highlighted in gold in Table 20 above). Both clinicians and non-clinicians demonstrated 6 instances each of this pattern. Finally, in 4 instances for non-clinicians and in 2 cases for clinicians, patterns of complete disagreement occurred for this question.

Across all 58 documents, perfect agreement took place 32% of the time (43% or 10 of the smoking documents, and 19% or 5 of the asthma documents) The “general agreement pattern” (noted in gold in the table above) comprised 42% (24) of all documents: 39% (11) were asthma documents, and 43% (13) were smoking documents. In these cases, there was no identifiable dispute as to whether or not the patient smoked (e.g., even if the “not sures” were eliminated, the subjects generally agreed on the answer). The data suggests that subjects differed in the internal degree of “sure enough” in reaching a conclusion, that is, the cognitive scales did not dip conclusively towards one category. The patterns of perfect agreement and “general” agreement accounted for 74% (43) of the documents. There was no single instance of subjects providing an answer pattern

including only “yes” and “no” (e.g., without an associated “not sure”) in any of the document sets. The following table summarizes these results:

Answer Patterns Within Documents				
Subject Type	**Agree[1]**	**Agree/Not Sure[2]**	**Disagree[3]**	**Row Totals**
Smoking Documents				
Non-clinicians	14	10	6	**30**
Clinicians	12	15	3	**30**
Totals	**26**	**25**	**9**	**60**
Asthma Documents				
Non-clinicians	5	12	11	**28**
Clinicians	6	12	10	**28**
Totals	**11**	**24**	**21**	**56**
Grand Total	**37**	**49**	**30**	**116**

[1] Agree: 100% agreement (all answers the same)
[2] Agree/Not sure: General agreement (one of “yes” or “no” and “not sure”)
[3]Disagree: (all answers given)

Table 20: Patterns of answers within the document sets

This table demonstrates the distribution of answer patterns among the document sets as total number of documents in which each answer pattern occurred. There is a significant distribution in the totals between the document sets ($\chi^2(1,N=16) = 10.8, p = .005$) with pattern one (agreement) higher in smoking (simple) documents, and pattern three (full disagreement) higher in asthma (complex) documents.

The qualitative analysis employed content analysis and card sorting. As discussed earlier, each comment was read with its associated snip, question, question answer, and document type. Thus, each comment was contextually bound. In addition, because this study wished to inform the cognition of agreement, the patterns in the question answers served as a convenient, preliminary grouping for the comments. Lacking better

terminology, these groups are called *perfect agreement*, *fair agreement*, and *disagreement*.

Iterative reading of these comment groups yielded significant insight into the data. For both smoking and asthma documents the following discussion will provide examples of the thematic content that emerged from this analysis. User comments are both quoted and italicized. Some spelling errors have been corrected for readability.

5.4.2. Smoking Documents

When there is perfect agreement

There are 26 out of 60 (43%) smoking documents (thirty documents in each set for each study group) in which the answers perfectly agreed across both questions (e.g., "clearly state" or "in your opinion"). When looking for what these 26 documents have in common, explicit statements of smoking status emerge as consistent points of agreement, as evidenced by how frequently snips such as "*now smoking ½ pack/day*" or "*He continues to drink alcohol and smoke tobacco*" or "*he has cut down on his smoking and is thinking about quitting*" are selected. When reviewing the comments associated with these snips, the comments almost universally state that the selected snips is "*definitive,*" represents "*clear evidence,*" or "*completely supports*" the selected answer.

A closer inspection of the phrases identified as "explicit" by users indicates, without exception, that statements considered explicit (with regard to smoking status) contain a) some notion of time, related to "now", b) the word "tobacco" or "packs" or "packs per day" (and its variants) as an indication of tobacco use, especially cigarettes, or c) some

statement indicating the patient "smokes." Very commonly, though not always, some reference to "how long" is also included.

In the absence of explicit statements in documents, both non-clinicians and clinicians demonstrate that they use of an *almost identical set of heuristics* to determine if a patient smokes. Both clinicians and non-clinicians highlight all three of types of snips listed below, together, when they occur:

1. A Diagnosis Code of "*Tobacco Use Disorder*"
2. Some mention about smoking, quitting (a desire to, success at, difficulty with, etc.), "cessation" (e.g., "*Having a tobacco cessation reinforced indicates to me that the person is still smoking.*"), or the string literal "*Tobacco Use: education and counseling provided*"
3. Chantix (or its trade name equivalent, look up)

In the snips reviewed for smoking documents, there is *no exception* to the above two rules for smoking documents where all users agreed on both questions.

When there is fair agreement

The next group of documents consists of those where all users divided between one of "yes or no" and "not sure". This second group comprised 25 documents (or, 42% of all smoking documents). As before, comments were reviewed in the context of their associated snips, user type, document id, question id and question answer. Very quickly, the reason there were only two types of answers for these documents was because the group that answered "yes" or "no" was *sure*, and the other group was "not sure" *enough.* In each case, on review of the comments, the single cause for lack of agreement among

subjects in this category involved term precision, or how a concept was defined, and what bounded or clarified it endpoints. This is best demonstrated by example. Consider smoking document #61, below, and the distribution of answers among subjects:

Smoking Document			Question Answer		
Doc ID	**Question**	**User Type**	**Yes**	**No**	**Not sure**
61	**1**	**Non-clinician**		7	1
		Clinician		7	1
	2	**Non-clinician**		8	
		Clinician		8	
		Totals	**0**	**30**	**2**

Table 21: An example of "fair" answer agreement among subjects

The counts for this document reveal very little disagreement (14 of 16 subjects thought the patient didn't smoke, and two weren't sure). Cells highlighted in green represent instances of 100% agreement. Cells in gold demonstrate generally good answer agreement.

Deeper investigation of the text snips and their associated comments reveals a single point of confusion: the document phrase "*Tobacco Use: Quit five yrs ago.*" One subject answered "not sure" and commented "*…only tobacco use [was] mentioned,*" and the other subject answering "not sure" offered "*[it] does not state clearly either way whether she smokes*." We can reasonably assume the first comment referred to non-tobacco substances that could be consumed by smoking. For the other subject, perhaps the issue is one of *currency*, implying that quitting *5 years ago* does not constitute *current* abstinence. Neither subject offered these explanations for the statements they made, but regardless of whether these presumptions are true, the subjects *carefully considered the*

text they read and wrestled with its meaning - sufficiently enough to propose different answers to the question.

In *every instance* in this group of documents, simple imprecision appears to explain the disagreement among subjects: Subjects comment when terms or concepts do not appear precise. In addition, subjects, lay and expert alike, comment on confusion when they lack heuristics to create an *ad hoc* classification system (e.g. decide if a patient's asthma is well-controlled or not). The following comment, though not from Smoking Document #61, beautifully captures this conundrum, and demonstrates a point of cognitive indecision deciding between "no, this person does not smoke" and "I'm not sure if this patient smokes."

> *She isn't currently smoking, but she has smoked recently and still has cravings, so it isn't clear to me whether to consider her a smoker or not. At what point does someone become a former smoker?*

Every document in this category (the category of pretty good agreement) includes comments exposing the need for clarification regarding all types of terms, such as:

- What defines "smoke" or "smoker"? That is, if a person had, *"…not smoked for 6 1/2 months, other than a couple of drags last week,*" is this person a smoker? Does this patient smoke?
- By "smokes" do we mean "tobacco smoke"? Do you count "*using medical marijuana in the evening*" as smoking? And, is "*chewing tobacco, tobacco use*?"

- What does "*Tobacco Dependence 305.1M*" mean? One user says, *"This suggests that he is currently smoking, although sometimes you carry this DX for life, even if you quit."*
- What does "*Tobacco Use Disorder*" mean? A user offered, *"…tobacco use disorder I take to be an "official" term for smoking, although it weirdly implies that there might be some adaptive use of tobacco use that might not be a disorder. Like eating disorder implies a mal-adaptive form of eating, tobacco use disorder implies that it's not so much that existence of tobacco use that is bad, but that somehow "you're doing it wrong*."
- What does "*Tobacco Dependence*" mean "*with no qualification?"* one user asked, then added, *"Not clear what form of tobacco he uses, so he may not smoke, but since that's the most common form of tobacco use, I'm guessing he does*."

All of these comments illustrate points of disagreement identified by the subjects. The above examples are but a few representing this overarching theme. For smoking, the points of confusion appear to be currency (e.g., Does the patient smoke *now*?), length of time since last cigarette, if smoking means only cigarettes, how long a patient might have a diagnosis code assigned that no longer applies, what some diagnosis codes mean, and whether or not chewing tobacco part of the tobacco disorder umbrella. These documents have been grouped into a single category here because the comments point to confusion caused by imprecise terminology, or poor heuristics for improvising in these cases.

When there is disagreement

Please note the level of disagreement demonstrated in Smoking Document #63, below:

			Question Answer		
Doc ID	**Question**	**User Type**	**Yes**	**No**	**Not sure**
63	**1**	**Non-clinician**	4	1	3
		Clinician	4	3	1
	2	**Non-clinician**	4		4
		Clinician	4	2	2
		Totals	**16**	**6**	**10**

Table 22: An example of disagreement among subjects

For this document, the level of confusion changes depending on the question and all subjects are now divided among the three types of answers. Patterns highlighted in blue show pure disagreement that is, all possible answers were given by the subjects; the cells highlighted in gold demonstrate, at least for non-clinicians for this question, there was "general" agreement, where the pattern of a single definitive answer, coupled with "not sure" represents the unique answers for this question.

The final group of smoking documents consisted of a 9 documents (15% of the total). For this group, subjects have disagreed on their answers completely, that is, all possible answers appear.

For this final document set, imprecise language, again appears to be the primary cause of the disagreement; subjects wrestle with how long a person has to stop smoking to be considered a "*former smoker*." The differences here represent no more than a difference of opinion regarding what that time period is. This document contains the following snip text, which is identified by users and forms the basis of the subsequent comments:

Smoked 4 cigarettes several weeks ago when she was caring for her sick mother. No cigarettes for the last 2 weeks but having cravings. Wondering about options to decrease cravings but worried about cost.

Below is a small selection of the comments elicited by this snip:

- *Text clearly states that she has smoked recently and still has cravings. Not clear whether it's time to consider her a former smoker or not*
- *2 weeks sounds good*
- *I would say she still smokes 2 weeks (especially with cravings) isn't long enough to say she has quit smoking*
- *I read this and it seems as though she does not currently smoke but is struggling with that, she seems to be in a zone between smoking and not smoking*
- *this one is difficult - I guess since her last cigarettes were only 2 weeks ago, I would consider her a current smoker*
- *pt does not smoke at time of visit, but has in recent past and has currently cravings, which mean that she could smoke again tomorrow - to me, that's a smoker*
- *I would treat her as a former smoker at this point, unless she continues to relapse.*

This particular document demonstrated a tremendous difficulty with an issue of time; specifically, how long a patient has to stop smoking before that patient is considered a non-smoker. Because these comments come from both clinicians and non-clinicians it

becomes clear there is confusion about what defines a person as a "*former smoker*", and how one indicates whether or not a patient smokes *now.* However, despite the fact that this document displays the confusing answer characteristics, the issue that remains, is one of concept precision.

Summary

One document in the smoking collection contained multiple, contradictory statements. Smoking Document #90 stated: "No tobacco, ETOH, drug abuse history," "patient admitted to being a closet smoker (used 7 cigarettes)," and "will have patient try and quit smoking". Despite the contradictory evidence, users agreed that the patient either smoked or it was not clear if the patient smoked or not. This finding would be interesting to pursue because it implies that people possess heuristics to weigh evidence towards a reasonable conclusion, even when faced with contradictory evidence. However that remains outside the scope of this study.

Finally, in the smoking document set, subjects tend to repeatedly select very similar text; there is almost always overlap. Thus, subjects also don't appear to disagree about the textual data *per se*, only that imprecise language can be confusing. Subjects identify and describe where topics aren't clear, but otherwise do not demonstrate strong cognitive differences in how they select text to support question answers.

5.4.3. Asthma Documents

Introduction

The three categories used to describe the smoking documents above divide the asthma documents into three groups for qualitative analysis. Of 56 total asthma documents (28 evaluated by non-clinicians and 28 evaluated by clinicians), 11 demonstrate perfect agreement, 24 show overall agreement in combination with "not sure" responses, and the remaining 21 documents comprise the third group demonstrating the lowest level of agreement.

