EFFECT OF A METACOGNITIVE INTERVENTION ON COGNITIVE HEURISTIC USE DURING DIAGNOSTIC REASONING

by

Velma Lucille Payne

Bachelor of Science in Computer Science, Oral Roberts University, 1984

Master of Science in Computer Information Systems, Robert Morris University, 1996

Master of Business Administration, Robert Morris University, 1997

Master of Science in Biomedical Informatics, University of Pittsburgh, 2008

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation proposal was presented

by

Velma Lucille Payne

To the following committee members:

Claudia Mello-Thoms, MSEE, PhD,

Assistant Professor of Biomedical Informatics and Radiology, University of Pittsburgh

Mark S. Roberts, MD, MPP,

Professor of Medicine, Health Policy and Management, and Industrial Engineering, University of Pittsburgh

Cleotilde Gonzalez, PhD,

Associate Research Professor, Department of Social and Decision Sciences, Carnegie Mellon University

Pat Croskerry, MD, PhD,

Professor, Department of Emergency Medicine, Dalhousie University

Dissertation Advisor: Rebecca S. Crowley, MD, MS,

Associate Professor of Biomedical Informatics, Pathology and Intelligent Systems, University of Pittsburgh

EFFECT OF A METACOGNITIVE INTERVENTION ON

COGNITIVE HEURISTIC USE DURING DIAGNOSTIC REASONING

Velma Lucille Payne, M.S., MBA, M.S.

University of Pittsburgh, 2011

Abstract

Medical judgment and decision-making frequently occur under conditions of uncertainty. In order to reduce the complexity of diagnosis, physicians often rely on cognitive heuristics. Use of heuristics during clinical reasoning can be effective; however when used inappropriately the result can be flawed reasoning, medical errors and patient harm. Many researchers have attempted to debias individuals from inappropriate heuristic use by designing interventions based on normative theories of decision-making. There have been few attempts to debias individuals using interventions based on descriptive decision-making theories.

Objectives: (1) Assess use of *Anchoring and Adjustment* and *Confirmation Bias* during diagnostic reasoning; (2) Investigate the impact of heuristic use on diagnostic accuracy; (3) Determine the impact of a metacognitive intervention based on the Mental Model Theory designed to reduce biased judgment by inducing physicians to 'think about how they think'; and (4) Test a novel technique using eye-tracking to determine heuristic use and diagnostic accuracy within mode of thinking as defined by the Dual Process Theory.

Methods: Medical students and residents assessed clinical scenarios using a computer system, specified a diagnosis, and designated the data used to arrive at the diagnosis. During case analysis, subjects either verbalized their thoughts or wore eye-tracking equipment to capture eye

movements and pupil size as they diagnosed cases. Diagnostic data specified by the subject was used to measure heuristic use and assess the impact of heuristic use on diagnostic accuracy. Eye-tracking data was used to determine the frequency of heuristic use (*Confirmation Bias* only) and mode of thinking. Statistic models were executed to determine the effect of the metacognitive intervention.

Results: Use of cognitive heuristics during diagnostic reasoning was common for this subject population. Logistic regression showed case difficulty to be an important factor contributing to diagnostic error. The metacognitive intervention had no effect on heuristic use and diagnostic accuracy. Eye-tracking data reveal this subject population infrequently assess cases in the Intuitive mode of thinking; spend more time in the Analytical mode of thinking, and switches between the two modes frequently as they reason through a case to arrive at a diagnosis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DEDICATION

Life threatening harm brought to my parent is the reason I began the pursuit of patient safety and diagnostic errors research. It is my hope that some aspect of this research will one day significantly impact patient safety and enhance patient care so that others do not have to suffer the harm my Father did as a result of flawed judgment during the diagnostic and therapeutic process.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Rebecca S. Crowley, M.D., M.S. for her advice, support and guidance during this research. I also thank my committee members, Pat Croskerry, M.D., Ph.D., Cleotilde Gonzalez, Ph.D., Mark S. Roberts, M.D., MPP, and Claudia R. Mello-Thoms, Ph.D., for their support and guidance. I acknowledge the financial support of the National Library of Medicine and the National Institute of Health; without this support I would not have been able to pursue this lifelong goal.

Heartfelt thanks go to my very special friends and fellow students Chad Kimmel, Himanshu Grover, and Nara Um, who spent countless hours offering me words of encouragement and trying to convince me that one day this pursuit would be worth it. Thanks Chad, Himanshu and Nara. You are special and I will forever have a spot in my heart for you.

Words do not come close to adequately expressing my gratitude to the three men in my life that, along various stages in my life, have always been there for me, have always believed in me, never lost faith in my ability to excel, and who were the few that have touched my life, each in their own special way. To *Ty*, who has by far been the most faithful and loving presence throughout my entire life. I dedicate this work to two very exceptional and special men. I gladly went through the pains of the completion of this degree all to honor my late *Father, Earvin A. Payne* and my late *Uncle, Lee R. Gill* who never lost faith in me and my abilities. I would never have embarked upon this venture had it not been for the foundation you both built for me to climb upon, the soft arms you provided for me to fall into when times were tough and discouragement overcame me, and your faith in me that never waivered. **Dad and Unk, it is an honor to have had you in my life. I dedicate this work to you.** I only wish you had been alive to see the goal accomplished; but somehow I know that you are aware of what has transpired. I hope I have made you proud.

1 INTRODUCTION

1.1 BACKGROUND

According to the Institute of Medicine's (IOM) report *To Err is Human: Building a Safer Health System*, **medical errors** are among the top ten causes of death in the United States.[1] The IOM report mobilized the healthcare industry to address the problem of preventable medical errors. "**Diagnostic errors** comprise a substantial and costly fraction of all medical errors,"[2] represent the second largest cause of adverse events,[3] and are the second leading cause of malpractice suits against hospitals.[4] A study conducted in the United States (U.S.) of autopsy findings identified diagnostic inconsistencies in nearly twenty percent (19.8%) of cases, including missed diagnoses of malignant tumors, misidentification of the location of primary malignant tumors, lack of evidence of malignancy when it was thought the patient had one, and unknown infections.[5]

A variety of attempts have been made to classify diagnostic errors.[6-11] Based on a review of 100 cases of suspected diagnostic errors collected from five large academic tertiary care medical centers over five years, Graber et al. proposed a taxonomy of diagnostic errors including *no-fault errors*, *system-related errors* and *cognitive errors.*[2] Graber et al. defined **cognitive-based diagnostic errors** as errors resulting from "inadequate knowledge or faulty data gathering, inaccurate clinical reasoning, or faulty verification."[2] Examples of cognitive errors provided in their report are "flawed perception, faulty logic, falling prey to biased heuristics, and settling on a final diagnosis too early."[2]

The title of IOM's report *To Err is Human* implies that errors are an inevitable part of human nature and therefore difficult to eliminate. Given the human and economic impact of diagnostic errors, the reduction of these errors is an important and necessary goal of healthcare. This research will address **the reduction of cognitive-based diagnostic errors**, focusing on physicians' misuse of cognitive heuristics during diagnostic reasoning.

1.2 PROBLEM DESCRIPTION

One of the challenges of medicine is decision-making under uncertainty. Based on empirical studies, Kahneman and Tversky found that when people make judgments under uncertainty they rely on identifiable heuristics which reduce complex tasks to simpler judgmental operations.[12-14] A heuristic is a rule-of-thumb, mental shortcut, or guideline that is applied to make complex tasks simpler. Heuristics can lead to appropriate judgments**;** however, they can also lead to severe and systematic errors when not used properly.[15] Kahneman and Tversky refer to such systematic errors as *cognitive biases* or departures from the normative rational theory.[15] Physicians are no exception to Kahneman and Tversky's findings. During the diagnostic process, physicians are required to make critical decisions under conditions of uncertainty, which have been shown to result in cognitive biases.[2,16-25] These biases can produce a variety of medical errors including incorrect diagnosis, delayed diagnosis, and inappropriate treatment. Even though the term 'bias' does not automatically correlate with the occurrence of an 'error', I will follow Kahneman and Tversky's convention by referring to systematic errors resulting from inappropriate use of cognitive heuristics as 'biases'.

1.3 SIGNIFICANCE OF THIS RESEARCH

Many investigators have addressed the question *Can cognitive biases be overcome?*. Several debiasing techniques have been tested with varying success.[9,118,122-132] Many of these debiasing techniques are based on normative decision-making strategies such as instructing subjects on statistics and probability,[127] training them to use Bayes' Theorem to compute the probability of hypothesized diagnoses,[123,130] didactic instruction on cognitive heuristics and biases,[9,118,123, 126,127,129], etc. Published editorials suggest the use of metacognition, or thinking about how one thinks, as a cognitive heuristic debiasing technique.[48,54,56,119,120] An extensive literature review on cognitive debiasing has revealed only one published study attempting the reduction of suboptimal decisions by having clinicians examine their own decision-making processes.[118] Empirical evidence is needed to validate the use of metacognition as a successful means of

debiasing physicians and improving diagnostic reasoning. This research will use a metacognitive intervention as a novel approach to debiasing physician judgment.

This research extends existing debiasing studies by applying principles of the Mental Model Theory (MMT) during diagnostic reasoning, which have not been previously tested. The MMT was selected because there is limited empirical evidence supporting published editorials that metacognition is a valuable tool for improving diagnostic reasoning. In accordance with the MMT, errors occur during reasoning due to (1) inappropriately constructed mental models, and (2) failure to consider all applicable models. This research will utilize an innovative approach to reduce biased judgment by providing feedback regarding the mental models physicians construct during diagnosis and by providing strategies to accurately reason with these models. The feedback will trigger the examination of cognitive processes used during diagnostic reasoning. Basing the feedback on principles of the MMT to cause subjects to examine the mental models they construct and reason with during diagnostic reasoning is a technique designed to induce them to think about how they think in order to determine if metacognition improves diagnostic reasoning. Most interventions to date were based upon normative models of decision theory, as opposed to descriptive models. Approaching the problem of flawed and biased judgment by starting with how people actually think, rather than how they should think, has not been investigated while clinicians go through the diagnostic reasoning process. Empirical evidence is needed to determine the impact of descriptive theories of reasoning and metacognition on such reasoning. This research has the potential to provide this empirical evidence.

A statement made by researchers that have extensively studied heuristics and biases is "there is little direct evidence of the extent to which cognitive biases are leading to diagnostic errors".[151] This research has the potential to contribute to the study of medical decision making by providing insight into the use of heuristics and biases during diagnostic reasoning, and by testing a novel intervention based on a well-described theoretical foundation.

1.4 GUIDE FOR THE READER

Chapter 1 consists of the introduction, background and significance of this research.

Chapter 2 provides a discussion of medical errors, including definitions used in the medical error literature. An overview of research studies identifying types of medical errors is also provided. Since this dissertation focuses on the impact that inappropriate use of cognitive heuristics have on diagnostic accuracy, this chapter also includes a discussion on diagnostic errors and cognitive-based diagnostic errors.

Chapter 3 describes the seminal work of Kahneman and Tversky when assessing the use of cognitive heuristics, biases associated with heuristics, and clinicians' use of heuristics.

Chapter 4 describes decision theories including the Expected Utility Theory, Bayes' Theorem, Mental Model Theory and the Dual Process Theory, and their application in diagnostic reasoning.

Chapter 5 includes an overview of the literature of empirical studies attempting cognitive heuristic debiasing. Since the research will utilize an intervention of feedback, this chapter also discusses previous studies that used feedback.

Chapter 6 includes the research statement including the research objectives and questions, cognitive heuristics studied, research design, instrument, methods, intervention assessed.

Chapter 7 describes the study's results.

Chapter 8 discusses the study's findings, limitations and the conclusions of this research.

2 MEDICAL ERRORS

Medical errors are a significant problem. Since the release of the IOM report estimating deaths from medical errors to be between 44,000 and 98,000, the healthcare industry has made an effort to identify the source of medical errors. This chapter provides definitions of medical error terms; describes research studies that have investigated medical errors; outlines categories and frequency statistics of medical errors; provides diagnostic error statistics; and describes cognitive-based aspects of diagnostic errors.

2.1 MEDICAL ERROR DEFINITIONS

A standard nomenclature of terms related to medical errors does not exist. The terms, definitions and categories of medical errors (Figure 1) developed by the IOM are as follows.[1]

- **Error** - the failure of a planned action to be completed as intended (i.e., error of execution) or the use of a wrong plan to achieve an aim (i.e., error of planning)
- **Adverse Event** – an injury caused by medical management rather than the underlying condition of the patient
- **Preventable Adverse Event** – an adverse event attributable to error
- **Non-Preventable Adverse Event** – an adverse event not attributable to error
- **Negligent Adverse Event** – a subset of preventable adverse events that satisfy legal criteria used in determining negligence, i.e., whether the care provided failed to meet the standard of care reasonably expected of an average physician qualified to take care of the patient in question
- **Near Miss** – Errors that do not result in harm

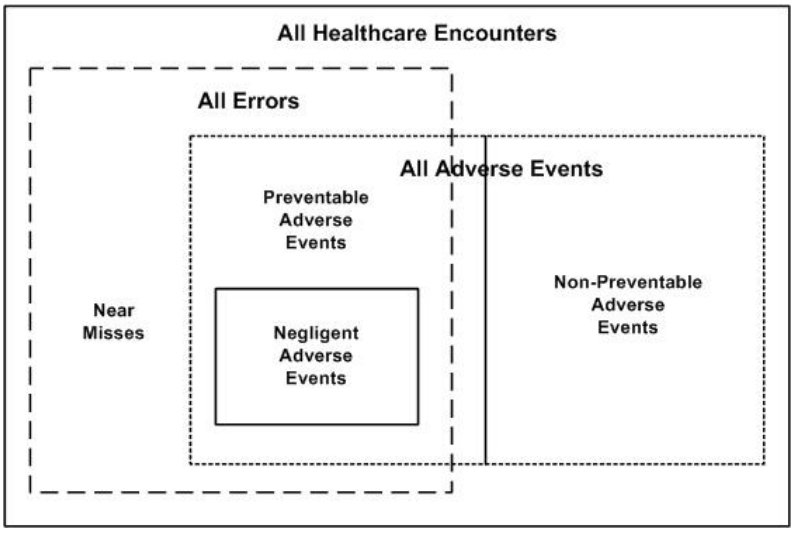


Figure 1 Medical Error Categories[1]

Some *adverse events* are not preventable and are therefore not considered medical errors. *Preventable adverse events* are considered medical errors; and some of these result from negligence. Another category of medical errors are *near misses* which are medical errors that do not harm a patient, but are an error of execution or planning.

2.2 MEDICAL ERROR RESEARCH STUDIES

Several research studies have been conducted to investigate medical errors. The often quoted studies listed in Table 1, provide frequency statistics for various categories of medical errors.[26] These studies can be further distinguished by whether they measure all adverse events, or only medication-related errors.

Table 1 Medical Error Studies [26]

Study	Primary Measurement	Other Measurements	Definitions Researchers Utilized
Harvard Medical Practice Studies (HMPS) – U.S. Based Study			
Brennan et al., 1991 [27]	Adverse events	Negligent adverse events	**Adverse event** - an injury caused by medical management **Negligent adverse event** - injury caused by substandard medical management

Study	Primary Measurement	Other Measurements	Definitions Researchers Utilized
			Both types are, in theory, preventable
Leape et al., 1991 [3]	Classify errors potentially causing adverse events		Study sought to determine if an adverse event could have been caused by **reasonably avoidable error** defined as a mistake in performance or thought.
Colorado and Utah Study – U.S. Based Study			
Thomas, et al., 2000[28]	Adverse events	Negligent adverse event	Same definitions as HMPS.
Thomas, 2000[28]	Preventability		Preventability assessed by two investigators.
Quality in Australian Healthcare Study			
Wilson, et al.,1995 [29]	Adverse event	Preventable adverse event	**Adverse event** - injury caused by medical care rather than the disease process. **Preventability** - an error in management due to the failure to follow accepted practice at an individual or system level.
Wilson, et al., 1999[30]	Identify and classify errors		**Error** - an act of commission or omission that caused, or contributed to, the cause of the unintended injury.
Adverse Drug Event Study Group – U.S. Based Study			
Bates, Leape, et al., 1995 [31]	Adverse drug event Potential adverse drug event	Preventable Adverse Event Severity	**Adverse drug events** were judged preventable if they were due to an error or were preventable by any means currently available (included potential adverse drug events).

The *Harvard Medical Practice Studies*[3,27] (U.S. based study) serves as the benchmark for estimating the extent of medical injuries occurring in hospitals. These United States investigators reviewed a random sample of over 31,000 records from 2,671,863 patients discharged from acute-care hospitals in New York in 1984 (non-psychiatric, pediatric and ob/gyn patients were excluded in this review). They identified 98,610 *adverse events* defined as "an injury that was caused by medical management (rather than the underlying disease) and that prolonged the hospitalization, produced a disability at the time of discharge, or both" and 27,177 *negligent adverse events* defined as "care that fell below the standard expected of physicians in their community".[27]

The *Colorado and Utah Study* (also a U.S. based study) measured adverse events in a representative sample of Colorado and Utah hospitals using non-psychiatric 1992 discharge data. This study identified 5,614 adverse events occurring in Utah (32.6% due to negligence) and 11,578 (27.5% due to negligence) in Colorado.[28] Extrapolating the results to over 33.6 million hospital admissions in the United States in 1997, implies that at least 44,000 Americans die each year in hospitals as a result of preventable medical errors.[1]

Based on these studies, the IOM released their report estimating the number of Americans that die each year as a result of medical errors to be between 44,000 and 98,000, with an annual cost of over 9 billion dollars.[1] These figures only took into account *hospitalized patients*, which represent only a small proportion of the total population at risk, and *direct hospital costs*, which are only a fraction of total costs.[1] Prior to the IOM report the issue of medical errors was not widely acknowledged. With greater understanding of the significance of the problem, the federal government and many professional organizations have pledged to reduce the number of errors in healthcare.[32]

Some investigators have disputed the accuracy of the adverse event figures presented by the IOM, the Harvard Medical Practice Studies and the Utah and Colorado study. Several researchers have pointed out flaws in these studies.[32-34] Regardless of this debate, medical errors continue to be recognized as an important problem, worthy of addressing within the healthcare community.

2.3 DIAGNOSTIC ERRORS

The definition of *diagnostic error* used by the Australian Patient Safety Foundation, and adopted by the members of the inaugural conference on Diagnostic Error in Medicine – the first national effort to address diagnostic errors – is *"a diagnosis that was unintentionally delayed (sufficient information was available earlier), wrong (another diagnosis was made before the correct one), or missed (no diagnosis was ever made), as judged from the eventual appreciation of more definitive information"*.[35,36] Graber postulates the definition of diagnostic error as *"a diagnosis that is missed, wrong, or delayed, as detected by some subsequent definitive test or finding"*.[37]

Statistics show diagnostic errors to be a significant problem resulting in increased morbidity and mortality. In his classic studies of clinical reasoning, Elstein estimates the rate of diagnostic errors to be approximately 15%.[7] In the Harvard Medical Practice Studies, of the adverse events that resulted in serious disability, diagnostic mishaps ranked highest at 47%.[3,27] In the Utah - Colorado study, incorrect or delayed diagnosis represented 20.1% of negligent adverse events resulting in permanent disability.[28] Diagnostic errors are the second leading cause of malpractice suits against hospitals.[4] A study of autopsy findings (based on data from the U.S.) identified diagnostic discrepancies in 20% of cases including undiagnosed infections and malignant tumors as well as false-positive cancerous tumor diagnoses.[5] Roughly 5% of autopsies reveal a lethal diagnostic error where a correct diagnosis coupled with treatment could have averted death.[38] In reviewing closed malpractice claims from four liability insurers, Kachalia, et al. found that in the last decade diagnostic errors have become the most prevalent type of malpractice claims in the United States.[39] Leape, et al. estimate that 40,000 to 80,000 annual U.S. hospital deaths within the U.S. are a result of misdiagnosis.[40]

Based on a review of literature published between 1966 and 1998, Bordage classified diagnostic errors into three categories including "*data gathering errors* (from observation to findings); *data-integration errors* (from findings to diagnoses); and *situational factors.*"[6] A breakdown of each category is listed in Table 2. Based on review of one-hundred cases of suspected diagnostic errors from five large academic tertiary-care medical centers in the U.S. over a five year period, Graber, et al. classified errors by etiology as "*no-fault errors*; *system-related errors* and *cognitive errors.*"[35] A high level definition of each category is as defined in Table 3; a more detailed definition of system-related and cognitive errors is provided in Tables 4 and 5.[35] Of the one-hundred cases, 28% are solely due to cognitive factors, 19% are solely due to system-related factors, 46% involve both cognitive and system-related factors, and 7% are due to no-fault errors not related to cognitive or system-related factors (Figure 2).[35] Sixty-five percent of the cases involve system-related factors and 74% involve cognitive factors.[35] In cases where a *wrong diagnosis* occurred, cognitive factors contributed to the error in 92% of the cases, and system-related factors contributed to a wrong diagnosis in 50% of the cases.[35] In cases where a *delayed diagnosis* occurred, 36% were associated with cognitive factors; 89% were related to system-related factors.[35] The *no-fault diagnostic errors* were a result of patient related factors (missed appointments, deception, etc.) and masked disease presentation.[35] *System-related*

diagnostic errors were related to organizational, technical and equipment problems; policies and procedures; inefficient processes and difficulty with teamwork and communication.[35]

Table 2 Diagnostic Error Types [6]

Data Gathering (from observation to findings)
1. Incomplete history of present illness or history and physical
2. Ineffective questioning (interviewing)
3. Failure to gather useful information to verify diagnosis
4. Faulty detection
5. Excessive data gathering
6. Failure to validate findings with patient
7. Misidentification of symptoms or signs
8. Faulty or improper physical examination techniques
9. Failure to screen
10. Over reliance on someone else's history and physical
11. Poor etiquette leading to poor data quality
12. Misled by the way the information presented itself
Data-Integration (from findings to diagnoses)
13. Failure to consider a finding(s)
14. Over or underestimating the usefulness or meaningfulness of a finding(s)
15. Faulty context formulation
16. Faulty estimate of prevalence
17. Failure to periodically review the situation
18. Over reliance on someone else's opinion
19. Reporting findings not gathered
20. Faulty causal model: ignorance or misconceptions
21. Failure to ask for advice (consultation)
22. Failure to act sooner
23. No-fault error: atypical case, extremely rare, or rapidly evolving
Situational Factors
24. Stress
25. Fatigue, too many hours
26. Excessive workload, not enough time
27. Physician uncomfortable with own feelings toward the patient
28. Physician's mood or personality
29. Work environment: equipment, support, peer pressure, rewards and punishment

Table 3 Diagnostic Error Etiology [35]

No-fault Errors – 7% of all diagnostic errors
Masked or unusual presentation of disease
Patient-related error (uncooperative, deceptive)