When there is perfect agreement

There were 11 examples (19%) of 100% agreement in answers among subjects reviewing asthma documents. One hundred percent agreement in subject responses results, *without exception,* when subjects can identify an explicit statement that answers the question. So, for the simple asthma question, "Is this patient's asthma well controlled?" text associated with an answer of "yes" for documents in this category included examples such as: "*60 year hx of asthma, which is currently completely controlled,*" "*stable controlled asthma*", "*asthma, mild, intermittent, well controlled: yes*", and "*in regards to her asthma, the patient seems to be quite stable.*" These same text selections are also highly associated with the second, or more complex question ("In your opinion, is the patient's asthma well-controlled?"). In addition, documents containing results of the Asthma Control Test, especially when the test results demonstrated scores above 20, appeared in this category. As with smoking, subjects commonly use terms such as "*definitive,*" "*clear*" or "*irrefutable*" to support snip selections.

Unlike smoking, and likely due to the small sample size comprising this category of asthma documents, no consistent pattern of heuristics such as "this symptom, plus this medication, plus this statement is equivalent to well controlled asthma" emerged from the data. This finding does not prove the lack of such heuristics, only that our sample size was not large enough to see such patterns emerge, if they do exist.

An interesting phenomenon did emerge in the investigation of the comments in these instances of perfect agreement: despite identifying explicit and irrefutable text, subjects almost always select additional text to validate the answer. Even in the presence of a statement such as "*the patient's asthma is well controlled,*" subjects add snips such as "*he isn't having any problems with his asthma*", "*the pt states she is feeling much better*", or "*he isn't using his inhaler as much*." These additional strings are certainly additional evidence, and provide a sort of linguistic insurance, but are not entirely necessary given the other evidence selected by subjects. This observation is simply noted; it does not appear to directly apply to either hypothesis motivating this study.

When there is fair agreement

Twenty-four asthma documents (41%) demonstrated high levels of agreement in this review. Not surprisingly, as with smoking documents, the pattern of observed disagreement appears to result from confusion about the meaning of terms. A simple example demonstrates this point: An expert user has selected two text snips:

- *with regards to her asthma, she does not require ongoing maintenance therapy*, and
- *30 y.o. female with mild, intermittent asthma.*

When asked if the text states if the patient's asthma is well controlled, the expert answers "not sure" and adds these comments:

- *[It] does not state that she does not use meds or have flare-ups*",
- "*Is mild, intermittent asthma the same as well controlled asthma?*

This particular expert does not demonstrate willingness to make a sure decision based on the selected evidence. The problem appears to result from term imprecision: "well controlled asthma" consistently needs clarification by *both* subject groups.

Again, as with smoking documents, many points of uncertainty demonstrated in the comments revolve around temporality, more specifically, "What window of time constitutes 'now'?". Similarly, a text snip such as "*she has asthma, which has been quiescent recently*" causes subjects to comment "*I have no idea what 'recently' means*", "*this isn't a particularly strong statement*" or "*this is an empty statement, essentially…meaningless.*" One subject even comments after selecting the snip "*asthma – severe and persistent*" that "*[the selected snip] states the patient has asthma and it's severe and chronic, but not whether it is under control.*" I assume the point of this subject's comment is that there is no statement of level of control *now* otherwise I cannot make sense of the response. For this reason, I see this comment as reflective of an issue of time precision.

The comments for asthma documents clustering in this category reveal some very explicit heuristics used by subjects *from both study groups* for determining whether a patient's asthma is well controlled, when no otherwise clear and explicit statement exists. Examples of these carefully defined rules follow:

1. If there is no change in treatment for asthma, then the asthma is well-controlled
2. If there is no mention of the asthma, when some form of asthma is listed in the diagnosis list, then the asthma is well-controlled
3. Make sure it's really asthma first, and then answer the question.
4. Patients using oral rescue-type inhalers must not be well-controlled

Overall, the analysis reveals confusion about concept definitions and difficulty in determining what constitutes an appropriate time window representing "current status". There does not appear to result from domain confusion, where lack of knowledge of what constitutes level of asthma control might explain the disagreement, but instead from confusion of interpretation and the inability to find complete, clear, and explicit language. For this document set, as well as for the smoking documents, domain expertise appears to play little role in decision-making for either subject group when the point of contention is an imprecise or undefined term or a question of time boundary.

When there is disagreement

Twenty-one (36%) asthma documents revealed poor inter rater agreement among subjects. One might expect disagreement about the complex clinical concept "well controlled asthma" to explain the largest amount of subject disagreement. This does not, however, appear to be the case.

Interestingly, both lay subjects and experts appear to understand that an exacerbation of asthma is a constrained event, often occurring as a result of exposure to a specific allergen (e.g., a single ski trip where mother was exposed to cold", or the first cut grass of

the season, a friend's cat, and so on). Comments indicate that subjects know that patients with asthma can have trouble in these "*novel*" or "*rarely occurring*" situations. What subjects don't agree on is whether a single exacerbation means that a patient's asthma is well-controlled or not, especially if the patient's asthma is otherwise well controlled. This is not a disagreement among lay people and experts as groups - even the experts are not sure! Somehow this does not appear to be a clinical question, but instead a question of concept definition regarding what constitutes the bounds on "normal' and how much a deviation from that norm is "acceptable". Of the nine documents in this category confusion about whether or not a single exacerbation affects the status of well controlled asthma accounts for confusion among subjects in 5 documents. Despite this consistent issue, a small sample size makes it difficult to gain deeper knowledge as to why this occurs as frequently as it does.

Of the remaining documents in this group, some interesting observations emerge. In general, the primary concern among subjects is that insufficient information is present to make a current asthma status determination. The first example, a single document in the asthma document set lists "Asthma" in the problem list, then provides no subsequent mention of anything related to the patient's pulmonary status. Subjects *always* highlighted the problem, and then disagreed completely on the level of asthma control, commonly noting that in the absence of any substantive information related to the (perhaps incorrect) diagnosis, making a determination as to the patient's level of control was difficult at best. There were several non-clinician respondents who indicated that the absence of any mention of the current clinical expression of the diagnosis was the

equivalent of good control[xiv]. Similarly, several clinicians expressed concern that the patient had a diagnosis in the problem list that was not addressed, and for that reason, were unsure if the original ICD-9 code (e.g., Asthma: Unspecified) was actually correct. This single document demonstrated what may be a pervasive problem in clinical documentation: the presence of a problem or diagnosis statement without supporting information or comment. This example suggests a strong need for greater exploration of how diagnoses are verified, carried forward, and addressed in clinical documentation. This topic will be explored further in the discussion section.

In five documents, both lay persons and experts questioned whether or not the patient really had asthma, thus making it very hard to answer a question about the level of asthma control. This is somewhat different from the document discussed in the previous paragraph, as in these five documents, there is considerable information about the patient's pulmonary status, however, the provided information is not sufficient for clinicians to determine what is really going on. The following comments shed light on this finding: "*sounds more like a URI [Upper Respiratory Tract Infection]*", or "*I think this is COPD [Chronic Obstructive Pulmonary Disease]*." One of the two documents also contained the phrase "*asthma r/o severe pneumonia*" which was always selected by subjects. In this case, the reporting clinician (e.g., the clinician creating the original documentation) appears to be stating a differential diagnosis; that is, because the clinician lacks sufficient evidence to state a single diagnosis, s/he has narrowed the list to a set of candidate diagnoses (e.g., "asthma" or "severe pneumonia"). Some subjects questioned

[xiv] This represents an example of questionable heuristics. The subjects may be right, but there is a substantial probability that this is an incorrect assumption.

whether such documents belonged in the data set; others considered this a statement of the patient having asthma with superimposed pneumonia, still others identified the statement as a differential diagnosis, and, finally, others wondered if the patient had asthma at all. Again, such cases demonstrate a need for better clarification of the underlying status of a clinical condition, even if it is not precisely diagnosed, should a differential diagnosis be listed, subsequent clinicians can build a more detailed and appropriate picture of the patient's current status in the context of the patient's history. Even though the clinicians identified the differential diagnosis, one clinician subject commented "*it is difficult to figure out if this doc[tor] has any idea what is going on with the patient, and has ordered a chest film, without doing a complete pulmonary assessment. This is a horrible note.*" This statement demonstrates that clinicians can and do wrestle with making sense of clinical documentation after the fact, and what may have appeared clearly described by one clinician can read as highly imprecise if not utterly confusing to another.

Two other documents in this set demonstrate an interesting similarity. Both say quite a bit about the patient's asthma, but neither explicitly states the current status of the disease. In other words, despite a large quantity of descriptive information (e.g., what tends to exacerbate the patient's asthma, how the patient has gotten rid of her cats to see if this helps reduce episodic wheezing, how the patient's *child's* asthma is progressing, etc.) no set of descriptive data forms a complete picture of the patient's *current* status. For example, in one document, the writer states "*Reactive Airway Disease – has been using her inhaler more – will plan to check PFTS and renew inhaler. No signs of wheezing on exam today."* Clearly there is a great deal of detail about this patient's asthma care in this

quote, but one subject stated "*this means nothing to me*" when asked if the patient's asthma is well-controlled, adding that "*reactive airway disease can mean something other than asthma, though they may be treated similarly*."

Finally, the remaining 13 documents in this set represent what might best be called irresolvable ambiguity. Complex statements such as: "*Unfortunately, the patient has a hard time understanding the need for maintenance therapy to avoid exacerbations*" are difficult to assess, even for the expert clinician, who comments: "*This becomes a semantic question. During the time of the visit, the patient is nearly symptom free. Only a slight itch, clear lungs and near normal PFT. However, her failure to take her medicines place[s] her at risk for an attack. As asthma is an intermittent disease, the absence of symptoms at any moment in time is not evidence of being well controlled."*

This last group of documents elicited what might be called the most "frustrated" set of comments. Although such comments were few, both subject groups did comment on poor documentation style ("*This is a horrible note*" or "*all of this is just fluff, it tells me nothing*" or "*if this patient has asthma, there should be some mention of the level of use of rescue inhalers.*" The emergent impression from his review is that when readers become frustrated in their attempts to find desired information they may focus the "blame" for that lack of clarity on the previous, documenting clinician. Part of this blame may be misguided; this issue will be explored in greater depth in the discussion section.

Summary

There are similar overall patterns in the asthma set as in the smoking set: subjects tend to repeatedly select the same text snips and there is almost always overlap between subject pairs. The comments associated with the text help us understand when and often why points of disagreement emerge. For asthma documents, subjects disagree about what "well controlled asthma" is, and what "currently" or "now" means in precise temporal terms. This information is neither new nor surprising. What is interesting though, is that the qualitative analysis of the comments reveals that both non-clinicians and clinicians disagree in the same ways and for the same reasons, and that in only a very few cases – two asthma documents out of the combined set of 58 documents –does clinical or domain-specific knowledge appear necessary as input to forward the decision-making process. But, because in these two cases, the clinicians split evenly on their answers, it is hard to determine if this constitutes evidence of the need for clinical knowledge to answer the questions.

5.4.4. Qualitative Analysis Summary

Although the three distinct answer patterns noted in the data allowed for a preliminary grouping of the comment data for qualitative analysis, five themes emerged when the data were then re-analyzed as a whole. These very general themes include 1) explicit text, 2) the use of ad hoc heuristics, 3) issues related to time boundaries, 4) inability to precisely define terms, and 5) irresolvable ambiguity. Each of these themes is discussed in detail in next chapter. However, in general, there appears to be a clear set of distinguishing traits that identify an explicit statement, at least among the documents used

here, and, when these statements appear in documents, agreement is usually very high. The qualitative analysis also suggests that clinicians and non-clinicians don't disagree as "rigidly" as pure quantitative statistics might show. Indeed, the additional qualitative evidence presented here suggests that in general both types of subjects would improve their ability to decide on a "yes" or "no" answer, if terms were, if not more explicitly stated, better defined from the onset. It appears that all subjects struggle with imprecision in definition of time boundaries as well. Finally, for this study, domain expertise rarely appears necessary to resolve the disagreements observed among subjects or between groups. Again, this can be stated only with regard to these specific data sets. Regardless, this is something of a surprise.

When returning to the hypotheses motivating this study, despite the lack of expected quantitative support for the hypotheses, the qualitative evidence offers tremendous insight into the reasons for disagreement among subjects. This evidence therefore drives the discussion that follows.

6. Discussion

6.1. Introduction

This mixed-method, hypothesis-driven study explored the differences in levels of agreement between non-clinicians and clinicians when answering questions about and annotating text in ambulatory care encounter notes. The study also explored the differences in levels of agreement within and between subject groups for these same measures for "simple" and "complex tasks". The goal of the study was to gain insight into the reasons coders disagree when identifying concepts in clinical documents.

The discussion begins with an examination of the quantitative results and highlights several study design issues that may have confounded the results. This is followed by an examination of the qualitative themes that emerged from subject comments. Finally, the results are synthesized as a framework for conclusions and recommendations for leveraging computerized clinical systems (such as EHRs) to help mitigate or eliminate some of the causes of disagreement among document coders.

6.2. Results

6.2.1. Hypothesis #1

Introduction

When examining question answers, this study provides generally encouraging evidence in support of Hypothesis #1 (e.g., clinicians will agree with one another more than non-clinicians), With regard to snip selection, the results are somewhat less clear, as some of

the statistically significant results counter the proposed hypothesis. These results are discussed separately.

Question Answers

A simple summary table of the results of the quantitative analysis of question answers for Hypothesis #1 is useful as a reference for this discussion:

Question	% Observed Agreement	Kappa
Smoking Q1	N/S	N/S
Smoking Q2	C > NC	C > NC
Asthma Q1	C > NC	C > NC
Asthma Q2	N/S	N/S

Table 23: A summary of answer data for Hypothesis #1

For this table, 'C' represents "Clinician" and "NC" represents "Non-clinician". Cells highlighted in green represent significant results in support of the hypothesis; with clinicians agreeing significantly more than non-clinicians (e.g., C > NC). The remaining cells, containing "N/S" represent non-significant results in measures of agreement within groups.

Reviewing this summary table reveals that clinicians equal or significantly exceed measured agreement in answers when compared to non-clinicians, lending substance to the hypothesis. In the case of Asthma Question #2, clinicians demonstrate higher agreement overall than do non-clinicians though the results are not significant. Both study groups perform (statistically) equally in terms of agreement for Smoking Question #1.

The lack of *significant* differences between study groups for two of the four questions requires investigation. First, it is noteworthy that the subject groups do not differ on a simple question (e.g., Smoking question #1, which asks if the text "clearly states" whether or not the patient smokes) as well as a complex question (e.g., Asthma Question #2, which asks the subject's opinion). Had the non-significant results occurred with simple (e.g., "clearly state") question types *only*, it would have been easy to conclude that the document sets contained sufficient, explicitly stated data to enable both non-clinicians and clinicians to agree on the answer: that is, the exact phrases needed were present, and the subjects found them. However, this is not the correct explanation; the subject groups showed no significant differences in agreement for a complex (e.g., "in your opinion") question as well. This suggests that the results may be potentially confounded by the structure of the questions themselves.

Do the Questions Discriminate Clinical Expertise?

First and foremost, it is important to consider whether or not the questions themselves sufficiently discriminate clinical expertise. Both hypotheses for this study posit clinical experts will out-perform non-clinicians, despite different task objectives. However, the topics chosen may represent topics for which the non-clinician study group possessed sufficient *general knowledge* to answer uncomplicated clinical questions. For example, if we consider smoking documents, the language used to describe smoking habits comprises a relatively constrained set of phrase patterns when compared to more complex clinical concepts, such as "pulmonary fibrosis", or "right-sided heart failure and its pulmonary complications", etc. Furthermore, the documents used in this study were presented in a

consistent format, as they were extracted from the hospital clinical information system, and all subjects reviewed the documents in the same order. It is therefore reasonable to assume that a) subjects rapidly determined where smoking information was likely located in documents, b) what language was commonly used to describe smoking behavior in patients, and c) what pieces of explicit information, when woven together, represented sufficiently "clearly stated" information to answer a question. These three assumptions are extremely reasonable in light of the fact that the study comprised a very small set of documents pulled from two clinics over a short interval of time and most importantly *many documents were created by a small set of clinicians,* meaning that the document set might lack sufficient internal variability to reveal the intended distinctions.

Smoking questions represent rather "general knowledge" questions, so one suspects that differences in agreement on smoking questions might be insignificant, as the descriptive language about smoking tends to be non-medical in nature.[xv] The questions asked about asthma contain a similar, though less-obvious general knowledge component. Indeed, non-clinician subjects in the study demonstrate by their answers choices and associated text selections that they possess a solid baseline understanding of asthma, e.g.: it can interfere with activities of daily living, often requires inhalers, steroids, and possibly other medications, manifests as exacerbations or flares, may result in visits to the Emergency Department, etc., and that if someone's asthma is well controlled, the patient does not usually exhibit these behaviors.

[xv] With the obvious exception of common abbreviations like "ppd" ("packs per day"), or expressions such as "pack year history" which may not be known to non-clinicians.

However, asking if the patient had "well controlled asthma" was intended to distinguish between clinical and non-clinician knowledge, based on the assumption that well controlled asthma is multifaceted concept requiring in-depth understanding about respiratory management status based on a number of clinically relevant variables. The desired distinction between non-clinician and clinician groups did not emerge, at least for the asthma question. Here it is possible that the documents themselves did not vary enough in clinical information regarding the level of control (that is, too many documents may have described generally well-controlled asthma cases, and few documents presented complex or refractory asthma management). As a result, both subject groups may have found it somewhat possible to intuit the level of control, without additional or specific clinical knowledge.

We must also consider the educational level of the lay subjects recruited for this study. The subjects were college-educated and all worked at this researcher's hospital in various capacities. Despite the fact that lay subjects were non-clinicians, the fact that the group included medical librarians and clinical informatics students clearly may have biased the results. That is, the questions may possess sufficient discriminatory power if the lay subject group is not so well educated or obviously associated with a clinical environment. This suggests that repeating the study, with more stringent lay group inclusion criteria (e.g., cannot be college-educated, may not work in a clinical setting, etc.) might reveal more of a distinction in clinical knowledge requirements between groups.

How Does the Operational Specificity of Terms Affect Answers?

Another issue to consider when evaluating these results is the operational specificity of the terms used in the questions. As previously discussed, question terms were not defined for subjects. So, when the question, "Does the text *clearly state* that the patient smokes?" was asked, the lack of an explicit definition of "*clearly state"* may have created cognitive distress for some subjects. The phrase "clearly state" was intended to require the subject to find the exact words that expressed the desired concept. For example, a statement such as "patient denies smoking" might represent an explicit statement of the patient's current smoking status. However, many subjects took the "clearly state" somewhat *too literally*. When creating document-identifying icons for the document types, a smoking cigarette was used to represent smoking documents. Interestingly, nowhere in the instructions or questions were subjects asked specifically if the patient smoked *cigarettes*. The question simply asked if the document clearly stated whether or not the patient *smoked*. From the comments associated with the answers, it is clear subjects assumed the question related to cigarette smoking, though a few subjects mentioned patients could smoke other substances, and wondered if chewing tobacco should be considered when answering this question. So, the study may have unintentionally planted a cognitive seed (the smoking cigarette icon) which primed all subjects to look for explicit statements related to cigarette smoking.

There may be less of an issue with operational specificity regarding the phrase "*in your opinion*", as in "*In your opinion*, is this patient's asthma well controlled?" though many

of the clinicians did ask if the study required a medical or personal opinion.[xvi] "*In your opinion*" appeared to offer subjects a great deal more cognitive latitude in interpreting the data, as evidenced by the comments, which demonstrated considerable reasoning about the presented evidence in aid of answering the question. There may have been less anxiety about the correctness of an answer when asked an opinion question, perhaps affording the subject greater freedom or comfort in stating an opinion, though it is difficult to make such a determination from this study. Finally, the study asked the "*in your opinion*" question first, followed by the "*clearly states*" question. This approach was intended to allow subjects to form a gestalt of the document in the context of the question, first, followed by a more precise identification of text afterwards. It is difficult to determine if this ordering confounded the operational specificity of terms. In addition, it is difficult to determine how subjects personally interpreted "*in your opinion*" when answering the question. This consideration is discussed in more detail in the qualitative analysis that follows.

The phrase "well controlled" was also very operationally unclear; and, again, this was intentional. As previously discussed, when asked what this meant, subjects were presented a metaphor: "If you had chronic pain, and it was well controlled, what would that mean?" It was the author's intention to avoid prescribing a definition that might alter an existing mental model representing asthma control for any subject, and thus introduce a strong confounder to the hypotheses. The notion of well-controlled asthma, and the factors that contribute to this label, is relatively new and, outside of the score from the

[xvi] I did not select for the clinician. I asked them to make a decision one way or the other and be consistent in its application.

Asthma Control Test, perhaps not well defined. Furthermore, the penetration or dissemination of this knowledge to the broader clinical community has not been assessed but is likely limited owing to the length of time it takes to move clinical evidence into practice. If clinicians do not have clear mental models of this complex concept (by current definitions) it is certainly reasonable that non-clinicians lack this internal representation as well. The intention to avoid defining well controlled asthma was well thought out; whether or not this lack of specificity confounded the results is unclear, without a repeat study, where all terms are formally operationalized, to compare results.

The Text Snips

As with the answer results above, a simple summary table of the results of the quantitative analysis of snip agreement for hypothesis #1 is useful as a reference for this discussion:

	% Observed Agreement	Kappa
Smoking Q 1	C > NC	N/S
Smoking Q2	NC > C	NC > C
Asthma Q1	C > NC	N/S
Asthma Q2	N/S	NC > C

Table 24: A summary of agreement among snip data for Hypothesis #1

For this table, 'C' represents "Clinician" and "NC" represents "Non-clinician". Cells highlighted in green represent significant results that support the proposed hypothesis; cells in pink represent significant results that counter the proposed hypothesis. Cells highlighted in green support the hypothesis. The remaining cells, containing "N/S" represent non-significant results in measures of agreement within groups.

Visual review of this table demonstrates some confusion among the results with regard to the hypothesis that clinicians will agree more than non-clinicians in terms of text selection. When percent observed agreement is used as the measure of agreement, clinicians agree significantly more on the *simple* questions for both document sets. Non-clinicians agree at least as well or significantly more than clinicians for the *complex* questions. This combination of mixed results counters the proposed hypothesis Under Kappa, the performance of the clinician group is worse; here, non-clinicians agree at least as much or significantly more than clinicians. Thus, the hypothesis is not supported. As with the question answers, the inconsistent results require investigation.

Does Clinical Experience Matter?

We have already discussed that the study questions may not be sufficient to discriminate between clinical and non-clinical knowledge. Clearly, clinicians have greater experience diagnosing and treating asthma than do non-clinicians by virtue of years of training and experience, and must therefore possess some mental model of asthma control, whether or not it is precisely framed or bounded. Furthermore, clinicians must include more *clinical* concepts in this mental model than non-clinicians. Indeed, this is well demonstrated by the very high frequency with which clinician subjects select specific, *medically-related* asthma topics (e.g., results of pulmonary function studies, oxygen saturation on activity, use of rescue medication, x-rays results, etc.) as evidence to support an answer. In terms of the proposed hypothesis, therefore, we expect clinicians to out- perform non-clinicians. But this is not the case.

Again, the comment data reveal an interesting phenomenon regarding text selection, particularly within the asthma document set: many non-clinicians selected text because they thought it *might* be important. For example, in virtually all instances where the results of pulmonary function tests (PFTs) were listed in a document, non-clinicians selected the text then indicated in the associated comments that *they were not sure what the results meant, but that the data were likely important*. This presented an unexpected confounder for differentiating between clinical and non-clinical cognition, as one would expect PFT results to represent knowledge requiring greater clinical training or expertise. Interestingly clinicians revealed through their comments that they carefully weighed the PFT results to make a determination of the level of asthma control. Thus, some text was selected by non-clinicians because it *might* be important whereas clinicians selected the same text because it was *considered* in the decision of how to answer a question This presents a conundrum for interpretation because then *motivation* for text selection differs among individual subjects, independent of the instructions provided. Thus, the measures of agreement selected contain a strong amount of potential noise which may account for the results. It is intriguing to consider if asking the subject to rate his or her level of certainty on the questions might have provided more insight into this distinction. Perhaps this additional data might account for the results seen here.[xvii]

Early in the results section, in Figure 19 (Comparison of total selected characters between study groups), it was noted that non-clinicians select as least as much or significantly

[xvii] It is intriguing to consider if asking the subject to rate his or her level of certainty on the questions might have provided more insight into this distinction. Perhaps this additional data might account for the results seen here.

more text overall than do clinicians on 2 of the 4 study questions. This suggests that non-clinicians may arbitrarily select any text involving smoking (or, asthma) concepts or statements, whether or not it actually contributed to the question decision or not, thus resulting in a higher overall amount of text selected with the associated increased likelihood that much of this text would overlap with other subjects. However, this result, too, was inconsistent. When the subjects were compared by mean total selected characters and mean percent overlap characters, non-clinicians selected significantly more characters for asthma question #1 than non-clinicians, but overlapped significantly more on smoking question #2 and asthma question #2; that is, a selection of a high number of characters overall did *not* appear to correlate with an increased level of agreement as measured by mean total character overlap. So, the question remains open as to whether text overlap represents true agreement between subjects, despite the fact that it may be our current best measure of this complex cognitive concept.