System-related Errors – 19 % of all diagnostic errors
Technical failure and equipment problems
Organizational flaws
Cognitive Errors – 28% of all diagnostic errors
Faulty knowledge
Faulty data gathering
Faulty synthesis

Table 4 System-Related Aspects of Diagnostic Errors [35]

Type (number of cases)	Definition
Technical	
Technical failure and equipment problems (13)	Test instruments are faulty, miscalibrated or unavailable
Organizational	
Clustering (35)	Repeating instances of the same error type
Policy and procedures (33)	Policies that fail to account for certain conditions or that actively create error prone solutions
Inefficient processes (32)	Standardized processes resulting in unnecessary delay; absence of expedited pathways
Teamwork or communications (27)	Failure to share needed information or skills
Patient neglect (23)	Failure to provide necessary care
Management (20)	Failed oversight of system issues
Coordination of care (18)	Clumsy interactions between caregivers or sites of care; hand-off problems
Supervision (8)	Failed oversight of trainees
Expertise unavailable (8)	Required specialist not available in a timely manner
Training and orientation (7)	Clinicians not made aware of correct processes, policy or procedures
Personnel (4)	Clinician laziness, rude behavior, or known to have recurring problems with communication or teamwork
External Interference (0)	Interference with proper care by corporate or government institutions

Table 5 Cognitive Aspects of Diagnostic Error [35]

Type (number of cases)	Definition
Faulty Knowledge	
Knowledge base inadequate or defective (4)	Insufficient knowledge of relevant condition
Skills inadequate or defective (7)	Insufficient diagnostic skill for relevant condition
Faulty Data Gathering	
Ineffective, incomplete, or faulty workup (24)	Problems in organizing and coordinating patient tests and consultations

Type (number of cases)	Definition
Ineffective, incomplete, or faulty history and physical examination (10)	Failure to collect appropriate information from the initial interview and examination
Faulty test or procedure techniques (7)	Standard test / procedure is conducted incorrectly
Failure to screen (pre-hypothesis) (3)	Failure to perform indicated screening procedures
Poor etiquette leading to poor data quality (1)	Failure to collect required information owing to poor patient interaction
Faulty Synthesis: Faulty Information Processing	
Faulty context generation (26)	Lack of awareness / consideration of aspects of patient's situation that are relevant to diagnosis
Over estimating or under estimating usefulness or salience of a finding (25)	Clinician is aware of symptom but either focuses too closely on it to the exclusion of others or fails to appreciate its relevance
Faulty detection or perception (25)	Symptom, sign or finding should be noticeable, but clinician misses it
Failed heuristics (23)	Failure to apply appropriate rule of thumb, or over application of such a rule under inappropriate / atypical circumstances
Failure to act sooner (15)	Delay in appropriate data-analysis activity
Faulty triggering (14)	Clinician considers inappropriate conclusion based on current data or fails to consider conclusion reasonable from data
Misidentification of a symptom or sign (11)	One symptom is mistaken for another
Distraction by other goals or issues (10)	Other aspects of patient treatment (e.g., dealing with an earlier condition) are allowed to obscure diagnostic process for current condition
Faulty interpretation of a test result (10)	Test results are read correctly, but incorrectly conclusions are drawn
Reporting or remembering findings not gathered (0)	Symptoms or signs reported that do not exist, often findings that are typically present in the suspected illness
Faulty Synthesis: Faulty Verification	
Premature closure (39)	Failure to consider other possibilities once an initial diagnosis has been reached
Failure to order or follow up on appropriate test (18)	Clinician does not use an appropriate test to confirm a diagnosis or does not take appropriate next step after test

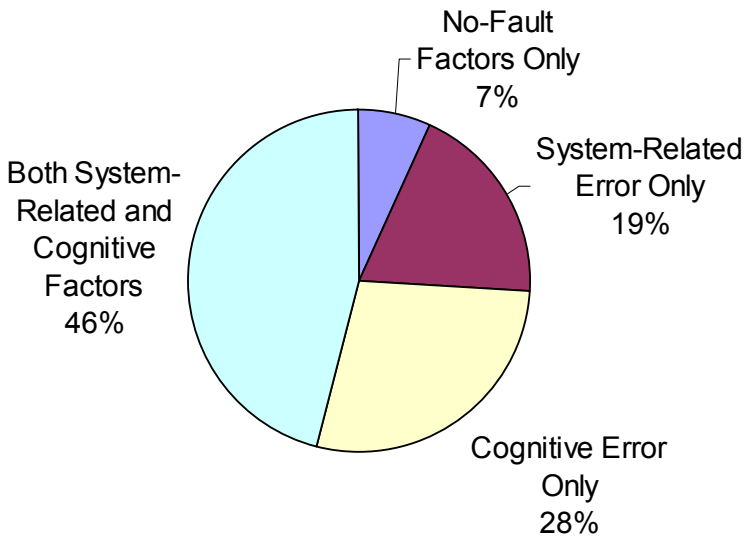


Figure 2 Factors Contributing to Diagnostic Errors [35]

Even though the diagnostic error statistics and categories provided are solely from the Graber, et al. study, numerous researchers support these findings. The categories identified by Graber, et al. are consistent with the categories derived by Bordage,[6] Chimowitz, et al.[41] and Kassirer and Kopelman.[8] The frequency of errors resulting from faulty data gathering, and/or insufficient knowledge and skills identified by Graber, et al.[35] are consistent with studies conducted by Bordage,[6] Weinber and Statson[42] and a publication over 50 years ago by Gruver and Freis.[43]

3 COGNITIVE HEURISTICS AND BIASES

People often rely on heuristics, or general rules of thumb, when confronted with a complicated judgment or decision.[44] Use of heuristics can result in "a close approximation to the optimal decision suggested by normative theories".[44] However, heuristic use can also lead to predictable biases and inconsistencies.[44] If biases are predictable, this implies that strategies can be employed to detect and correct biasing, resulting in improved judgment.

3.1 STUDY OF COGNITIVE HEURISTICS AND BIASES

Daniel Kahneman and Amos Tversky spent nearly three decades studying how people use cognitive heuristics when making judgments under conditions of uncertainty. They proposed that people rely on a limited number of heuristic principles to reduce the complex tasks of probabilistic assessment and prediction to simpler judgmental operations.[14] An advantage of using heuristics is **"**they reduce the time and effort required to make reasonable good judgments and decisions.**"** [44] A disadvantage of using heuristics is "there are certain instances in which they lead to severe and systematic errors or biases, which are deviations from normatively derived answers".[44] Kahneman and Tversky studied three cognitive heuristics – *Availability*, *Anchoring and Adjustment*, and *Representativeness* – identifying common ways in which people misuse them, resulting in less than optimal decisions.

The *Availability* heuristic is "a rule of thumb in which decision makers assess the frequency of a class, or the probability of an event, by the ease with which instances or occurrences can be brought to mind".[14] *Availability* is used to estimate the frequency and probability of an event by how easy it is to remember that event. Using *Availability* simplifies

what could be a difficult judgment by exploiting the phenomenon that common events are easier to remember and uncommon events are more difficult to remember. However, *Availability*, results in a systematic bias "when events are more available than others not because they tend to occur frequently or with high probability, but because they are inherently easier to think about because they have taken place recently, because they are highly emotional, etc."[44] For example, when asked if more people die from being murdered or from being involved in an automobile accident, most people would answer that more people die from being murdered because the media tends to report murders at a greater frequency than it reports automobile accidents. According to Kahneman and Tversky, "these kinds of statistics are counterintuitive because most people estimate the frequency of an event by how easy it is to bring instances of the event to mind".[14] "*Availability* can lead to biased judgments when examples of one event are inherently more difficult to generate than examples of another".[44] When assessing *Availability*, Kahneman and Tversky presented subjects with the following scenario: "In a typical sample of text in the English language, is it more likely that a word starts with the letter K or that K is the third letter (not counting three letter words)."[45] Out of 152 subjects, 105 thought words starting with K were more probable.[45] Actually, there are approximately twice as many words with K as the third letter than words starting with K.[45] People tend to overestimate the relative frequency of words starting with K since it is easier to generate words in this format.[44]

The *Representativeness* heuristic describes the judgment of probabilities "by the degree to which A is representative of B, that is, the degree to which A resembles B".[14] A common technique Kahneman and Tversky used to assess *Representativeness* was to provide a description of an individual and ask subjects to select the most likely classification of the individual. For example, "Linda is a 31 year old single female, outspoken and very bright and majored in philosophy. As a student she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations. Is Linda most likely (a) a bank teller or (b) a bank teller that is active in the feminist movement".[46] In a Kahneman and Tversky study, 9 out of 10 respondents thought Linda was more likely a bank teller that is active in the feminist movement than just a bank teller.[46] This response is an example of the common *Representativeness* bias that Kahneman and Tversky called *conjunction fallacy* in which "the conjunction or co-occurrence of two events cannot be more likely than the probability of either event alone".[47] Other common biases associated with *Representativeness* include (1) the *law of*

small numbers which is the belief that random samples of a population will resemble each other more closely related than statistical sampling theory would predict, (2) the *gambler's fallacy* which is the belief that a series of independent trials with the same outcome will soon be followed by an opposite outcome, (3) the tendency to *neglect base rate information*; and (4) the *tendency to overlook regression* by neglecting the diagnosticity of information from which people base their predictions.[44] In order to make better judgments when using *Representativeness*, one should "not be misled by highly detailed scenarios, pay attention to base rate information, remember that chance is not self-correcting, and don't misinterpret regression toward the mean."[44]

Finally, *Anchoring and Adjustment* is "the insufficient adjustment up or down from an original starting value, or anchor."[44] To determine if humans commit *Anchoring and Adjustment* they had subjects spin a wheel with numbers around the wheel. Once the wheel stopped, they asked subjects if the percentage of African countries in the United Nations is more or less than the number the wheel landed on. When the wheel landed on 65, subjects gave a median estimate of 45%. When the wheel landed on 10, subjects gave a median estimate of 25%. Subjects insufficiently adjusted up or down from the anchor, or original starting value.[44] Common biases of *Anchoring and Adjustment* are (1) *insufficient adjustment*; (2) biases in the *evaluation of conjunctive events* (drawing a red marble seven times in succession, with replacement, from a bag containing 90% red marbles and 10% white marbles) and *biases in the evaluation of disjunctive events* (drawing a red marble at least once in seven successive tries, with replacement, from a bag containing 10% red and 90% white marbles) in which people tend to overestimate the probability of conjunctive events and underestimate the probability of disjunctive events.[44]

3.2 APPLICATION OF HEURISTICS TO DIAGNOSTIC REASONING

Many researchers have investigated physicians' behavior during diagnosis to determine if they utilize cognitive heuristics and biases. Findings show that physicians do use rules-of-thumb, or

heuristics, while gathering and interpreting information during the diagnostic process, and are vulnerable to cognitive biases.[2, 9, 16, 18, 21-24, 48-53, 57-59, 62-74]

Using the Inventory of Cognitive Biases in Medicine (ICBM), a series of twenty-two medical scenarios developed from actual clinical experiences where answer choices represent either statistically based decisions or cognitive biases, Hershberger, et al., found that practicing physicians demonstrated considerable susceptibility to cognitive biases varying by medical specialty.[17,18] Practicing physicians were slightly less impacted by cognitive biases, but susceptibility to cognitive biases was observed among both novices and experts.[17]

During each intermediate step of diagnostic reasoning there is the potential for physicians to use cognitive heuristics and biases (Table 6). During **hypothesis generation** when diagnoses are generated, physicians are susceptible to biases based on *Representativeness* and *Availability*. *Representativeness* can be used to determine how closely a patient's findings resemble the prototypical manifestations of diseases.[56] Use of such pattern-recognition methods can lead to errors when the physician does not consider atypical representations.[56] *Availability* occurs in diagnostic reasoning when a diagnosis is triggered by recent cases similar to the current case. Diagnostic errors can occur if a diagnosis is made because it readily comes to mind, but does not completely fit the current case. Conversely, if a disease has not been seen for a long time, it may be less available.[56] A misdiagnosis can occur if the physician assumes this patient cannot possibly have the same diagnosis as the last three patients they have seen (gambler's fallacy).[56]

A number of cognitive biases such as *Confirmation Bias*, *Anchoring and Adjustment*, *Search Satisficing, Premature Closure* and *Overconfidence* bias can prompt clinicians to make errors when **pruning, selecting and/or validating a diagnosis**.[56] *Search Satisficing,* or calling off a search once something is found, may occur when a physician arrives at an initial diagnostic hypothesis based on the review of only a portion of the clinical data available, and does not review additional clinical data once their initial diagnosis has been specified. *Premature Closure* is when a physician accepts a diagnosis before it has been fully verified. *Confirmation Bias* is the tendency to look for confirming evidence to support a diagnosis rather than look for disconfirming evidence to refute it even when the latter is persuasive and definitive.[56] When a physician does not review additional data or order additional tests because they are confident in their diagnosis, they may be committing the *Overconfidence* bias, which is a "tendency to act on

incomplete information, intuitions or hunches; or when too much faith is placed in an opinion instead of carefully gathered evidence".[56]

When **selecting a course of action**, the *Omission Bias* and *Outcome Bias* can adversely influence treatment decisions if the physician focuses too heavily on what *could happen*, rather than what is most *likely to happen* once a treatment or therapy is initiated.[53] Physicians can underutilize preventive interventions in order to avoid having a direct role in bad outcomes.[51,53,55] Death by natural causes can be viewed as better than death by prescription.[53] Physicians may rather risk a patient dying because nothing was done than have to bear the responsibility of a patient dying because of a treatment they prescribed. *Outcome bias* is when a physician places too much emphasis on patient outcomes, and does not consider the rationale and evidence underlying medical decisions.[53, 58]

Table 6 Heuristics and Biases within Diagnostic Reasoning [53]

Diagnostic Step	Heuristic / Bias *	Definition
Generating Differential Diagnosis (generating hypothesis)	Availability [2, 21,23, 48, 49, 51, 53, 61, 63, 67] (Recall)	"Differential influenced by what is easily recalled, creating a false sense of prevalence"[53]
	Representativeness [2, 48, 49, 51, 53, 63] (Judging by similarity) Base Rate Neglect, Insensitivity to Sample Size[19, 60, 61]	"Representativeness drives the diagnostician toward looking for prototypical manifestations of disease…Restraining decision-making along these pattern-recognition lines leads to atypical variants being missed".[56] "Clinical suspicion influenced solely by signs and symptoms, and neglects prevalence of competing diagnosis" [53]
	Gambler's Fallacy [48, 61] (Monte Carlo Fallacy, Law of Averages, Sequence Effect)	"The belief that the next toss of the coin cannot possibly be the same outcome as it has been in the last 6 tosses. That is, a patient that presents with the same symptoms as the last 5 patients cannot possibly have the same disease – the physician doubts that the sequence can continue to be the same as previous patients" [53]
	Framing Effect [23, 60]	"How diagnosticians see things may be strongly influenced by the way in which the problem is framed"[56]
Pruning Differential Selecting Diagnosis Validating Diagnosis	Confirmation Bias [2, 9, 48, 51,53, 63, 64,74] (Pseudo-diagnosticity)	"The tendency to look for confirming evidence to support a diagnosis rather than look for disconfirming evidence to refute it, despite the latter often being more than persuasive and definitive" [56]
	Anchoring and Adjustment [2, 16, 23, 24, 48, 51,53,66,68, 69, 70]	"Inadequate adjustment of differential in light of new data results in a final diagnosis unduly influenced by starting point" [53] "The tendency to perceptually lock onto salient features in the patient's initial presentation

Diagnostic Step	Heuristic / Bias *	Definition
		too early in the diagnostic process, and failing to adjust this initial impression in the light of later information"[56]
	Search Satisficing [2, 48, 53] Premature Closure [2, 23, 65]	"Clinician stops search for additional diagnoses after anticipated diagnosis made" [53]
	Overconfidence[2, 16, 48, 57, 63, 71]	"When one puts too much faith in their diagnosis and does not gather sufficient information to confirm the diagnosis. Those who are overconfident spend insufficient time accumulating evidence and synthesizing it before action. They tend to act on incomplete information and hunches" [53]
Selecting a Course of Action	Outcome Bias [48, 51, 53, 58]	"Clinical decision is judged on the outcome rather than the logic and evidence supporting the decision" [53]
	Omission Bias [22, 48, 51, 53, 65]	"Undue emphasis on avoiding adverse effect of a therapy results in under-utilization of beneficial treatment" [53]
After receiving feedback	Hindsight Bias[59, 61, 71, 72]	"When knowing the outcome profoundly influences the perception of past events and prevents a realistic appraisal of what actually occurred"[56]

* Terms used synonymously in literature are shown in parentheses

There are those that have classified heuristics and biases in terms of the errors that may result when using various heuristics and biases; errors referred to as cognitive dispositions to respond.[222] Table 7 classifies heuristics in terms of their associated cognitive disposition to respond. Definitions of each heuristic can be found in Campbell, et al.[222]

Table 7 Classification Scheme for Cognitive Dispositions to Respond [222]

Cognitive Disposition to Respond	Heuristic / Bias
Error of over attachment to a particular diagnosis	Anchoring Confirmation Bias Premature Closure
Error due to failure to consider alternative diagnoses	Multiple alternatives bias Representativeness Restraint Search satisficing Sutton's slip Unpacking principle Vertical line failure
Error due to inheriting someone else's thinking	Triage cueing Diagnosis momentum Framing effect Ascertainment effect

Cognitive Disposition to Respond	Heuristic / Bias
Errors in prevalence perception or estimation	Availability bias Base-rate neglect Gambler's fallacy Hindsight bias Playing the odds Posterior probability error Order effects
Errors involving patient characteristics or presentation context	Fundamental attribution error Gender bias Psych out error Yin-yang out
Errors associated with physician affect or personality	Commission bias Omission bias Outcome bias Visceral bias Over and under-confidence Belief bias Ego bias Sunk costs Zebra retreat

The use of cognitive heuristics is widely accepted in the medical decision making community. However, further investigation is required to determine the impact of cognitive heuristics and biases on diagnostic errors. One objective of this dissertation project is to determine how frequently a diagnostic error is due to use of the cognitive heuristics *Anchoring and Adjustment* and *Confirmation Bias*.

4 THEORIES OF DECISION MAKING

Previous research interventions attempting to debias individuals used techniques based on normative theories of decision making. This research differs from that previous work by attempting to debias individuals using techniques based on a descriptive theory of decision making. In this chapter, I will outline the differences between the normative and descriptive approaches and describe specific theories in both categories of decision-making.

4.1 NORMATIVE AND DESCRIPTIVE DECISION THEORY

The study of decision theory began in the 20th century and has progressed to its current state with contributions from economists, statisticians, psychologists, political and social scientists and philosophers.[197] There are two primary categories of decision theory - normative and descriptive theories of decision-making. **Normative Decision Theory** is about how decisions *should* be made, and **Descriptive Decision Theory** is about how decisions are *actually* made.[197]

Normative decision theory is concerned with identifying the optimal decision to make, assuming an ideal decision maker who is fully informed and fully rational, and is able to process the information with perfect accuracy.[198] The practical application of how people make decisions is called decision analysis and is aimed at finding tools, methodologies and software, or decision support tools, to help people make optimal decisions.[199] Since individuals do not always behave in optimal ways, researchers began the study of how people actually behave (descriptive decision theories). These two areas of study are closely linked due to normative theory researchers creating hypotheses that are tested against actual human behavior. The Expected Utility Theory and Bayes' Theorem are normative theories which are discussed in the next section. Descriptive

theories of decision making, discussed in section 4.3, include the Mental Model Theory and the Dual Process Theory. These theories have been selected since they are well developed and discussed theories related to decision-making.

4.2 NORMATIVE THEORIES OF DECISION MAKING

4.2.1 Expected Utility Theory

The Expected Utility Theory (EUT) was proposed as a theory of behavior by John von Newmann and Oskar Morgenstern in 1947.[76] This classical utility theory was intended to describe how people *would* behave if they followed certain requirements of rational decision-making, not how they *actually* behave.[76] This theory is based on an explicit set of principles or axioms that underlie rational decision-making including alternative ordering (preferring one alternative over another); alternative dominance (alternatives are weakly dominant, strongly dominant, etc.); cancelation (equal alternatives cancel each other out); transitivity (if you prefer A to B and B to C, then you prefer A to C); continuity (the decision maker should prefer a gamble between the best and worst outcome to a sure intermediate outcome if the odds of the best outcome are good enough); and invariance (the decision maker should not be affected by the way alternatives are presented).[44] It has been mathematically proven that when decision makers violate principles such as these, expected utility is not maximized.[76]

4.2.1.1 Application of Expected Utility Theory to Diagnostic Reasoning

The practical application of the EUT when determining the best option for patients is highly debated. Some feel decision trees and the calculation of expected utilities are not clinically useful,[78,181] and others view these techniques as extremely helpful remedies for the complex medical situations physicians face.[77] A major impediment of EUT may be that it requires knowledge of highly specific conditional probabilities, when little or no objective data actually exists.[78] When clinicians substitute inaccurate estimates of prior and conditional probabilities,

the resulting posterior probabilities will be invalid.[78] In a research study conducted by Eddy in which he assessed the use of probabilistic reasoning in clinical medicine, he concluded that in everyday clinical practice physicians rarely use mathematical probabilities or compute expected utilities during reasoning.[79]

The manner in which likelihoods are mentally represented may also contribute to the mismatch between observed behavior and EUT. People may have knowledge of likelihoods, but without representing them numerically.[50,78] For example, Kuipers et al. found that subjective probabilities or likelihood values are stored as symbolic descriptions of numbers which are expressed as categorical or ordinal relations with bounding values.[78]

It is also believed that people do not utilize decision trees and expected utilities in decision-making because we have a limited supply of working memory, which is a "system for temporarily storing and managing the information required to carry out complex cognitive tasks such as learning, reasoning, and comprehension".[198] A concept developed by James March and Herb Simon called ***Bounded Rationality*** states that deviations from the optimal are a result of people not having the capacity to compute optimal solutions due to our working memory imposing limits on how much information we can use.[80]

4.2.2 Bayes' Theorem

When a preliminary hypothesis is followed by new information, judgments need to be updated based on the new information. For example, when a physician makes an initial diagnosis then receives new clinical information, reassessment of the initial diagnosis must occur. **Bayes' Theorem** can be used to arrive at an updated diagnosis or treatment decision based on newly available data.[53,81]

The probabilistic approach to diagnostic reasoning based on Bayes' Theorem involves three steps: 1. Assigning a probability to each diagnosis or treatment being considered – these are known as *prior probabilities*; 2. Identifying and collecting information for testing of competing hypotheses; 3. Calculation of post-test or *posterior probabilities* for each diagnosis or treatment under consideration given a new finding or test result.[53] This process continues serially until a diagnosis or treatment decision has been reached.[53] The normative procedure of Bayes' Theorem

used to update probabilities is the most general rule for updating beliefs about a hypothesis given new evidence.[75]