6.2.2. Hypothesis #2

Introduction

This study fails to support Hypothesis #2. As with hypothesis #1, question answers and snip selections are discussed in separate sections.

The Questions

As with Hypothesis #1, a summary table helps compile the data into a manageable unit for review. The following table summarizes Hypothesis #2 question data

Subject Type	Comparison	% Observed Agreement	Kappa
Non-clinician	S1-S2	S2 > S1	N/S
	A1-A2	N/S	A1 > A2
Clinician	S1-S2	S2 > S1	S2 > S1
	A1-A2	A1 > A2	A1 > A2

Table 25: A summary of agreement among subject question data for Hypothesis #2

For this table, 'S' represents "Smoking" and "A" represents "Asthma. S1 and A1 are the simple questions, as S2 and A2 represent the complex questions. Cells highlighted in pale green represent statistically significant results consistent with the proposed hypothesis. Cells highlighted in pink represent significant results that *contradict* the proposed hypothesis. Cells with a white background and the value "N/S" are non-significant results.

In terms of mean percent observed agreement, at the answer level, both non-clinicians and clinicians agreed significantly more on the *complex* smoking question, contradicting the hypothesis. However, for all remaining questions subjects in each study group agreed at least as much or significantly more on the simpler question. Both measures of agreement (percent observed agreement and Kappa) correlated well. Because these results fail to support the proposed hypothesis entirely, the results require deeper investigation. Not surprisingly, many of the previously identified study issues emerge as potentially confounding concerns.

Defining Simple versus Complex Tasks

The definitions of "simple" and "complex" questions used for this study were based on I2B2's use of the terms "textually-based" and "judgment-based" as previously discussed. In light of the results for this hypothesis, it is reasonable to question whether or not this distinction in concept identification tasks is appropriate to use as a means of differentiating between simple and complex *cognitive* tasks. In adopting these distinctions, we have assumed that explicitly stated concepts are cognitively easier to resolve than concepts requiring additional inference. This may not be the case; inference is likely required to make sense of even the most explicitly stated text, as details must be weighed in context, ambiguity must be resolved (as in the case of explicit but contradicting statements within a single document), and the user must derive meaning from this combination of cognitive activities. So, explicitly stated concepts may not be cognitively simpler than less explicitly stated ones and the presence of easily identifiable, explicit text may serve only to lessen the obvious differences between clinicians and non-clinicians when the concepts themselves are relatively simple, as in this study. That is, if we ask if a patient smokes, it is not especially difficult for a subject in either group to answer "yes, the patient smokes" when phrases such as "*patient continues to smoke and is unwilling to consider stopping at this point*" or "*patient smokes about a pack a day*" appear in the text.

So, there are two conclusions to draw regarding these results and the nature of simple vs. complex cognitive tasks. The first is that we cannot determine from this study if an explicitly stated concept is cognitively easier to identify than another concept requiring inference. Secondly, our ability to draw this conclusion is likely confounded by the relatively simple concepts we asked subjects to identify in the text: that is, this study did

not include documents of sufficient clinical complexity to determine if the distinction was useful. We did not, for example, give the subjects questions requiring obvious clinical training, such as stating a diagnosis based on a group of complex symptoms, or some other, clearly more difficult clinical task. So, both the usefulness of the I2B2 task definitions as discriminators of cognitive work remains an open question.

Term operationalization

We have already alluded to the problems that may emerge when terms are not well operationalized and that the intentional decision to refrain from defining specific terms might have introduced strong confounding effects. This portion of the discussion is not repeated under this hypothesis. However, it is important to address an additional effect of lack of term operationalization that re-evaluation of Hypothesis #2 textual data and comments revealed.

Questions asking the subject if a concept was "clearly stated" represented the textual or simple question. The lack of term operational specificity may have unintentionally obfuscated what *might* have been a simple question by asking the question *vaguely*. On the other hand, this specific problem may not be resolvable. If "clearly state" was fully defined – leaving no room for doubt as to its meaning – subjects would need lists of "acceptable" combinations of words in every position they may occur in every document. At this point, humans would become unnecessary to code documents because the meaning of the words would lose relevance; the task of identifying the explicit task would be one of direct pattern matching – something computers do more efficiently than humans. Furthermore, should this exhaustive list not be available, subjects would require

heuristics for compiling discrete data points into meaningful concepts. Automated methods have not yet completely mastered these tasks, though work is ongoing. So, what definition of "clearly state" suffices? Humans code text by translating symbols on a page into bits of information and synthesizing this with personal experience and education to form meaning. From this meaning, we form mental models, and our models are universally unique. Thus, there does not appear to be a sufficient definition of "clearly states" that suffices if we are to examine cognition, and we subsequently have to make do with this limitation.

The Text Snips

The following summary table displays the results of the quantitative analysis of the selected snip data for Hypothesis #2:

Subject Type	Comparison	% Observed Agreement	Kappa
Non-clinician	S1-S2	S2 > S1	S2 > S1
	A1-A2	N/S	A1 > A2
Clinician	S1-S2	S2 > S1	S2 > S1
	A1-A2	A1 > A2	A1 > A2

Table 26: A summary of agreement among subjects for snip data for Hypothesis #2

For this table, 'S' represents "Smoking" and "A" represents "Asthma. S1 and A1 are the simple Cells highlighted in pale green represent statistically significant results consistent with the proposed hypothesis. Cells highlighted in pink represent significant results that *contradict* the proposed hypothesis. Cells with a white background and the value "N/S" are non-significant results..

When reviewing these results, it is noteworthy that both study groups demonstrate consistent behavior across questions and measures of agreement, despite the fact that these results do not support the proposed hypothesis. Both non-clinicians and clinicians demonstrate higher agreement on the more complex smoking question. Similarly, both groups demonstrate at least equal or significantly higher agreement on the simpler of the asthma questions. Again, these mixed results (in terms of the proposed hypothesis) warrant investigation.

We have already discussed many potentially confounding effects in this study: 1) that the questions asked of subjects may not have sufficiently discriminated clinical knowledge, 2) that intentionally limited term operationalization may have hidden potential differences among subjects, 3) that the document sets selected for the study may not have been clinically complex enough to demand clinical expertise to judge, and 4) that the I2B2 definitions of textual and inferential concept identification tasks may not be suitable as the distinction between simple and complex cognitive tasks, respectively. Each of these issues may have effects here, and for the same reasons as previously discussed, so these issues are not repeated.

However one final point warrants notice. For the smoking documents, both subject groups agreed significantly more on the more complex of the two questions. A likely explanation for this emerges upon review of the qualitative data associated with these questions, and is discussed in greater detail below. However, briefly, this result may reflect a combination of a basic ability to identify smoking concepts (e.g., what it means for someone to smoke, and how that is often represented in text) in combination with

additional knowledge about smoking, specifically (e.g., that it is difficult to quit this most addictive habit, that relapses are common, that many factors play into relapse, etc.). This does not represent clinical knowledge *per se*. In addition, when subjects are asked their opinion about whether or not a patient smokes, the subjects are no longer limited to locating explicit statements within the text, and can leverage their general knowledge along with any selected text related to smoking to form a reasonable opinion. If the gold standard for agreement is a majority vote of experts, then non-clinicians demonstrate performance consistent with the gold standard for this question, which suggests that a) explicit statements about smoking in the study documents leave a great deal of room for interpretation, and b) smoking status is neither fully-assessed nor consistently documented, at least in the study document sets, and c) somehow, all subjects are able to form consistent gestalt about smoking status despite these concerns. We explore this further in the qualitative analysis that follows.

6.2.3. Qualitative Discussion

When qualitative analysis is performed along with quantitative analysis, the former can often shed light on the latter. In the case of this study, the qualitative analysis produced invaluable insight into the nature of disagreement when coding medical documents. First, three patterns of answers emerged among the study subjects, regardless of study group, and these patterns included perfect agreement, fair agreement, or disagreement. Using these preliminary groupings to organize the comments data then reviewing the data in the context of the associated answer and snip(s), five consistent themes related to inter-rater agreement emerged: 1) specific and literal (e.g., explicit) statements lead to very

high agreement, 2) when explicit statements are not present, the application of *ad hoc* heuristics can create high agreement, 3) subjects have trouble clarifying time boundaries that are clinically relevant and can lead to disagreement, 4) exceptions to rules cause confusion, and 5) some ambiguity is inherent in the source text and is irresolvable . These themes are not surprising, nor new, but do highlight the fact that inter-rater agreement, though somewhat easily measured in quantifiable terms, may not be best explained "by the numbers" alone. In addition, these thematic findings suggest that there is considerable room for improvement in reducing ambiguity in the clinical record. The following discussion is again organized by the three question pattern classifications used for assessing snips, because this provides a consistent presentation framework. The themes are discussed as they cross-cut this organizational framework.

Conditions for Perfect Agreement

For a certain set of documents, all subject pair-wise comparisons, regardless of group, select the same answers to both questions and also select overlapping snips. That is, all users answer "yes" and select "*the patient smokes 0.5 pks/day*". When all users select the same or similar text snips and answer both questions in a document the same way, perfect agreement has occurred, in the context of this discussion. This pattern of perfect agreement was startlingly high for smoking questions: 47% for non-clinicians 40% for clinicians. This indicates that a large portion of these documents contain text detailed and specific enough to definitively determine if a patient does or doesn't smoke.

Subjects uniformly identify explicit statements of status in the text when they are able, and identify these snips as "*definitive*" or "irrefutable *evidence*" to answer a question. All subjects locate phrases such as "*Tobacco Use: Never*", "*smokes ½ pack per day*", "*not ready to quit smoking*", "*asthma well controlled*", "*asthma completely stable*", etc. Even when subjects make comments indicating that some snips are not entirely explicit, subjects often provide additional evidence (such as medications) in snips, as well as more detailed comments on the reasoning underlying an answer. For example, the statement "*Tobacco Use: Quit 1995*" elicited many comments. Subjects commented that there was room for doubt in such a statement: "*I mean, she could be smoking again, but it doesn't seem that they ever ask that, so I assume if the patient was smoking it would be noted.*" Despite the statement for potential doubt, the subject decided the patient did not smoke. This particular comment came from a non-clinician and there is no example in the study of a comment of this style made by a clinician. This raises the question if it is commonly *understood* by clinical personnel that "*Tobacco Use: Quit 1995*" implicitly indicates the patient does not smoke *now*. This is a risky assumption. Without addition information to clarify the currency of smoking status, there is significant room for doubt for any coder.

Given the lack of explicit statements regarding a patient's asthma control in all but a few documents, it is necessary to look for other factors explaining agreement. In the few documents where clear statements such as "*stable controlled asthma*" or "*asthma, mild, intermittent, well controlled: yes*", do occur, agreement is complete, as might be expected. However, even when explicit statements are absent, the presence of an Asthma

Control Test[xviii] score also accounts for a large portion of the perfect agreement among both subject groups. Asthma Control Tests or ACTs (http://www.asthmacontrol.com) may not be well known to clinicians, given the difficulty in disseminating new clinical information in today's climate of data-overload. In this study, in those cases where the ACT was used to assess a patient's level of asthma control, there was complete agreement between study groups, with a singular exception: In this case, a non-clinician noted that the ACT score was 11, but had no idea if the score was, in their words: "*good or bad, since I don't know what the total score can be, or if high is good or low is good. Seems like control should be a low score, but I am not sure.*" In this case, it is disappointing, at best, that the score was reported without a denominator, or some modifier indicating the *meaning* of the score of "11".

As discussed in the results section, when subjects cannot find explicit text within a document they often devise *ad hoc*, document-specific sets of heuristics to reach a conclusion regarding an answer. For smoking documents, some simple algorithms appear to meet a commonly and tacitly agreed-upon set of rules that equate with "*patient currently smokes*". Both non-clinicians and clinicians demonstrate that certain combinations of facts constitute sufficient evidence of smoking, and the groups agree on these facts an equal amount of the time.

Examples of asthma heuristics are less frequently observed (recall comments were not required and are often omitted), though several did emerge, such as "*use of an inhaler*

[xviii] The Asthma Control Test (or, ACT) and the Childhood Asthma Control Text (both available at http://www.asthmacontrol.com) were first included in the National Institute of Health's Asthma Guidelines in 2007. The test consists of 5 questions, each asking the respondent to indicate how often in the last 4 weeks asthma has impacted their lives and in what ways. The user answers with Likert-scale type responses covering a 5-point scale. An ACT score of ≥ 19 is consistent with poor asthma control.

means the asthma can't be well controlled." There is insufficient data to explore this theme in depth as, although similar comments from both study groups occur, they occur infrequently. However, "use of an inhaler" must represent some baseline knowledge about asthma for this study's subjects and must carry associated implications of level of control. In addition, in one case of perfect agreement, all subjects picked some portions of this piece of contiguous text whether as separate snips or as a complete chunk:

> *Asthma - getting better. Has finally stopped sleeping in the recliner. Gets SOB occasionally. Is taking albuterol every 2-4 hrs still. Peak flows 230-270*
> *Has had two occasions where she woke at night and felt that she couldn't move air in or out….Prednisone is down to 1/2 tab*
> *Has taken albuterol for SOB primarily.*
> *Doesn't notice much wheezing. Feels as if she is not getting good deep breaths. Is definitely subjectively better than she was.*

It certainly would be intriguing to understand the heuristic at work in this example, where a statement of asthma control is clearly missing. However, the two statements about difficulty breathing ("…*where she woke at night and felt that she couldn't move air in or out*…" and "…*she is not getting good deep breaths.*…") were selected in *all* cases, so level of asthma control may directly relate to ease of breathing for the study subjects, even in the presence of other potentially informative surrounding data.

Conditions for Fair Agreement

In this category there are two sets of data for each document: One set with all "yes" and "not sure", and another with all "no" and "not sure" as answers. For this group, no other combination is possible. If the comments are combined, it rapidly becomes clear that the points of disagreement are best described by the comments provided when subjects answer "not sure." In these cases, the distinction between those who are sure and those who are not is not between "yes" or "no" but between "yes" and "yes *enough*", or "no" and "no *enough*". Thus, subjects are not on opposite ends of the spectrum, just somewhere in the boundary area between "yes, absolutely" and "yes, probably."[xix] This suggests that the level of agreement between clinicians and non-clinicians may be rather large after all, if some of the "not sure" can be resolved, or, more specifically, if the fuzziness in concept boundaries can be tightened. Indeed, the comments grouped in this category demonstrate not only why subjects disagree, but also that subjects differ on concepts in very similar ways.

For example, questions about time emerge as an important reason for answering "not sure". Subjects find it particularly difficult to properly bind the notion of time into an appropriate interval or window that fits the question. "*What does 'now' mean?*" asks one subject, '…the *doctor said things had been good 'recently'. How long is recently?*" In addition, given that the question concepts (e.g., "*smokes*" or "*has well controlled asthma*") are both imprecise and context-dependent, the confusion about the precise meaning of "currently" or "recently" or "now" becomes reasonable. And, because

[xix] It must be emphasized that this conclusion emerges from the comments the subjects entered, and not as an assumption of the author.

"current status" is so important to our assessment *of another's assessment of a patient* in a study such as this, subjects demonstrate disagreement when they are unclear about where and how big the difference or boundary between "yes" and "no" really is. Time certainly helps set this boundary, but significant differences surround what level of precision is necessary, or if the precision has true relevance. This following example, previously mentioned, demonstrates the problem. Many subjects had trouble determining if this patient was a smoker or not:

> …s*moked 4 cigarettes several weeks ago when she was caring for her sick mother. No cigarettes for the last 2 weeks but having cravings. Wondering about options to decrease cravings but worried about cost*.

There is no question (based on the comments) that this patient has smoked before. In fact, other portions of the text support the fact that the patient has had a difficult history attempting to quit the habit. Subjects find many interesting things to say about this snip, but most commonly raise the issue of whether smoking 4 cigarettes several weeks ago constitutes smoking "*now*". Several subjects are hard-line: "*…if she has smoked before and smokes under stress, she is likely to smoke again. I say she is a smoker*." Other subjects are somewhat more forgiving: "*Sounds good enough for me*." Regardless, the absence of some well-defined concept of time, in the context of tobacco use, appears to be unclear to subjects. Furthermore, the confusion may result from a combination of textual ambiguity (e.g., it is unclear if the patient smokes now) or the lack of specificity the terms used in the study question, as previously discussed. It is not possible, in this study, to determine which or both of these issues most influence the subjects' answers.

However, one suspects that greater clarity in documentation might alleviate some of the ambiguity introduced by asking potentially confusion study questions.

Conditions for Poor Agreement

Please recall that documents in this last group demonstrated patterns of answers showing complete disagreement, that is, some subjects responded “yes”, some answered “no”, and the remainder “not sure.” It appears that the level of inter-rater disagreement placing documents in this final category has to do with a variety of factors. In the absence of explicit statements or workable heuristics to answer questions in the moment, agreement in answers erodes every time a new type of imprecision is added. For both smoking and asthma documents, term imprecision in combination with unclear time intervals seems a sufficient combination to cause some confusion, resulting in fair rather than perfect agreement. However, for this final document cluster, a different kind of problem emerges: the patient presents an exception to a rule: an exacerbation to an otherwise stable diagnosis (e.g., a former smoker who had a cigarette two weeks ago or an asthmatic patient who experiences a flare when exposed to her neighbor’s cat). Perhaps this latter category is a bird’s eye view of the previous problem: an issue of how to interpret the single exception (which represents an incomplete or unclear concept definition) in the context of time. Regardless, these exceptions create disagreement among non-clinicians and clinicians alike, suggesting that we do not yet have a good understanding about how single clinical exacerbations impact overall disease state, at least for whether or not to classify a person as to whether or not they smoke, or their level of asthma control.

Finally, some ambiguity is irresolvable *post hoc*. Medicine is often a science of exceptions to rules: clinical documentation must sometimes reveal more about what a condition isn't than what it is until a final diagnosis can be made. If clinicians do not make clear statements, or contradict themselves when documenting, it is particularly difficult to deduce intended meaning at a later date. For one document, one clinician subject commented "*this really is a horrible note…"* The document in question caused confusion among all subjects. Only the person who created the note can determine its meaning, and over time, with many intervening patient encounters, that meaning can be quickly forgotten. Thus, every effort must be made to support the clinician in entering clear, consistent, and understandable data. In addition, we must improve our ability to effectively render that data back to others so that its appropriate and intended meaning may be derived. We discuss this especially important point, and recommendations for how to mitigate some of these issues later in this document.

6.2.4. Implications for Inter-rater Agreement

Introduction

When reviewing the quantitative and qualitative results together, this study suggests that agreement is a matter of degree, and in that context, non-clinicians and clinicians appear to agree with each other a large part of the time. When the study groups disagree, they do so for very similar reasons. Most surprising is that the *issues causing disagreement have little to do with clinical expertise*, but instead have to do with how people resolve perceived ambiguity. Simply put, in this study, *agreement is higher when ambiguity is low.*

As EHRs increasingly become the norm for data collection in clinical care, we must continue to tune these systems to assure that the *data that is collected is as correct and unambiguous as possible* and furthermore, that the collected *data is rendered back* to the user both correctly and unambiguously. Many examples from the qualitative analysis demonstrate that systems can collect and propagate poor data collection and subsequently re-render it unintelligibly. Sometimes the problem is with the data collection, sometimes it is with the rendering, and sometimes it is with both. What this suggests is that we may miss opportunities to leverage computational systems to mitigate some of these problems. Clearly, the documenting clinician plays a huge role in the quality of the data collected and disseminated to subsequent users. This important concern is not dismissed, but is set aside as we focus on computational levers for improving the quality, usefulness, comprehensiveness, and clarity of collected and rendered clinical data. The qualitative themes identified in this study provide a useful framework for discussing these recommendations and therefore guide the remainder of the discussion.

Agreement is high when concepts are explicitly stated

If the goal is to collect unambiguous data, then efforts directed toward mitigating the introduction of ambiguity at the point of data collection may help resolve some of the issues related in inter-rater disagreement. Based on this study, several recommendations for data collection can be suggested.

Computerized data collection currently accommodates a wide range of data collection from highly structured (e.g., menu-driven) to free text entry. Despite a built-in capacity for a high level of discrete data collection, the system used to collect documentation for this study set is newly operational and thus does not yet leverage many of these features. Indeed, the vast majority of clinical data, outside of discrete reports (e.g., labs, x-ray results, ECG results, etc.) remains textually based. Thus, it remains difficult to determine just where the balance needs to be set between requiring specific, discrete data for explicitness and permitting textual entry so clinicians may express nuance that cannot be captured through a rigidly prescribed set of menus. As discussed, agreement among subjects was generally high for this study, suggesting that free-text collection and rendering of data is not inherently bad; that is, subjects can make sense of and "correctly" interpret these data rather consistently.

With regard to the smoking documents, we must consider what we wish to accomplish with data collection about a patient's current smoking status. If smoking cessation is a major quality goal for health care, then it is reasonable to expect that, at a minimum, the necessary data related to a patient's smoking status must be collected, including whether or not the patient smokes, how much and what the patient smokes, whether or not the patient wishes to quit, if the patient is offered counseling or other support to quit, and if follow-up visits are scheduled to track the progress of these efforts. However, the data necessary to meet *quality goals* may differ subtly from data considered *clinically relevant at this visit*, leaving the clinician with the problem of making the determination where to focus limited time and energy in documenting the care given. For example, if a patient is seen for a broken ankle, the need for the evaluation of smoking status may be superfluous

in the given context. In addition, as discussed, we do not have a firm grasp on how to formally define the case of "smoker" vs. "non-smoker" in light of the interesting exceptions highlighted in the qualitative discussion (e.g., a patient who quit smoking two months ago but had "a couple of drags" last week).

For this study, a computerized data collection interface requiring the clinician to answer the question "Does this patient smoke?" would certainly help raters agree on the answer.[xx] But even this single question lacks specificity, because it does not provide any information about what or how much the patient smokes and whether or not the issue was addressed during the visit. Indeed, this study demonstrated that the concept of "smoking" was open to wide interpretation, depending on how subjects derived the answer from the documentation. Some subjects assumed the question referred to smoking cigarettes, though this was never clearly stated. Other subjects wondered if the question about smoking included non-tobacco products such as marijuana or methamphetamines. Still others asked in comments if this might include chewing tobacco as well. This suggests that including a roll-over pop-up window providing inclusion criteria for what constitutes smoking might provide clarification on the data entry side so that all clinicians would have the same reference point from which to make a diagnosis. Similarly, this pop-up would help raters understand the question asked of them, and might therefore resolve some confusion. As always, providing these tools is one thing; having users use them, and use them consistently, is another. And, deciding on whose definition to use, how to keep these definitions current with clinical evidence, and how to

[xx] Importantly, the list of acceptable answers, besides "yes", "no", "unable to determine" should include "unable to assess due to more urgent concerns". The clinician must be allowed to triage the patient and address clinical concerns in the order of their importance.

embed them in clinical information systems in real time, without significant computational overhead, remains an even more complex issue.

This single study cannot resolve these concerns. Clearly, requiring detailed, directed data collection about smoking status is not feasible at every ambulatory visit. However, this study suggests that if a patient has documented pulmonary complications (such as asthma) or has a history of tobacco use (in whatever form), the clinician must be required to re-assess smoking status completely, if other, more clinically urgent issues do not take precedence at this visit. And, if this documentation is required, the collected data must be a) unambiguous, b) complete, c) clinically relevant, and d) sufficient to measure quality of care for these patient populations. For smoking, this requires that the "Five A's of Smoking" be assessed according to the United States Public Health Service guidelines for Treating Tobacco Use and Dependencies (available at http://www.surgeongeneral.gov/tobacco/tobaqrg.htm), including:

1. Asking the patient if they use tobacco
2. Advising anyone who smokes to quit
3. Assessing whether or not the smoker is ready to quit
4. Assisting the smoker with treatment (e.g., prescribing medication, offering counseling, etc.)
5. Arranging follow-up contacts.

As for asthma documents, a great deal of ambiguity in terms of the level of control of asthma might be resolvable with adoption of the Asthma Control Test as a regular assessment of control at every visit for patients with a history of asthma. This test would

at least afford raters a quantifiable and consistent measure of control to compare visit to visit. This approach places the burden of data collection on the clinician, as directed by the electronic system. As with any requirement of this type, great care must be taken to balance the need for complete information against the precise purpose of the ambulatory care visit, the severity of the patient's clinical presentation, and the time the clinician has to address what s/he may consider to be extraneous issues at this encounter.