4.2.2.1 Application of Bayes' Theorem to Diagnostic Reasoning

There has been much debate regarding the feasibility of using Bayes' Theorem in the clinical setting. Many of the beliefs regarding use of the EUT apply equally to the use of Bayes' Theorem. Kempainen et al. state that even though Bayes' Theorem represents "the idealized means of moving from clinical uncertainty to a definitive diagnosis or optimal treatment, its application is fraught with difficulty [in that] Bayes' Theorem requires specific information such as accurate estimates of prior probabilities that may not be available".[53] Others indicate that the problem with Bayes' Theorem is that the sensitivity and specificity of diagnostic tests vary in different patient populations, making it difficult to use the theorem.[82,83]

These concerns are shared by Graber who's opinion is that "despite the allure of normative approaches, such approaches are infrequently used in clinical practice because the approach is complex, [that is] using a normative approach requires definitive data on the base rates of disease in our population of patients, and the characteristics of every diagnostic test that could be applied, data that is typically not available."[84] Eddy believes that "the evidence shows that physicians do not manage uncertainty very well; many physicians make major errors in probabilistic reasoning; and that these errors threaten the quality of medical care"[79]. He studied the use of probabilistic reasoning when analyzing how physicians process information related to mammographic diagnose of breast cancer. When given all the information required to compute the probability of breast cancer for a given case, 95% of the practicing physicians (in a context where these physicians deal with this type of judgment on a daily basis) estimated the probability of a positive mammographic test result to be 75% (the actual probability was 7.5%, that is, physicians extremely over-estimated the actual probability.[79] In a study conducted by Poses et al., in which they sought to determine physicians' ability to provide accurate probability estimates required when using Bayes' Theorem, they prospectively obtained estimates of the probability of streptococcal pharyngitis from ten experienced physicians to determine the accuracy of these unaided probability estimates.[85] With only clinical data to assess (the results of throat cultures were not given to the physician), it was found that physicians overestimated the rate of positive cultures for sore throats for 81% of the patients.[85] Another study conducted by

Christensen-Szalanski and Bushyhead found that physicians overestimated the probability of pneumonia when assessing patient cases, but were sensitive to the predictive value of symptoms and appeared to use base-rate information correctly when making clinical judgments.[93]

4.3 DESCRIPTIVE THEORIES OF DECISION MAKING

4.3.1 Mental Model Theory

The *Mental Model Theory* (MMT) asserts that people construct mental representations of a situation and apply a form of logic to them to draw conclusions regarding the situation. Here, I describe principles and predictions of the theory.

4.3.1.1 Principles of the Mental Model Theory of Deductive Reasoning

In describing the Mental Model Theory of Deductive Reasoning, Philip Johnson-Laird states:

> "According to the mental model theory of deductive reasoning, reasoners use the meanings of assertions together with general knowledge to construct mental models of the possibilities compatible with the premises. Each model represents what is true in a possibility. A conclusion is held to be valid if it holds in all models of the premises. Evidence shows that the fewer models an inference calls for, the easier the inference is. Errors arise because reasoners fail to consider all possible models, and because models do not normally represent what is false, even though reasoners can construct counterexamples to refute invalid conclusions".[94]

As people make judgments, they reason. *Reasoning* is the deduction of inferences from premises which are accepted as true for the circumstance in which they are stated. *Deductive reasoning* is reasoning from the general to the specific or from cause to effect. When deductive reasoning is performed, one takes a general rule, or premise, and deduces a particular, or specific, conclusion.

The reasoning process starts with *premises*, which are declarative statements, propositions, perceptions or beliefs.[94] When reasoning is applied to the premises, ideally, it yields a valid conclusion that is not explicit in the premises.[94] The intervening processes that

occur during reasoning are mysterious.[94] Some theorists presume that the human mind constructs logical syntactic representations of assertions, then the mind applies rules of formal logic to these representations to arrive at a conclusion.[94,96,97] Whereas, others believe people apply logic to assertions to derive conclusions, relying on their general knowledge, understanding of the meaning of the premises, and similar principles.[98] Johnson-Laird, Byrne and Bara believe that when people reason, they construct mental models of the premises representing the situation, and draw conclusions from these models[94,99]

The idea that thinking depends on mental models originated from a Scottish psychologist named Kenneth Craik, who suggested that "our perception constructs 'small-scale models' of reality that are used to anticipate events and to reason".[94,101] In the fifth chapter of Craik's book *The Nature of Explanation*, he wrote:

> "If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and the future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it".[101]

The Mental Model Theory of deductive reasoning is based on three principles, which distinguish models from syntactic representations of logical form, semantic networks, and other sorts of proposed mental representations.[94] The principles of the MMT include the following:

Principle 1. Each Mental Model Represents a Possibility. Each mental model represents one and only one possibility and captures what is common to the various ways that the specific possibility might occur.[94] This may sound contradictory – it represents only one possibility, but yet it captures all ways the possibility can occur. A representation of a situation may result in the construction of multiple mental models. This is best explained by an example. The exclusive disjunction *'Either TCE is in the river or else it doesn't come from the river'* follows the form A or else B but not both. From this disjunction two mental models are constructed to represent both possibilities: the possibility "TCE is in the river" can be represented as a mental model in the form "*TCE-in-river"*. The possibility "it doesn't come from the river" can be represented as a mental model in the form "¬ *TCE-comes-from-river"* (¬ denoting negation).[94] Mental models

are not necessarily made up of words; in this case they represent relations between the TCE and the river.

Principle 2. Principle of Truth. "Mental models represent what is true according to the premises, but by default do not represent what is false".[94] According to the principle of truth, mental models represent only the possibilities that are true given a premise, and represent a clause in the premises only when it is true in the possibility.[94] Given the exclusive disjunction "not A or else B", the mental model that corresponds with 'not A' is '¬A', and the model that corresponds to 'else B' is simply 'B'.[103] The first mental model '¬A' does not represent 'B' which is false in this possibility. The second model 'B' does not represent 'not A' which is false in this possibility.[103] People tend to not consider what is false in each possibility (model); they tend to consider only what is true. Using the disjunction example presented in the first principle, 'Either TCE is in the river or else it doesn't come from the river', the first premise 'TCE is in the river' is represented by the model '*TCE-in-river*' - this model does not represent explicitly that in this possibility it is <u>false</u> that TCE does <u>not</u> come from the river[94] – since this is essentially a double negative, it converts to 'TCE comes from the river'.

Even though the principle of truth proposes that by default individuals do not represent what is false, there are exceptions that overrule the principle. Individuals can make *mental footnotes* about the falseness of the premises and if they retain these footnotes within their minds they can construct *fully explicit mental models* which represent the premises even when they are false.[94] Even though individuals can derive fully explicit models, representing only what is true is the norm.[94] Mentally it is easier when reasoners do not have to bother with what is false.[94]

Principle 3. Reasoning with Mental Models. Deductive reasoning depends on the mental models that have been constructed.[94] When reasoning with mental models, there are four possible outcomes[94]: 1. If a conclusion holds in all the models of the premises, that is, it has no counter-examples, it is *necessary* given the premises; 2. If it holds in a proportion of models, its *probability is equal to that proportion*, given that the models represent equi-probable alternatives; 3. If it holds in at least one model, it is *possible* given the premises; and 4. If it does not hold in any of the models, it is *impossible* given the premises. To illustrate, consider the following:

Type of Premise	Premise	Mental Models Yielded from Premise
Disjunction	Either TCE is in the river or else it doesn't come from the river	TCE-in-river ¬TCE-come-from-river
Categorical Assertion	TCE does come from the river	TCE-come-from-river

As already established, the disjunction yields two mental models *TCE-in-river* and ¬*TCE-come-from-river*. The categorical assertion yields one mental model *TCE-come-from-river*. Combining the model of the categorical assertion with each of the disjunction models, results in the second disjunction model ¬*TCE-come-from-river* being eliminated since the two models are contradictory. The only model remaining is the first disjunction model *TCE-in-river*. Therefore, the conclusion is TCE is in the river.

4.3.1.2 Predictions of the Mental Model Theory

Many experiments have corroborated the Mental Model Theory. A complete bibliography can be found on Ruth Byrne's website (http://www.tcd.ie/Psychology/other/Ruth_Byrne/mental_models/index.html). The MMT yields some critical predictions as follows: [94]

Prediction 1. One model is better than many. "The fewer the number of models needed for an inference, and the simpler the models, the easier the inference should be". [94] With fewer and simpler models, inferences should take less time and reasoning should be less prone to error.[94] This prediction of the MMT is a consequence of the limitations of our working memory,[94,104] because multiple models can overload the processing capacity of working memory and lead to errors due to reasoners failing to consider all the models of the premises.[94] Several researchers have investigated the impact of multiple mental models on reasoning. Halford et al. has shown that the fewer the entities in a model of a relation, the easier it is for individuals to make inferences.[105] Schaeken et al.[106], Carreiras and Santamaria[107], and Knauff et al.[108] have all obtained comparable findings when investigating problems based on temporal relations. Vandierendonck et al. have shown the same effect in temporal and spatial reasoning.[109]

Prediction 2: Reasoners make errors due to failing to consider all mental models. When a premise, or set of premises, should result in the construction of multiple mental models, an individual is likely to not construct all the applicable models because they fail to foresee all the

models.[110] This in turn will cause them to draw a conclusion corresponding to only some of the models of the premises; they will not consider all the applicable models.[111] For example, when given the syllogism '*None of the As are a B*' and '*All the Bs are Cs*', the conclusion an individual might draw is that '*None of the As is a* C'. This conclusion is compatible with only one model of the premises of the syllogism.[112] The individual that has drawn this conclusion has failed to realize that the Cs that are not Bs could be As - as a result they do not draw the valid conclusion relating As and Cs which is '*Some of the Cs are not As*'.[94]

Just as with the previous prediction, it has been proposed that due to the limitations of working memory, reasoners construct as few models as possible – often times only a single model is constructed.[94] Ormerod and his colleagues proposed that reasoners construct the minimal set of models needed to infer a conclusion – they call this *minimal completion,* i.e. they "construct only the minimally necessary models".[113] Sloutsky et al., have observed that reasoners commit reasoning errors due to basing conclusions on a single model of the premises because they fail to foresee and/or consider all the applicable models.[114]

Prediction 3: When falsity matters, fallacies occur. Reasoners focus on what is true and neglect what is false.[94] When reasoners do not consider what is false, reasoning errors occur. A classic method used to confirm this prediction is the Wason Selection Task where a participant is given four cards that have a letter or number that is visible to the participant - cards such as A 2 B 3. The participants are told that each card has a number on one side and a letter on the other side. The participant has to choose which card to turn over in order to determine if the following rule is true or false: *If a card has a 'A' on one side, then it has a '2' on the other side*. Most people think about the situation when the conditional is true (A on front, 2 on back) and indicate that the 'A' card and the '2' card should be turned over to prove the rule. In order for the participant to make the proper selection of the card to turn over, they need to consider the counter-example to the conditional; that is they need to construct the A ¬2 mental model and select the 'A' and '3' cards. According to the MMT, people reason through the Wason Selection Task by only constructing a mental model regarding what is true, and do not construct mental models on what is false (especially if they lack cognitive ability).[115] It has been shown that "any manipulation that helps the participant to consider counter-examples, and to match them to the corresponding cards, should improve performance" in this and similar tasks.[116]

Prediction 4: Content and background knowledge modulate the interpretation of assertions and the process of reasoning. Since meaning is central to interpreting premises and the construction of mental models, the subject matter (content) of inferences and background knowledge can affect reasoning by influencing the interpretation of the premises.[94] For example, the following inference is valid in form: '*Eva is in Rio or she's in Brazil; She's not in Brazil; therefore, she's in Rio'.* However, it is impossible for Eva to be in Rio without being in Brazil since Rio is in Brazil. Since Rio is not in Norway, the following inference is easy *'Eva is in Rio or else she's in Norway; She's not in Norway; therefore, she's in Rio'*. In order to make the proper conclusion in the Rio / Brazil example, one has to know that Rio is in Brazil. If they do not know this, their conclusion will be incorrect. Knowledge and beliefs also influence the process of reasoning in that individuals search harder for counter-examples to conclusions that violate their knowledge.[94]

Prediction 5: With experience, reasoners develop tailor-made strategies for particular sorts of problems. When individuals carry out a series of inferences, they develop strategies for coping with them.[94,117] Johnson-Laird states that "an earlier version of the Mental Model Theory implied that reasoners start reasoning with the most informative premise - but this claim is not always true - reasoners' strategies determine which premise they take into account first".[94] Some reasoners develop a strategy that is based on suppositions, i.e. reasoners might say 'suppose … it then follows that …'. Another strategy is to make an inference from a pair of premises, and then to make another inference from its conclusion and a third premise. Another strategy is to write down the possibilities compatible with premises and work through them in the order they are stated. A final strategy is to search for counter-examples for the models under consideration.