Often, however, our computational solutions to resolving data ambiguity appear to stop at the data collection level. This, however, is only a part of the problem. Even when data are collected, their subsequent display can be confusing, and more attention to optimal data rendering might help mitigate some of these issues. For example, perhaps it is time to leverage many of the advances in NLP in real-time in the clinical record. One of the documents in the smoking set explicitly stated the patient had no smoking history, and then proceeded to discuss smoking cessation and a prescription for Chantix for the patient. Given this is a relatively straight-forward contradiction, one wonders if it might be possible to notify the documenting clinician of the obvious lack of congruence in the data, so that it might be rectified before the document is closed. In addition, it may be reasonable to build a vocabulary to express smoking status derived from mining the data in the document. For example, if we are interested in the 5 A's as previously discussed, then is it reasonable to summarize the collected information in some standard way, such as "The patient smokes 2 packs of cigarettes a day, was advised to quit, and asked if she wished for assistance. She declined at this time." One wonders if such simple data summaries for topics such as smoking could be developed to these ends.

Ad hoc heuristics can help resolve ambiguity

This second qualitative theme suggests that there are multiple ways to determine facts, if sufficient data points are available to reach a conclusion. This study has demonstrated that users often form heuristics to accomplish this task, and that the quality of the heuristics can vary dramatically. The nature of this theme implies that perhaps we can computationally leverage simple heuristics to assist the clinician or subsequent document coder. We have alluded to this in the mention of data summaries with standardized expressions in the paragraph above. Again, the question becomes one of whether or not it is possible to leverage such heuristics in real time to support the user in making the best determination based on the data. This would require a separate study and one more rooted in formal NLP testing and analysis.

Temporal issues can create confusion

As long as free text continues to be used for data collection, the likelihood of imprecise terms entering the clinical record remains high. As noted in this study, concept modifiers representing indeterminate time (e.g., “recently”, “in the last little bit”) are difficult to interpret in context. In such cases, perhaps the best that can be done is to educate clinicians as to how difficult such terms are to understand at a later time, and to discourage use of these words. This requires yet another level of conscious attention on the part of the documenting clinician which may receive much lower priority than other, more pressing issues. Thus, we look to potential computational solutions to help mitigate some of these problems. For example, use of the ACT should help eliminate terms such as “recently” for the duration of current asthma control. In addition, roll-over definitions

may resolve the confusion when a clinician or coder cannot determine if a person who has quit smoking for two weeks is actually a smoker or not, because the definition should hopefully prescribe some concrete bounds on the length of time a person has to have gone without smoking anything (or been exposed to second or third-hand smoke) to be considered a non-smoker. Clinical definitions, available in real time, are already a reality with Info Buttons[95], on-line clinical databases and other lookups. It is time to embed this sort of functionality in all clinical systems. Such solutions help get the data in the system cleanly.

In terms of rendering data with its temporal context, we may miss significant opportunities to *enhance* data display to support general cognition regarding time. For example, most would agree that reading that a patient is 43 years old is easier than having to determine the patient's age on a given day using only the patient's date of birth. Indeed, for this reason, most clinical information systems perform such real-time calculations for the clinician and display the age in the interface. However, this represents only one of a large collection of time-related supports that electronic systems could provide. For example, it seems reasonable that any date (e.g., date of a study, sample collection, surgery, etc.) should provide a roll-over indicating how long ago, in years, months, weeks, days, or hours a data point was collected. Furthermore, should three or more of the same data points exist in the system, the roll-over might graph them so that the points could be visually trended. Finally, the clinician may wish to know where this patient falls in the distribution of patients (of this age, gender, etc.) at this

institution.[xxi] Finally, it seems reasonable that electronic systems should be able to leverage hospital-specific information about patient populations to offer the clinician support. For example, it should not be difficult to determine how many patients with a similar diagnosis have, say, elevated blood pressure, and to provide information about how it is treated.[xxii]Such information may provide enough contextual information to more adequately inform decision making based on previously collected data.

Exceptions to rules can create confusion

In general, we have addressed these issues by stating that both clinicians and coders need clear definitions of clinical conditions. We do not yet know how exceptions fit within normal clinical presentations and how these exacerbations affect levels of agreement among coders. If such exceptions have significant relevance for accurate coding of clinical documents, then efforts to address how exceptions (e.g., "smoked two cigarettes last week") or single exacerbations (e.g., "needed a rescue inhaler when visiting a friend who just got a kitten") affect the accuracy of the diagnosis must be clarified. This is not an easy task, particularly in light of the fact that many patients have multiple chronic and interacting diagnoses, making potential exceptions or exacerbations difficult to quantify by level of importance in a given situation. Again, it is worth investigating embedding simple NLP in real-time analysis as data is collected, to signal where exceptions or exacerbations might be clinically significant.

[xxi] For an excellent example of how this might look, please check http://www09.wolframalpha.com. In the search window, enter "blood pressure 180 over 100". The term is defined and the values are displayed relative to the normal population.

[xxii] In the old days, we used to run "canned" searches in the off hours to build this information for patients scheduled to attend ambulatory visits the next day. This does not seem unreasonable for current systems; it offloads processing overhead to night time hours and the collected data are saved in tables so that others may access them as well.

Some ambiguity is irresolvable post hoc

On one hand, there is little to say here, because this category seems somewhat obvious. On the other hand, if we return to the larger picture of data collection and rendering, this study suggests other potential mitigators for the ambiguity problem. For documents in this set, one issue consistently stands out as problematic: the existence of ICD-9 codes in clinical documentation. In addition, the qualitative data supports the use of standardized vocabularies for representing clinical data. Finally, this study suggests that several interventions in the electronic health record to support general cognition (as discussed in the section on temporal data) may mitigate ambiguous data interpretations. Each of these topics is discussed in detail, with recommendations for mitigating these issues.

6.2.5. Recommendations to Reduce Ambiguity

The use of ICD-9 Codes in clinical documents

Before this is discussed in detail, with one particularly salient example, it is worth noting how ICD-9 codes may enter clinical documentation in the first place.

When a new clinical information system is brought up at an institution, it makes great sense to leverage all available electronic information to pre-populate the clinical record with skeleton data, even if these data are less than perfect. For example, in the absence of an electronic problem list, a filtered[xxiii] list of billing codes (such as ICD-9s or procedure

[xxiii] By "filtered" we mean limiting the list to major diagnoses or procedures. For example, IV placement is a billable procedure (CPT) code. Populating a problem or procedure list with this information would likely

codes) serves as an excellent, preliminary proxy for this list. Despite the fact that patients may have been treated at other institutions, and this institution's billing code list may be incomplete, having this limited "history" is better than none at all. In the documents used in this study, ICD-9s were indeed referenced as problem lists. In addition, clinicians often appear to have copied and pasted these lists into the plan section of their notes to serve as templates for addressing patient care issues. Using these codes to make a new system useful at the start makes enormous sense, not the least of which is time-savings for clinicians. However, the persistence of ICD-9 codes over the long term can promote ambiguity if a) incorrect or completely outdated codes are not removed from the record or b) the codes are not validated and clarified. These codes, after all, represent abstractions of previous clinical encounters, entered into the system usually by specially trained non-clinicians, for the purposes of billing the patient for the visit. There is no question that data are lost in such abstractions. For this reason, this investigator considers ICD-9 codes in the clinical record, persisting long after their initial intent, as problematic.

The following provides an example of the ambiguity that can be created by the use of ICD-9 codes as proxies for problem lists. This statement occurred in several smoking documents:

"*Tobacco Use Disorder: In Remission.*"

This code actually represents a single ICD-9 code: 305.1 Tobacco Use Disorder[96] with the associated modifier "In Remission". In all likelihood, the patient is indeed a former smoker (though the definition of Tobacco Use Disorder does not mention smoking as the

---

overload the list with information which may not serve the intended purpose of building a preliminary problem or procedure history list.

mechanism of tobacco use), and at some point, a responsible clinician noted this in the clinical record. This information was likely correctly abstracted as "Tobacco Use Disorder" by billing personnel. At a later date, it is equally likely that a clinician documented that the patient quit smoking, and again, billing personnel correctly noted this and added "In Remission" to the diagnosis to reflect this new information. Thus, the data are likely correct, and, for the most part, this sort of phrasing made sense to the clinicians reading it. So, in this case, clinical domain knowledge, in the form of familiarity with ICD-9 code short forms and modifiers definitely gave clinicians an advantage over non-clinicians. This is evidenced by the fact than many non-clinicians found this statement to be especially confusing; one subject commenting "*who writes this way and what does this mean?"* Another commented they were unsure if "*the person used to smoke but doesn't now*" or "*the person is somehow cured of Tobacco Use Disorder, for say, chewing tobacco, but may or may not smoke now.*" This subject is theoretically correct, because the actual code description makes no mention of smoking, merely that tobacco is abused. It is only by tacit convention that clinicians likely agree that the complete reference (e.g. "Tobacco Use Disorder: In Remission") is a) to smoking, b) relates to cigarette or cigar smoking, and c) that it is no longer an issue (or wasn't at the last visit). Clinicians would likely expect to see documented drug use, even if it is represented as one or more ICD-9 codes, under some code other than "Tobacco Use Disorder."

There may be several approaches to resolving this issue. Again, it must be stated that the use of historic ICD-9 codes as proxies to establish a problem list for patients in the absence of other specific data is not inherently bad. Despite the fact that these codes

represent secondary data, they remain better than no data at all, and give clinicians at least a minimal foundation of clinical information on which to build. However, the persistence of these codes in the clinical record may unintentionally promote ambiguity. We offer several suggestions for mitigating some of these issues.

First and foremost, this investigator believes ICD-9 codes should be rendered in a visually distinct manner within the record. Often, but not always, in the documents in this study, ICD-9 codes were rendered in capital letters. This is an inconsistent visual cue, and could easily be confused with other data included in the record also rendered in capital letters, such as when the user inadvertently presses the "Caps Lock" button while typing, or when data from linked systems imports in full capital letters, or if (rarely) capitals are used for emphasis by the documenting clinician. Thus, a better visual cue and one reserved for ICD-9s alone is recommended. A likely candidate is background highlighting in a single color reserved for this purpose. With such a visual cue, all persons reading clinical documentation would know when the text they are reading represents an abstraction of historical data, and that this abstraction is highly codified.

Secondly, the fact that most (but not all) clinicians appeared to understand the meaning of "Tobacco Use Disorder: In Remission" suggests that there is potential for confusion in interpretation when the codes are used. Additionally, this can be confounded in those instances where many documents list a code, *but include no subsequent information to address it*, potentially leaving the reader wondering what purpose the code serves, if it actually applies to this patient, it is a current problem, or is a former issue. The asthma documents contained two examples of this issue. In both cases "Asthma: Unspecified"

(ICD 493.9) appeared in the list of problems but was never again mentioned or addressed in the remainder of the encounter note. In terms of currency at this visit, this presents an additional problem: historical data appears to be current, but may not be; indeed, the information may not even be correct. As one clinician stated "*…incorrect ICD-9s can and often do stick with patients for a lifetime.*"

There are several approaches that might be used to help resolve these problems. The first is to resolve the ICD-9 codes into a functional problem list, and the second involves requiring that problems on the list be addressed (always, if clinically feasible). One solution to the first of these issues, and perhaps the most onerous to clinicians, is to require, perhaps over the course of three clinical encounters with a patient, that ICD-9 codes be transformed to a standard problem list, including the date of diagnosis, and the data of resolution (e.g., "In Remission") should the problem be resolved. Much of this can be performed computationally buy locating and displaying this information with the ICD-9 code, as the date of initial assignment of the code and the date of the addition of the modifier are known. The requirement could be enforced by disallowing signature locking of a record until this issue is resolved.[xxiv] Specifically, this means that ICD-9 codes could be presented in problem list format, with the provision that clinicians could sort the list by currency, remove incorrect diagnostic codes (as opposed to merely indicating that the problem was resolved), comment upon codes to clarify meaning, and sort the list by relevance or clinical importance. For this last suggestion, it would likely make more sense were clinicians allowed to replace the text of the ICD-9 code with more

[xxiv] This is admittedly heavy-handed, but would serve to resolve the problem over time. This approach also assumes that a well-defined, standardized, and clinically useful problem list format exists.

natural English (e.g., “The patient does not currently smoke.”) or if the system replaced the short form of the code with its short or long formal definition (or create a roll-over over the code itself that provided the definition).

In addition, should ICD-9 codes represent a problem list in the record, after clinicians have been able to edit them to represent more of the patient’s current status, then each of these diagnoses must be addressed in the record (time and patient care issues considered) before the record can be signed (or, “locked”). Thus, an ICD-9 code of “Tobacco Use Disorder: In Remission” would require some annotation as to whether or not the patient remains “In Remission” as this is not a permanent state with this disorder. Similarly, a diagnosis code of “Asthma: Unspecified” would require commentary so that this data point does not persist as an unsubstantiated diagnosis.

The Use of Standard Phrases

Despite the preceding commentary about the dangers of using ICD-9 codes as proxies for other clinical data, there is strong qualitative evidence from this study supporting the use of standardized (e.g., automated or pre-coded) phrases. For example, in support of data standardization, some of the clearest phrases are also the shortest, and pack a great deal of information into a tidy package. Consider: “*smokes 1 ppd*”, “*reason for visit: asthma poorly controlled*”, “*Quit smoking tobacco, 1998. No smoking.*” This strongly suggests that clean, directive user interfaces that collect this information are perhaps the first of the strong clinical tailoring areas to target, particularly with clinical decision support. It should be possible to assure capture of this data in complete form. If, indeed, smoking status assessment is to become, for example, part of a pay for performance issue then the

information has to be *simple to collect completely*, and *explicitly rendered* leveraging the informatics and human computer interface knowledge produced over the last 40 years of research. Approximately 25% percent of the disagreement among subjects in this study appears resolvable by *precise data collection and representation to the user*. This is not to say that all clinical data can be collected in this manner, but that data that could be, should. Perhaps a reasonable starting point would be to align standardized or menu-driven data collection to pay-for-performance measures, or to specific quality improvement measures (such as those collected by the National Surgical Quality Improvement program, or NSQIP: http://www.acsnsqip.org), data collection by quality of care guidelines, or any of a number of institution-driven quality improvement efforts. Each of these approaches represents a starting point, which can be quantitatively investigated to determine if data quality in terms of precision and comprehensiveness after such standardization is implemented. As always data quality remains only a proxy for quality of care. These measures must be tied to actual clinical outcomes (e.g., Does the number of smokers decrease when the Five As are documented?) as forcing clinicians to enter data that ultimately serves little purpose for actual patient outcomes. The success or failure of such approaches can be used as a guide for subsequent implementations of this sort.

Clinical decision support, general cognition support or both?

The types of disagreement noted by all subjects in this study were *predominantly non-clinical in nature*. Clinical inference was apparently needed in the rarest of cases, and even in those cases, the experts were unable to agree on an answer among themselves.

As discussed, disagreement appeared to result primarily from confusion in language meaning. This can be said with a very strong qualification that the documents selected for the study may not have represented a set of sufficient complexity and differentiation to adequately confuse non-clinicians. However, this result was surprising. It suggests that we may need to reframe at least a portion of the clinical decision support domain as *cognition support*, independent of medical domain expertise.

It may be that we operate under a false assumption about how we can best support clinicians at the computer, because we inadvertently assume decision support must support *clinical* decisions only. Instead, much of our focus may need to return to how to support *general cognition* to reduce the sorts of ambiguity propagated in clinical records through less than optimal data collection techniques, a failure to leverage NLP to highlight ambiguous information within clinical documents in real time, and the problems that may arise from rote regurgitation of collected information without associated contextual information that our systems could bring to bear. This suggests that attention to general cognition support must accompany the design of clinical decision support; that is, the two must be designed and evaluated in tandem.

We have suggested several basic cognition support issues, such as the use of "roll-overs" to provide definitions or/ calculate the difference between the date an information item was collected and today's date. In addition, real time linkage to information support tools and systems (such as Info buttons, clinical databases, etc.) appears to be ready for prime time, in that it is time to embed these resources into systems. Although this study did not attempt to investigate these issues specifically, these topics did emerge as potential

solutions to some of the cognitive confusion subjects discussed. This suggests that other basic cognition support features, properly embedded in clinical systems may further help reduce ambiguity. For example, if copying and pasting text propagates errors, then highlighting text that has been copied and pasted may alert clinicians to the need to review and clarify the text appropriately for the current clinical encounter. In addition, we do not yet know how much of the long-term research in human computer interface design is leveraged in the design of our clinical information systems. For example, in Western cultures, where we normally read from left-to-right and top-to-bottom, it is commonly accepted that important information be placed to the top and left of a screen footprint where it will most likely be seen due to the natural inclination to begin reading in that location. There is no evidence that this simple heuristic is followed or indeed if it improves the clinician's ability to identify important information. There is much work left to do in this regard.

Similarly, can we find some way to relay specific information to subsequent clinicians when notes include statements such as "*I have reviewed the patient's history and note no significant changes*"? Although such statements may provide some legal protection for clinicians, these statements do little to forward clinical information, especially if the statement is made on review of paper record, and represents the first entry regarding a patient's history in the electronic health record! It is difficult to determine how to prevent this from happening, though it does suggest an interesting point of intervention which may help resolve the ambiguity problem. Work continues on summarizing clinical encounters in meaningful ways so that important information can replace statements such as the one above.

7. Study Limitations

I have already alluded to many of the issues that may have confounded the quantitative results, and will note readdress these issues here. It is noteworthy that very little research exists regarding the precise cognition of clinical concept recognition, and that an exploratory study such as this is bound to illustrate study design issues that can only best be understood with considered hindsight. This is a positive outcome actually, as it suggests many subsequent studies to address the types of potential confounders we encountered. For example, we have addressed the fact that the distinction between textual and inference-based tasks may not represent a valid distinction in cognitive task complexity. This encourages further study into methods not only for elucidating this distinction and determining whether or not it actually matters.

The lay person study group did not represent the standard lay population; all were highly educated (most with graduate degrees) and had significant exposure to the health care environment. This certainly biases the results towards greater agreement between the study groups. It is very reasonable to assume different results might emerge from this study if lay subjects were chosen from the general population or if there were no requirements for college-level education or fluency in English. For these reasons, any conclusions drawn regarding differences in expert and lay cognition in disagreement when coding clinical documents must be evaluated with some care. For example, these results might support the idea that no clinical expertise is required for accurate coding of medical documents. This study cannot be used to support that statement with any certainty. This study used ambulatory care documents collected from a single electronic

record system from two general medicine clinics. In addition, ICD-9 codes were used to identify the records, making them specific for this study. It is highly unrealistic to assume that these results are reproducible across medical specialties or into the inpatient and specialty service worlds, despite the fact that these differences are certainly intriguing and warrant further exploration.

No mixed methods study should report results as final when the analysis has been performed by a single researcher. This surfaces issues of rigor in the quantitative arena and depth and breadth of scope in the qualitative analysis. Though I have made every effort to be precise in my reporting of statistical significance, I have possibly erred. In addition, without either additional reviewer input or further confirmation with the subjects participating in the study, qualitative verification remains partially incomplete, due to an inability to formally triangulate the results. In general, I am convinced this area of inquiry warrants a more detailed study design with larger and more complex document sets.

8. Conclusion

This study began looking for the answers to several questions related to agreement between lay and expert raters. Although the quantitative results did not fully support either hypothesis, the qualitative analysis of subject comments offered valuable insight into steps we can take to reduce some of the ambiguity, confusion, and cognitive overhead built into clinical systems. Both clinical decision *and* cognition support can truly support clinical thinking and decision making, and when combined should serve as a precursor in efforts to make document coding more consistent. To these ends, further research into cognition and computers is definitely warranted. In addition we must integrate current knowledge about efficiency and usability in human computer interface design to reduce ambiguity beginning at the point of data collection. Finally, we must look for better ways to leverage computation integration of data so that it can be rendered back to the user in a cognitively precise manner.

In this study, clinicians appear to struggle cognitively with confusion the same way lay persons do. When clinicians lack explicit data, they must make inferences (if possible). Lay people must do the same things. Clinicians have trouble determining proper bounds on time. So do lay people. Exceptions to rules cause confusion within both groups. Simply put, in this study, clinicians and lay persons identify highly similar points of cognitive confusion overall and clinical knowledge plays a relatively small role in isolating points of disagreement. Significantly, all subjects appeared able to identify salient clinical details, however not all subjects knew how to interpret them.

Clearly, providing subjects with definitions of terms and heuristics for combining findings into a diagnosis could help resolve this problem, if sufficient information were present in the record. Other researchers, as mentioned in the background section, have demonstrated that non-experts can approach expert performance on coding tasks when given and trained on thorough coding schemas prior to the task. It is noteworthy that without such schemas, the subjects in this study still managed to identify relevant information, even if in some cases the relevance to the subject was not particularly clear.

Thus, this study suggests that improving the collection and dissemination of data to optimize the electronic health record as a rich communication among providers over time is one of the primary tasks requiring our attention now. Both general cognition and clinical decision support should make it easier to enter data, verify data associations, and render information to maximize its usefulness to the user, in context to support this focus. It is evident that we can help people agree more with careful consideration of the way in which we collect, analyze, and re-display data computationally. It is time to investigate what gets lost, added, modified, or mangled in translation, and how we can mitigate these effects. This study suggests that we should focus efforts in clinical decision support on the proper, precise, and complete collection of sufficient data to render a clinically relevant statement that is equally proper, precise, and complete for all to share for providing high quality health care.

References

1. U.S. National Library of Medicine and the National Institutes of Health. Pubmed. http://www.ncbi.nlm.nih.gov/sites/entrez/. Accessed June 13, 2008.

2. Hersh WR. *Information Retrieval: A Health and Biomedical Perspective*. 2nd ed. New York: Springer-Verlag; 2003.

3. Busemeyer JR, Myung IJ. An Adaptive Approach to Human Decision Making: Learning Theory, Decision Theory, and Human Performance. *Journal of Experimental Psychology: General.* 1992;121(2):177-194.

4. Chapman WW, Dowling JN. Inductive Creation of an Annotation Schema for Manually Indexing Clinical Conditions from Emergency Department Reports. *Journal of Biomedical Informatics.* 2006;39:196-208.

5. Chapman WW, Dowling JN, Hripcsak G. Evaluation of Training With an Annotation Schema for Manual Annotation of Clinical Conditions From Emergency Department Reports. *IJMI.* 2008;77:107-113.

6. Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics.* 2002;35:99-110.

7. Hripcsak G, Kuperman GJ, Friedman C. Extracting findings from narrative reports: Software transferability and sources of physician disagreement. *Methods Inf Med.* 1998;37:1-7.

8. Hripcsak G, Wilcox A. Reference Standards, Judges, and Comparison Subjects: Roles for Experts in Evaluating System Performance. *JAMIA.* 2002;9:1-15.

9. Smith MA, Atherly AJ, ;, Kane RL, Pacata JT. Peer Review of the Quality of Care: Reliability and Sources of Variability for Outcome and Process Measurements. *JAMA.* 1997;278(19):1573-1578.

10. Stemler SE. A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability. *Practical Assessment, Research & Evaluation.* Vol 9; 2004:1-19.

11. Wilcox A, Hripcsak G. The Role of Domain knowledge in Automating Medical Text Report Classification. *JAMIA.* 2003;10:330-338.

12. Yawn BP, Wollan P. Interrater Reliability: Completing the Methods Description in Medical Records Review Studies. *American Journal of Epidemiology.* 2005;161(10):974-977.

13. Hripcsak G, Rothschild AS. Agreement, the F-Measure, and Reliability in Information Retrieval. *JAMIA.* 2005;12:296-298.

14. Hripcsak G, Kuperman GJ, Friedman C, Heitjan DF. A reliability study for evaluating information extraction from radiology reports. *JAMIA.* 1999;6(2):143-150.

15. Opila DA. The Impact of Feedback to Medical Housestaff on Chart Documentation and Quality of Care in the Outpatient Setting. *Journal of General Internal Medicine.* 1997;12:352-356.

16. Peabody JW, Luck J, Glassman P, Dressehaus TR, Lee M. Comparison of Vignettes, Standardized Patients, and Chart Abstraction. *JAMA.* 2000;283(13):1715-1722.

17. Socolar RR, Raines B, Chen-Mok M, Runyan DK, Green C, Paterno S. Intervention to Improve Physician Documentation and Knowledge of Child Sexual Abuse: A Randomized, Controlled Trial. *Pediatrics.* 1998;101(5):817-824.

18. Vergis A, Gillman L, Minor S, Taylor M, Park J. Structured Assessment Format for Evaluating Operative Reports in General Surgery. *Am J Surg.* 2008;195:24-29.

19. Weiss KB, Wagner R. Performance Measurement Through Audit, Feedback, and Profiling as Tools for Improving Clinical Care. *CHEST.* 2000;118(2):53S-58S.