4.3.1.3 Application of Mental Model Theory to Diagnostic Reasoning

The *Mental Model Theory* (MMT) asserts that people construct mental representations of a situation and apply a form of logic to them to draw conclusions regarding the situation. Given that physicians make diagnoses at the bedside without referencing documented facts and evidence, it is a logical conclusion that they store disease information within their minds in some form. There has been much debate regarding the representation of medical knowledge within the minds of physicians. However, there is agreement that physicians store disease information in

some form, reason with the mental representations, and draw a conclusion when a diagnosis is made. Evidence exists that as the transition from novice to expert occurs, physicians progress through multiple transitional stages of knowledge structures, for example, from elaborated causal networks, abridged networks, illness scripts and instance scripts.[182,183,185] According to researchers that have studied this transition, "these representations do not decay or become inert in the course of developing expertise, but rather remain available for future use when the situation requires their activation".[183]

As medical students train in the classroom, they develop ***elaborated causal networks*** that contain disease causes and consequences in terms of the underlying pathophysiological processes.[183] These causal networks, a kind of *propositional network*, designate how objects, represented as nodes, are related to each other; the relations between objects are represented as links. Schmidt et al., using the think-aloud technique, suggested that during the first four years of medical training, students develop causal networks that contain disease causes and consequences, and that the networks become increasingly complex and elaborate as a result of learning.[185]

As students have the opportunity to apply their acquired knowledge through exposure to patients, the propositional networks are converted into high-level, simplified causal models, or **abridged networks**, that include disease signs and symptoms as encountered in real patients. As students see more and more patients, they begin to use shortcuts in their reasoning process, and transform the information in their extensive causal networks so that the information is efficiently accessible. Schmidt et al. noticed a significant difference in the propositional networks used by students that were in this stage of their careers.[185] Networks were less complex among third and forth year medical students who were seeing actual patients. In verbalizing their thoughts, students did not fully explain the entire pathophysiological processes, but immediately honed in on the critical aspects of the case that were pertinent to make the diagnosis.[185]

As students compile information in their networks, they transition from a causal type of knowledge organization to list-like structures called ***illness scripts***.[183] As students begin to see patients with the same or comparable symptoms, "their extensive pathophysiological networks convert to diagnostic labels or simplified mental models that sufficiently explain the phenomena observed".[183] Instead of causal processes, they begin to concentrate on the different features that characterize the clinical appearance of a disease which becomes their anchor points – the center

of their thought processes. Illness scripts, a concept adapted from Feltovich and Barrows,[186] contain items such as enabling conditions, predisposing factors, boundary conditions, faults and consequences, all of which are organized into a serial structure, appearing in the order a physician informs other physicians about the patient's condition.[183] Claessen and Boshuizen confirmed that information can be stored in this serial order within memory.[187] Groen and Patel,[188] Coughlin and Patel,[189] Norman et al.,[190] Schmidt et al.,[191] and Hobus et al.[192] reported similar findings regarding illness scripts.

As expertise increases, clinicians store patient encounters as ***instance scripts,*** individual instances that are not merged into a prototypical form.[183,193] Hassebrock and Pretula[194] and VanRossum and Bender[195] showed evidence of the ability of clinicians to remember specifics of cases they assessed up to 20 years ago. Allen, Brooks and Norman also described the use of instance scripts by experts.[196]

Even though the four mental representations described can be seen as a developmental sequence corresponding to the stages of education, previously acquired knowledge structures remain available allowing the expert clinician to move from one representation to another as required by the complexity of the case.[183]

4.3.2 Dual Process Theory

The Dual Process Theory (DPT) states that there are two fundamental approaches to reasoning - *Intuitive,* or heuristic reasoning and *Analytical*, or systematic reasoning.[152] Theoretical work and empirical research on the DPT has been underway for over a quarter of a century.[153] Richard Shiffrin has been studying dual processing for over 30 years. In his early work with Atkinson he detailed the role of controlled processing in studies of short-term memory and verbal learning.[154] In the early 1970s Shiffrin and Gardner, and Shiffrin, McKay and Shaffer performed attention studies indicating that multiple channels could be processed in parallel, a result that contradicted much of the research being conducted at that time.[153,155,156] From these findings, Shiffrin and Schneider published a set of papers detailing automatic and controlled processes – the dual processes used in human information processing.[157,158] These papers have become the theoretical and empirical basis for much of the work on automaticity that has been conducted in subsequent

decades - including over 4,800 citations of this work.[153,157, 158] Many other researchers have devoted extensive effort to develop an empirical and theoretical understanding of the dual modes of processing including Anderson in his study of automaticity and the ACT theory,[159] Logan's research on attention and automaticity in priming tasks,[160] Pashler, et al. in their study of attention and performance,[161] and Stanovich's work on the impact of the automaticity theory,[162] to name a few.

The dual process theory of human information processing has been studied in many disciplines including cognitive psychology and social psychology where researchers have studied higher cognitive processes such as thinking, reasoning, decision-making, social judgment, learning and memory, philosophy of the mind, and evolutionary psychology.[163] A common theme of the multiple dual-process theories is that there are two different modes of processing - processes that are *unconscious, rapid, automatic, and high capacity* versus those that are *conscious, slow and deliberative*.[163] The main characteristics of the two modes of thinking (often referred to as the two types or systems of the DPT) are listed in Table 8.[152] In this dissertation, I will refer to the two modes of thinking as 'Intuitive' (System 1) and 'Analytical' (System 2).

Table 8 DPT System 1 and 2 Characteristics [152]

Cognitive Style	**System 1 Heuristic, Intuitive**	**System 2 Systematic, Analytic**
Computational principle	Associative	Rule-based
Responsiveness	Passive	Active
Capacity	High	Limited
Cognitive awareness / control	Low	High
Automaticity	High	Low
Rate	Fast	Slow
Reliability	Low	High
Errors	Relatively common	Rare
Effort	Low	High
Emotional attachment	High	Low
Scientific rigor	Low	High

4.3.2.1 Intuitive Mode of Thinking

As described by Shiffrin and Schneider in 1977, intuitive processes have the following properties:[158]

- People's short-term working memory capacity does not impact our intuitive processes
- Intuitive processes are unconscious, do not require our attention, and appear to act in parallel and independent of each other
- Our intuitive processes may be initiated under our control, but once initiated they run to completion automatically
- Intuitive processes require considerable training to develop, and once learned, are difficult to modify
- The speed and automaticity of intuitive processes usually keep their components hidden from conscious perception
- Intuitive processes can indirectly affect learning through forced allocation of analytical processing
- Executing a task repeatedly results in learning the sequence; which leads to performing the task in an intuitive manner (tying one's shoes)

Intuitive processes are "rapid, contextual, holistic and unavailable for introspection; commonly associated with visual perception that enables rapid recognition and categorization of objects, but …is not limited to visual perception".[164] It has been shown in numerous experiments that once people acquire implicit knowledge they execute intuitive processes without being able to verbalize the explicit rules they used to accomplish the task.[163] Intuitive processes are based on prior experiences and are initiated because experienced decision makers recognize overall patterns (Gestalt effects) in the information presented, and act accordingly.[152,164-166] When in the intuitive mode of thinking, we typically use heuristics or mental shortcuts, and judgments are often made by relying on our instinctive first impressions.[50,152,167] This mode of thinking "proves to be **effective much of the time**, is highly context-bound, with the potential for ambient conditions to expert powerful influence over it, and ***occasionally*** **fails** (sometimes catastrophically)."[152] Even with the occasional error, there are many benefits to operating within the intuitive mode of thinking. Without this mode of thinking we would have to methodically think through, and possibly relearn, a task every time it was performed. The intuitive mode of

thinking allows us to acquire skills that are used as a basis for learning more complex tasks. Once the skills are learned, they become automatic and are performed without conscious thought. This frees our minds and permits us to focus our attention on tasks such as problem-solving. There are two major sources of input to intuition including our innate behavior that is 'hard-wired', and repetitive processing that takes place in the analytic mode. The analytical mode of thinking also plays a key role in the acquisition of skills and expertise. This mode is described further in the next section.

4.3.2.2 Analytical Mode of Thinking

As described by Shiffrin and Schneider in 1977, control processes (the term they use for analytical processes) have the following properties:[158]

- Analytical processes are 'limited-capacity processes' requiring our conscious attention. This prevents us from simultaneously processing multiple thoughts, causing us to process thoughts linearly
- Analytical processes are processed in this manner due the limitations of working memory. We can only maintain a limited amount of information in working memory without losing it unless it is utilized
- Analytical processes can be adopted quickly without extensive training and modified fairly easily
- Analytical processes control the flow of information between short-term memory, working memory and long-term memory, resulting in learning

Analytical thinking is "conscious, logical and a-contextual; places heavy loads on working memory; is energy-intensive (literally); epitomizes the kind of thinking that is usually associated with 'effective problem-solving'; operates on abstract concepts or rules that may involve logical combinations of individual features; and is abstract and inductive which allows us to deal with hypothetical situations where we have no prior experience".[164] Analytical thinking involves conscious activation, hypothesis testing and deductive reasoning, and is logically sound because it involves critical thinking.[42,152] When engaged in analytical thinking the brain only processes one channel at a time.[42,152] Our analytical thoughts become increasingly competent as we mature, socialize and go through formal education, and are refined by training in critical thinking

and logical reasoning.[152] Analytical thinking adopts a systematic approach which reduces uncertainty, and decisions made in this mode approach normative reasoning and rationality.[42,152] A disadvantage of analytical thinking is that it is resource intensive and it takes longer to reach a decision.[42]

4.3.2.3 Conversion of Analytical Thinking to Intuitive Thinking

The analytical mode of thinking carries with it the disadvantage of being a slow process.[158] The more times a decision maker comes into contact with the same situation that always requires the same sequence of processing, the situation becomes familiar.[158] At that point, the decision maker finds they can draw a conclusion on the situation much quicker and with less effort, the attention and analytical demands are eased, and intuitive thinking replaces analytical thinking for this particular situation. When this occurs, other situations can be carried out in parallel with intuitive processes.[158] The conversion of analytic processes to intuitive processes allows us to make efficient use of our minds and allows that system to be devoted to other types of processing necessary to learn new tasks.[158] We could not learn additional complex concepts without the conversion of analytic processes to intuitive processes because the more complex concepts often build upon our intuitive processes. Even with the occasional error, there are many benefits to operating within the intuitive mode of thinking. Without this mode of thinking we would have to methodically think through, and possibly relearn, a task every time it was performed. The intuitive mode of thinking allows us to acquire skills that are used as a basis for learning more complex tasks. Once the skills are learned, they become automatic and are performed without conscious thought. This frees our minds and permits us to focus our attention on tasks such as problem-solving.

4.3.2.4 Application of Dual Process Theory to Diagnostic Reasoning

A universal model of diagnostic reasoning based on the Dual Process Theory has been embraced by the members of the inaugural conference on Diagnostic Errors in Medicine – a group of researchers that are committed to the study of diagnostic errors. Figure 3 is an overview of this model. The numbers in parentheses in this figure correspond with the steps in the discussion below.

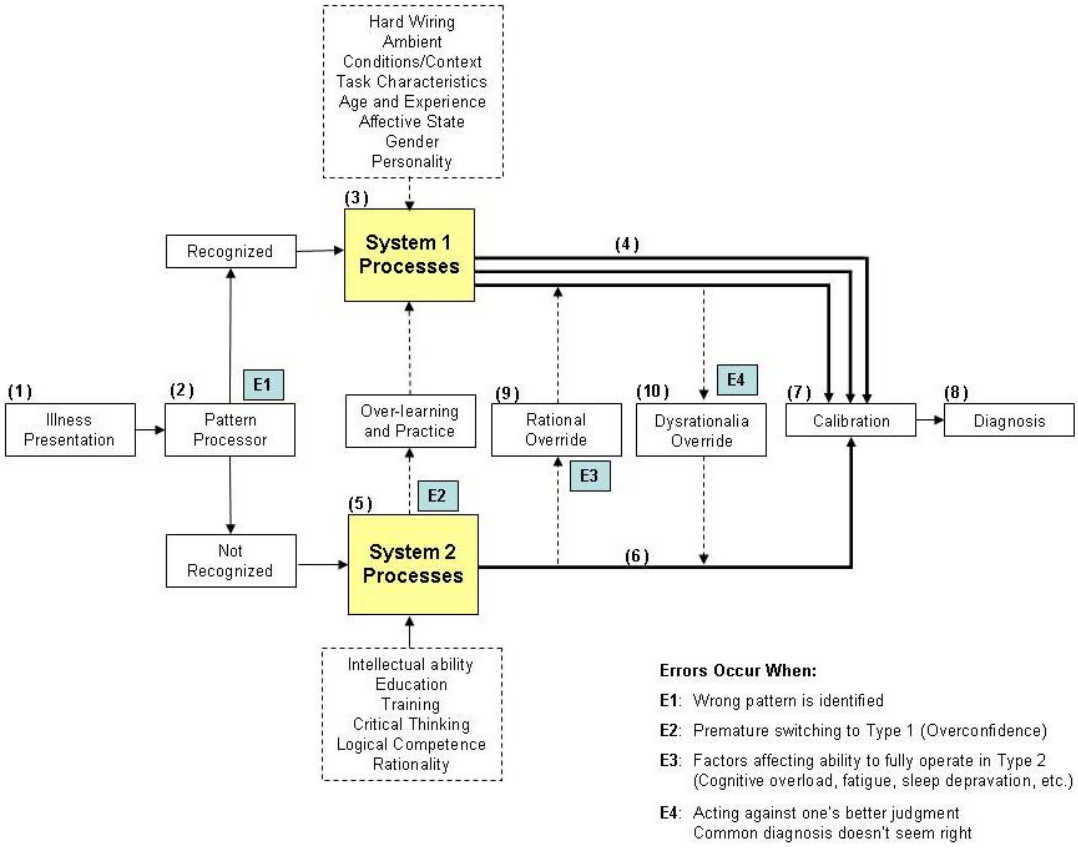


Figure 3 Universal Model of Diagnostic Reasoning [152]

Slightly modified from the figure presented in Croskerry, 2009

Step 1: The diagnostic reasoning process begins with the interaction between the patient and the physician. This interaction is normally by way of direct contact between the two; however the patient can also be presented to the physician through an intermediary such as family member or a junior physician.[152]

Step 2: As the physician assesses the clinical data, they run the data through a pattern recognition process (pattern processor) to determine if they can associate the data with a particular disease model.

Step 3: If the physician recognizes any salient features (or combination of findings), i.e., they recognize a pattern, their intuitive processes (System 1 Processes) are automatically initiated. This triggering of intuitive processes is an unconscious act - no deliberate thinking effort is involved. This automatic response can only occur if analytical processes have previously been

repeatedly engaged resulting in the physician learning the pattern (it becomes second nature to them).[152]

Step 4: While the pattern-recognition intuitive processes are active, other intuitive processes may be simultaneously generated. The physician may unconsciously have certain feelings toward the patient – they may have positive feelings toward some patients and negative feelings toward other patients.[152,169,170] Additional intuitive responses such as heuristics, intuitions, etc., can be triggered at the same time as the pattern-recognition response.[152]

Step 5: Analytical processes (System 2 Processes) are triggered if the pattern is not recognized or a disease presents in an unusual manner (is ambiguous). In this mode of thinking, the physician attempts to "make sense of the presentation through objective and systematic examination of the data, and by applying rules of reasoning and logic".[152]

Step 6: The analytic process is "a linear processing system, slower than intuitive processing, more costly in terms of investment of resources, relatively free of affective influences, and is considerably less prone to error".[152] Some of the factors that affect analytical reasoning are intellectual ability, education, training, critical thinking skills, logical competence and rationality.[152]

Step 7: If there are no subsequent modifications to the output of the intuitive and analytical processes (modifications are described below), the output is calibrated or tweaked as necessary.[152]

Step 8: Finally, the final diagnosis is made.

The model has several mechanisms for modifying its output.

Step 9: The intuitive and the analytical processes may interact with each other producing an output that is a synthesis of the two modes of thinking.[152,171] For example, the patients' initial presentation triggers an intuitive response, which subsequently triggers activation of analytical processes.[152] Analytical processes which monitor intuitive processes may result in the intuitive processes being rejected and overridden by the analytical processes (applying a rational override). For example, the initial review of a rash may result in a diagnosis of shingles. If atypical features exist, the analytical processes may override the initial diagnosis causing the physician to reassess the rash. Factors such as inattentiveness, distraction, fatigue and cognitive indolence may also detract from analytical processing, allowing intuitive processes "more

latitude than they deserve".[152] "The monitoring capacity of analytical processes depends on the deliberate mental effort and works best when the decision maker is well rested, has had enough sleep, free from distraction, and is focused on the task at hand".[152]

Step 10: Intuitive processes may also override reasoning being performed by analytical processes - a process labeled *dysrationalia*.[152] As a physician is analytically reasoning through a case (perhaps applying formal decision rules), intuition may take over resulting in them 'going with their gut' instead of continuing a fully analytical process.[152]

Those promoting and supporting this model of diagnostic reasoning do not suggest that all medical reasoning and decisions can be distinctly placed into either the intuitive or analytical mode of thinking. It is acknowledged that instead of a discrete separation of the two modes of thinking (systems), a "cognitive continuum with oscillation"[152] occurring between the intuitive and analytical processes quite possibly occurs, resulting in "varying degrees of efficiency and accuracy in judgment".[152,172]

There are several points within the diagnostic reasoning process where the process could fail leading to diagnostic errors.[173] These points are identified by the blue squares in Figure 3. The pattern the physician associates with the patient's presentation may me incorrect (E1).[173] Repeated exposures to a pattern in the analytical mode of thinking will eventually result in automatic triggering of intuitive processes. If the conversion of analytical processing to intuitive processing occurs prematurely (E2), a diagnostic error may occur.[173] The monitoring of analytical processes over intuitive processes may become compromised due to factors such as cognitive overload, fatigue, sleep deprivation, etc., all of which may diminish the capacity to provide adequate monitoring of intuitive processes (E3).[173] Finally, when intuitive processes are overridden by analytical reasoning, the result could be an incorrect diagnosis due to clinical decision rules being overridden by a physicians' gut feeling (intuition) (E4).[173]

4.3.2.5 Relationship of Heuristics to Mode of Thinking in DPT

A commonly held belief is that cognitive heuristics and biases occur during the intuitive mode of thinking. Pretz states "intuitive thinking has been generally thought of by cognitive psychologists in the decision-making tradition as synonymous with heuristic".[174] Typically, these heuristics will apply during pattern recognition. For example, the clinician may recognize the pattern

because they have recently seen a similar pattern in a previous patient – the *Availability* heuristic, or may make a diagnosis on a case because it resembles a disease model they believe they know well – the *Representativeness* heuristic.

As a result of heuristic use in the intuitive mode of thinking, it is also believed that more errors occur in the intuitive mode of thinking than in the analytical mode of thinking. "The dominant view in psychology, as in medicine, is that analytic reasoning is 'good' and intuitive reasoning is 'bad'".[164] However, there are those that believe heuristics and biases are also used in the analytical mode of thinking, that the two systems are equally effective, and that intuitive processing is no more prone to error than analytical processing.

By examining the frequency of use of *Anchoring and Adjustment* and *Confirmation Bias*, and their association to diagnostic errors in both modes of thinking, this research has the potential to provide insight on this much debated topic.

5 PRIOR RESEARCH ON COGNITIVE HEURISTIC DEBIASING AND COGNITIVE FEEDBACK

5.1 PRIOR RESEARCH ON COGNITIVE HEURISTIC DEBIASING

A number of researchers have attempted to debias individuals from committing heuristic errors using a wide variety of methods.[9,2-25,48,54,56,79,118,119,120-132] Table 9 highlights examples of some empirical debiasing studies. Debiasing studies based on normative decision theories such as providing training on the concepts of statistics, probability, Bayes' Theorem, and calculation of expected utilities, have had limited success.[123, 130] Other attempts have provided subjects with training on cognitive heuristics and biases, providing examples of actions that correspond to proper and improper use of heuristics.[9,118,123,126,127,129] Some researchers have gone a step further by requiring subjects to justify their choice,[131] rewrite the scenario to induce assessment of the entire problem,[131] consider opposite and alternative choices,[122,124,129] provide assessment of others' decisions[122]; take notes while assessing decision alternatives,[129] and engage in repeated practice.[126] To minimize biased judgment, during some empirical studies subjects have been trained on optimal decision strategies;[132] have been provided with feedback that details reasoning flaws; and have been given suggestions on ways to avoid biases.[122,129]

Empirical studies that have resulted in limited or no success in debiasing individuals have used techniques such as providing didactic instruction on reasoning strategies without providing concrete examples and the opportunity to practice the skills taught. Several studies that have provided instruction on the use of normative decision-making theories such as statistical concepts and Bayes' Theorem have resulted in limited reduction of biased judgment.[123,127,128,131] When subjects do not thoroughly assess the situation, or are not required to justify their decision, the reduction of biased judgment is minimal. In a study conducted by Arkes, Faust, Guilmette

and Hart, there were two study groups, one which was required to provide evidence from the clinical case to support their diagnosis, and the other group that was not required to provide a rationale for their decision.[72] The study results revealed that the subjects who did not provide a rationale for their diagnosis was significantly more susceptible to biasing than the group that provided a rationale.[72] A well known researcher of cognitive debiasing, Baruch Fischhoff, describes several techniques used in debiasing studies that have not resulted in successful debiasing.[122] These techniques include explicitly describing the bias to subjects and asking them to avoid it in their judgments, replacing diverse sets of general knowledge questions with homogeneous items or perceptual items, raising the stakes and making individuals accountable for decisions, and changing rating scales of degree of certainty by using verbal ratings instead of numerical.[122]