20. Baecker RM, Grudin J, Buxton W, Greenberg S. Chapter 9. Human Information Processing. In: Baecker RM, Grudin J, Buxton W, Greenberg S, eds. *Readings in Human-Computer Interaction: Toward the Year 2000*. San Francisco: Morgan Kaufman; 1995:573-586.

21. Gardner H. *The Mind's New Science: A History of the Cognitive Revolution*. New York: Basic Books, Inc.; 1985.

22. Green D. Introduction. In: Green D, et al, ed. *Cognitive Science: An Introduction*. Cambridge, Massacheusetts: Blackwell Publishers, Inc.; 1996:1-22.

23. Patel VL, Arocha JF, Kaufman DR. A Primer on Aspects of Cognition for Medical Informatics. *JAMIA.* 2001;8:324-343.

24. Hall JS. *Beyond AI: Creating the Conscience of the Machine.* Amherst, New York: Prometheus Books; 2007.

25. Lycan W, ed. *Mind and Cognition: An Anthology.* 2nd ed. Malden, Massacheusetts: Blackwell Publishers, Inc.; 1999. Lycan W, ed.

26. Chomsky N, Katz J. What the Linguist is Talking About. In: Garfield JL, ed. *Foundations of Cognitive Science: The Essential Readings*. New York: Paragon House; 1990:332-350.

27. Anderson JR. Chapter 5: Abstraction of Information Into Memory. *Cognitive Psychology and Its Implications*. 6th ed. New York: Worth Publishers; 2005:139-170.

28. Garfield JL. Convention, Context, and Meaning: Conditions on Natural Language Understanding. In: Garfield JL, ed. *Foundations of Cognitive Science: The Essential Readings*. New York: Paragon House; 1990:3-17.

29. Stedman TL. *Stedman's Medical Dictionary*. 26th ed. Baltimore: Williams & Wilkins; 1995.

30. Dudai Y. *Memory from A to Z: Keywords, Concepts, and Beyond*. New York: Oxford University Press; 2004.

31. Afifi AK, Bergman RA. *Functional Neuroanatomy: Text and Atlas*. 2nd ed. New York: Lange Medical Books/McGraw-Hill; 2005.

32. Ganong WF. Chapter 16. Higher Functions of the Nervous System. *Review of Medical Physiology*. 22nd ed: The McGraw-Hill Companies; 2005.

33. Nolte J. Chapter 23: Drives, Emotions, and Memories: The Hypothalamus and Limbic System. *The Human Brain: An Introduction to Its Functional Anatomy*. 5th ed. St. Louis: Mosby; 2002:576-578.

34. Ericsson KA, Kintsch W. Long-term Working Memory. *Psychological Review*. 1995;102(2):211-245.

35. Preston A. How Does Short-Term Memory Work in Relation to Long-Term Memory? *Scientific American* [September 26, 2007]; http://www.sciam.com/article.cfm?id=experts-short-term-memory-to-long-term. Accessed June 21, 2008.

36. Gaffan D. What Is a Memory System? Horel's Critique Revisited. *Behavioural Brain Research.* 2001;127:5-11.

37. Morris R. Elements of Neurobiological Theory of Hippocampal Function: The Role of Synaptic Plasticity, Synaptic Tagging and Schemas. *European Journal of Neuroscience.* 2006;23:2829-2846.

38. Hall AD. Chapter 2: Basic Percepts, Concepts, and Precepts. *Metasystems Methodology: A New Sysnthesis and Unitfication*. New York: Pergamon Press; 1989:53-124.

39. Lendaris GG. Systems, Man, and Cybernetics. *IEEE Transactions on Systems, Man, and Cybernetics.* 1986;SMC-16(4):604-610.

40. Atkinson R, Shiffrin R. Human Memory: A Proposed System and Its Control Processes. In: Spence K, ed. *The Pshychology of Learning and Motivation: Advances in Research and Theory*. New York: Academic Press; 1968:89-195.

41. Baddeley AD, Hitch GJ, eds. *The psychology of learning and motivation: advances in research and theory*. New York: Academic Press; 1974. Bower G, ed; No. 8.

42. Baddeley AD. The Episodic Buffer: A New Component of Working Memory. *Trends in Cognitive Sciences.* 2000;4(11):417-423.

43. Baddeley AD. Is Working Memory Still Working? *American Psychologist.* 2001;56(11):851-864.

44. Woltz DJ, Was CA. Available But Unattended Conceptual Information in Working Memory: Temporarily Active Semantic Content or Persistent Memory for Prior Operations? *Journal of Experimental Psychology.* 2007;33(1):155-168.

45. Newell A, Simon H. *Human Problem Solving*. Englewood Cliffs, NJ: Prentice Hall; 1972.

46. Gobet F. Some Shortcomings of the Long-Term Working Memory. *British Journal of Phsychology.* 2000;91:551-570.

47. Cowan N. The Magical Number 4 in Short-term Memory: A Reconsideration of Mental Storage Capacity. *Behavioral and Brain Sciences.* 2000;24:87-185.

48. Oberauer K. Access to Information in Working Memory: Exploring the Focus of Attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* 2002;28(3):411-421.

49. Halford GS, Baker R, McGredden JE, Bain JD. How Many Variables Can Humans Process? *Psychological Science.* 2005;16(1):70-76.

50. Card SK, Moran TP, Newell A. *The Psychology of Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1983.

51. Miller G. The Magic Number Seven, Plus or Minus Two: Some Limits on Our Capactity for Processing Information. 1956. *Psychological Review.* 1994;101(2):342-352.

52. Cooper R. Explanation and Simulation in Cognitive Science. In: Green D, ed. *Cognitive Science: An Introduction*. Cambridge, Massachusetts: Blackwell Publishers, Ltd.; 1996:416.

53. Just MA. *The Psychology of Reading and Language Comprehension*. Newton, Massachusetts: Allyn and Bacon, Inc.; 1987.

54. Turner RM. A View of Diagnostic Reasoning as a Memory-directed Task. Paper presented at: Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society, 1992; Bloomington, Indiana.

55. Bransford J, Johnson M. Consideration of Some Problems of Comprehension. In: Chase W, ed. *Visual Information Processing*. New York: Academic Press; 1973.

56. Sowa JF. Semantic Networks. 6/2/2006; http://www.jfsowa.com/pubs/semnet.htm. Accessed 9/1/2008.

57. Collins AM, Loftus EF. A Spreading-Activation Theory of Semantic Processing. *Phsychological Review.* 1975;82(6):407-428.

58. National Library of Medicine. Unified Medical Language System. *UMLS Knowledge Sources*: U.S. Department of Health and Human Services, National Institutes of Health; 2007.

59. Qillian MR. Semantic Memory. In: Minsky M, ed. *Semantic Information Processing*. Cambridge, Massachusetts: The MIT Press; 1968:216-270.

60. Anderson JR. A Spreading Activation Theory of Memory. *Journal of Verbal Learning and Verbal Behavior.* 1983;22:261-295.

61. Ratcliff R, McKoon G. Does Activation Really Spread? *Psychological Review.* 1981;88(5):454-462.

62. Anderson JR, Pirolli PL. Spread of Activation. *Journal of Experimental Psychology.* 1984;10(4):791-798.

63. Friedrich FJ, Henik A, Tzelgov J. Automatic Processes in Lexical Access and Spreading Actviation. *Journal of Experimental Psychology.* 1991;17(3):792-806.

64. Ratcliff R, McKoon G. A Retrieval Theory of Priming in Memory. *Psychological Review.* 1988;95(3):385-408.

65. Pusen C, Erickson JR, Hue C-W, Vyas AP. Priming from Category Members on Retrieval of Other Category Members: Positive and Negative Effects. *Journal of Experimental Psychology.* 1988;14(4):627-640.

66. Lewis RL. *An Architecturally-Based Theory of Human Sentence Comprehension*. Pittsburgh, PA: Computer Science, Carnegie Mellon University; 1993.

67. Clark HH, Clark EV. *They Phsyclology of Language: An Introduction to Psycholnguistics.* New York: Harcourt Brace Jovanovich.; 1977.

68. Posner MI, Pavese A. Anatomy of Word and Sentence Meaning. *Proceedings of the National Academy of Sciences of the United States of America.* 1997;95(3):899-905.

69. Rosch E. Natural Categories. *Cognitive Psychology.* 1973;4:328-350.

70. Rosch E, Mervis CB, Gray WD, Johnson DM, Boyes-Braem P. Basic Objects in Natural Categories. *Cognitive Psychology.* 1976;8:382-439.

71. Lakoff G. *Women, Fire, and Dangerous Things.* Chicago: The University of Chicago Press; 1987.

72. Brown R. How Shall a Thing Be Called? *Psychological Review.* 1958;65(1):14-21.

73. Brown R. *Social Psychology*. 2nd ed. New York: Free Press; 1968.

74. Rosch E, Simpson C, Miller RS. Structural Basis of Typicality Effects. *Journal of Experimental Psychology: Human Perception and Performance.* 1976;2(4):491-502.

75. Berlin B, Breedlove D, Raven P. Folk Taxonomies and Biological Classification. *Science.* 1966;154:273-275.

76. Berlin B, Breedlove D, Raven P. General Principles of Classification and Nomenclature in Folk Biology. *American Anthropologist.* 1973;75:214-242.

77. Lewis T, Amini F, Lanon R. *A General Theory of Love*. New York: Vintage Books; 2000.

78. Anderson JR. Chapter 9: Expertise. *Cognitive Psychology and It's Implications*. 6th ed. New York: Worth Publishers; 2005:279-311.

79. Charness N. Components of Skill in Bridge. *Canadian Journal of Psychology.* 1979;33(1):1-16.

80. Sloboda JA. Visual Perception of Musical Notation: Registering Pitch Symbols in Memory. *Quarterly Journal of Experimental Psychology.* 1976;28:1-16.

81. Charness N. Expertise in chess, music, and physics: A cognitive perspective. In: Obler LK, Fein D, eds. *The Exceptional Brain: Neuropsychology of Talent and Special Abilities*. New York, NY: Guilford Press; 1988:399-426.

82. McKeithen KB, Reitman JS, Rueter HH, Hirtle SC. Knowledge Organization and Skill Differences in Computer Programmers. *Cognitive Psychology.* 1981;13:307-325.

83. National Institute of Standards and Technology. The NIST Reference on Constants, Units, and Uncertainty. December, 2003; http://physics.nist.gov/cuu/Units/current.html. Accessed August 4, 2008, 2008.

84. Tinsley HE, Weiss DJ. Interrater Reliability and Agreement of Subjective Judgements. *Journal of Counseling Psychology*. 1975;22(4):358-376.

85. Rosner B. *Fundamentals of Biostatistics*. 6th ed. Belmont, CA: Thompson Brooks/Cole; 2006.

86. Duda RO, Hart PE, Stork DG. *Pattern Classification*. Vol 2008. 2nd ed. New York: John Wiley & Sons; 2001.

87. Smith LI. A Tutorial on Principal Components Analysis. http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf. Accessed August 8, 2008.

88. Linacre JM. *Many-facet Rasch Measurement*. Chicago: MESA Press; 1994.

89. Brorson S, Hróbjartsson A. Training Improves Agreement Among Doctors Using the Neer System for Proximal Humeral Fractures in a Systemic Review. *Journal of Clinical Epidemiology*. 2008;61:7-16.

90. LeBlanc A, Robichaud P, Lacasse Y, Boulet L-P. Quantification of asthma control: validation of the Asthma Control Scoring System. *Allergy*. 2007;62:120-125.

91. Chapman WW, Dowling JN, Wagner MM. Generating a Reliable Reference Standard Set for Syndromic Case Classification. *JAMIA*. 2005;12:618-629.

92. Hazlehurst B, Frost R, Sittig DF, Stevens VJ. MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. *JAMIA.* 2005;12(5):517-529.

93. Uzuner O. Recognizing Obesity and Co-morbidities in Sparse Data. *JAMIA.* 2009;16(4):561-570.

94. Ogren PV, Savova GK, Chute CG. Constructing Evaluation Corpora for Automated Clinical Named Entity Recognition. Paper presented at: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08); May 28-30, 2008.

95. Cimino J, Li J, Bakken S, Pate lV. Theoretical, empirical and practical approaches to resolving the unmet information needs of clinical information system users. Paper presented at: Proceedings of the American Medical Information Association, 2002.

96. National Center for Health Statistics (NCHS) and the Centers for Medicare & Medicaid Services (CMS). *The International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM)*. Sixth ed; 2008.