Many of the empirical studies that have been the most successful in reducing biased judgments used multiple techniques within a single study. For example, in a Wolf, Gruppen, Billi study, they provided subjects with training on selection of optimal diagnostic information; instructed them on the proper use of Bayes' Theorem; and provided an explanation on the proper use of heuristics.[132] This type of intervention provided subjects with skills that allowed them to apply multiple techniques to the problem to arrive at an optimal decision. Other techniques that have reduced biased judgment include providing subjects with the opportunity to reinforce concepts taught during the study, and using feedback to explain flawed reasoning. In a study conducted by Parmley, subjects were given examples of how to adjust reasoning strategies to avoid committing the *Confirmation Bias*; engaged in repeated practice on the techniques; and were provided with explanations when they arrived at incorrect answers.[9] This study provided subjects with the ability to reinforce valuable techniques through repeated practice and enhanced outcomes by providing feedback regarding errors. The same technique of repeated practice was used in a Koriat, Bjork study.[126] Another technique that has resulted in reduced biased judgment is to stimulate examination of the entire scenario or problem space. In a study conducted by Mumma and Wison, they used the techniques of considering an opposite alternative, taking notes, and causing the subjects to focus on certain aspects of the scenario by requiring them to order cues within the scenario.[129] Similar techniques of considering the alternatives were used in the Hirt, Markman study.[124] Techniques described by Fischhoff that have successfully debiased individuals include receiving feedback on large samples of responses and performance,

providing personalized feedback, discussing the relationship between one's subjective feelings of uncertainty and numerical probability responses, and having subjects list reasons why their preferred answer might be wrong.[122]

This research utilized some of the techniques used in previous studies that successfully reduced biased judgment. This study will use multiple techniques within a single study including requiring the subjects to engage in repeated practice to reinforce the diagnostic reasoning strategies learned during the intervention period; providing personalized feedback associated with the subjects' disease mental models; and stimulating examination of the entire clinical scenario by providing alternative reasoning strategies that incorporate clinical data that the subject did not take into consideration. This research used a form of the '*consider the alternatives'* approach in that it presents the subject with additional disease mental models that apply to the case under assessment. However, the research is slightly different than the technique used by previous studies in that those studies commonly require the subject to provide an explanation as to why they did not select other alternatives. This study does not require subjects to specify why they did not construct these alternative disease mental models.

Despite limited success in previous efforts, work continues to identify and test new methods with the hope of greater success. Several published articles and editorials suggest novel techniques, which have not yet been tested, to address the problem.[48,54,56,122] Additionally, there are recent working papers describing debiasing studies,[200] a poster describing recent research assessing clinicians' use of cognitive heuristics at the 2009 American Medical Informatics Association (AMIA)[201] conference, and a poster detailing a recent debiasing study that was presented at the 2009 Canadian Association of Emergency Physicians (CAEP) conference.[202] As described in Chapter 3, the negative effects of biased judgment in clinical reasoning are of such importance that they justify continued attempts to reduce their impact.

Table 9 Cognitive Debiasing Empirical Studies

Heuristic / Bias	Debias Technique / Study Design	Researchers	Subjects / Task Performed in Study / Results	Comments
Confirmation Bias (CB)	Training on heuristic and why it occurs. Examples of how to adjust for the bias. Required subjects to practice. Provided explanations on wrong answers. Study design: 2x2x2 mixed factorial design - Debiasing training vs. no debiasing training - Diagnostic change vs. no diagnostic change - Baseline vs. one week retest	Parmley[9]	Licensed psychologists completed 2 vignettes at baseline - additional information given a week later. Training group evidenced bias 23% of the time. Without training evidenced bias 33% of the time (not statistically significant) Over-correction: Training did not help or harm - Training group 5.7% vs. 6.1% non-training group Diagnostic accuracy with training – 77% Diagnostic accuracy with no training – 67%	PhD dissertation (2006) Hypotheses: 1. A significant number of clinicians will evidence the CB & fail to alter initial Dx even when new information presented later has disconfirming evidence. 2. The presence of training on the CB will mitigate the effects of the bias. 3. Training will not cause inaccurately alteration of diagnosis (over-correction)
Pseudo-Diagnosticity / Confirmation Bias	Training on selection of optimal diagnostic information (competing hypothesis heuristic). Instruction on Bayes. Explanation of proper use of heuristics. Study design: Intervention and control group	Wolf, Gruppen, Billi[132]	Medical students assessed 3 clinical cases. Measured % of subjects selecting optimal information. Results showed significant differences in intervention group. Case 1 – No intervention Case 2: 58% control vs. 95% intervention Case 3: 62% control vs. 87% intervention	Medically based study
Confirmation Bias	Provided subjects with varying levels of statistics and explanation for topic being tested. Study design: 3 groups of subjects – *Abstract plus Statistics* group; *Concrete plus Statistics* group; *Concrete only* group.	McKenzie[128]	Graduate students assessed scenarios. *Abstract plus Statistics* group not sensitive to intervention (line almost flat). *Concrete plus Statistics* group more sensitive than Abstract plus Statistics group but not as sensitive as *Concrete only* group (most sensitive)	The purpose of this study was to prove that confirmation bias is not as wide spread as people think
Anchoring / Primacy Effects	3x2 study - 3 debiasing and 2 study groups Debiasing Groups: 1. *Bias Inoculation* - subjects received heuristic training and engaged in practice trials using heuristic. 2. *Consider the Opposite* - subjects asked to consider the opposite alternative prior to making final judgment. 3. *Note Taking Group*- same as Consider the Opposite group except subjects took any notes on each category. Study Groups: 1. *Single Cue Anchoring* – subject given single, relatively extreme cue first (vs. toward the end) in a series of cues. 2. *Sequence Anchoring* – cues ranked according to applicability to diagnosticity (ascending vs. descending sequence)	Mumma, Wilson[129]	Undergraduate psychology students reviewed scenarios Study Group Effectiveness - Anchoring occurred in both groups. 1. Single cue anchoring effect - small to moderate 2. Sequence anchoring effect - moderate to large Debiasing Intervention Effectiveness - 1. *Single Cue Anchoring* effect – debiased successfully by *Consider the Opposite* and *Note Taking (*not significantly different). *Bias Inoculation* marginally statistical significant debiasing 2. *Sequence Anchoring Effect* – not successful for any of the debiasing interventions	Hypothesis: 1. Single cue and sequence anchoring manipulations will result in Anchoring 2. Anchoring effects will generalize to other domains
Availability	Used a *Consider the Alternative* debiasing technique Study design: Intervention and control group	Hirt, Markman[124]	Undergraduate students reviewed scenarios and generated explanation between personality trait and profession (risk-taker / fire-fighter). Availability was evident. Intervention group showed evidence of debiasing (correlation = 0.3 – the only statistic provided)	

Heuristic / Bias	Debias Technique / Study Design	Researchers	Subjects / Task Performed in Study / Results	Comments
Representativeness	Formal instruction - trained subjects on concepts of Law of Large Numbers. Four training groups and a control group.	Fong, Krantz, Nisbett[123]	Undergraduates solved statistical problems. Training increased frequency of statistic use and quality of statistically based responses	Multiple experiments – same domain & different domains – same results
Representativeness	Varying levels of training on statistical concepts & guidance on types of errors to avoid. Four training groups and a control group.	Kosonen, Winne[127]	Extension to Fong research – Results the same as Fong et. al	Multiple experiments –same domain & different domains – results were the same
Status Quo Bias / Satisficing	Study to determine how heuristic is used - not a de-biasing study. 2 groups – 1 vs. 2 medication alternative	Roswarski, Murray[25] Redelmaier, Shafir[22]	Mailed practicing physicians clinical vignette. Options: refer patient, no medication prescription or refer patient and prescribe new medication. Non-significant effect of multiple alternatives (contrary to previous studies). Both groups chose to refer without starting new meds. 1 medication group – 45.5% 2 medication group – 44.0%	Study purpose was to determine how professional characteristics (experience, workload, fatigue, continuing education, supervision) and practices of physicians alter the selection of medical treatments
Foresight Bias	Heuristic training and repeated practice	Koriat, Bjork[126]	Undergraduate students studied lists of paired associates then were tested on them. Results show both techniques reduced bias	Results were transferrable to additional domains (different lists)
Hindsight Bias	Consider the alternatives & provide a reason for selection Study Groups (2x2 study): *Provide Reasons* group: Foresight vs. Hindsight group *Don't Provide Reason* group: Foresight vs. Hindsight group	Arkes, Faust, Guilmette, Hart[72]	Neuropsychologists reviewed medical cases. Foresight groups provided Dx and probability of Dx. Hindsight groups provided probability for Dx they were given. *Provide Reason* groups provided evidence from case to support Dx. Results show hindsight bias significantly greater among hindsight-don't provide reason than hindsight-provide reason subjects.	
Hindsight and Over Confidence	Various techniques have been attempted	Fischoff[122]	Few strategies have reduced hindsight. None have eliminated it. When looking in hindsight people consistently exaggerate what could have been anticipated in foresight. Reference book chapter for what has worked, not worked, may work, but not yet tried.	Book chapter in *Judgment under Uncertainty: Heuristics and Biases* book discussing research studies attempting to reduce hindsight and over-confidence
Vertical Line Failure (VLF)	Mannequin-based simulation to introduce VLF cognitive failure, followed by debriefing session. Study design: Cognitive (training on heuristics / biases) vs. technical (technical aspects of performance) debriefing	Bond, et al.[118]	Medical personnel of varying levels participated in simulation. Results showed that some subjects (not all) in both debriefing groups increased awareness of widespread risk of error. Those who feel exercise did not increase their awareness of risk of error felt they already knew the high risk of error. Those in cognitive group felt information provided was a bit overwhelming.	
Framing Effect	Development of causal cognitive maps to capture strategic thought processes. Study design: Draw casual map before (prochoice group) vs. after (post-choice group) making decision	Hodgkinson, et al.[125]	Undergraduate students assessed elaborate strategic decision scenario. Results showed post-choice succumbed to framing bias. Prochoice group did not – framing effect was eliminated.	First study used undergraduate students; second used senior business managers – same results
Framing Effect	Pinged statistical training against deeper thought processing (Requiring subjects to provide a reason for their choice) Study design: Math skill group vs. Need-for-cognition group	Simon, Fagley, Halleran[131]	Undergraduate students assess math problems and situations involving life outcomes (having a disease, cancer treatment, etc.) Results showed both math and need-for-cognition groups attenuated framing effects. Deeper processing markedly reduced framing	No p-value given in article. Markedly reduced is terms used by author (no definition provided)

5.2 PRIOR WORK ON COGNTIVE FEEDBACK

5.2.1 Impact of Different Types of Feedback

When making decisions people try to make the best decision possible given the information available to them. Receiving feedback that the decision made resulted in the optimal outcome is not always possible, especially in the constantly changing situations commonly found in Medicine. Physicians do not always know if the diagnosis they made was accurate or the treatment they prescribed resulted in improvement of the patient's condition.

Empirical studies have assessed the impact of various types of feedback on outcomes including *performance (outcome) feedback* that designates if a problem was solved correctly or incorrectly; *cognitive feedback* which provides an explanation of how to perform a task; and *feedforward feedback* which enables the decision maker to examine the future effects of their actions.[134] Studies have shown that improved performance does not occur when receiving only *performance (outcome) feedback* within dynamic decision-making tasks such as medical emergency rooms.[134,139] Studies have shown limited effectiveness, even negative effects, of *performance (outcome) feedback* as a means of improving judgment in situations of uncertainty.[134,137,138] In contrast, the addition of the other forms of feedback results in enhanced performance. For example, in a study conducted by Sengupta and Abdel-Hamid, individuals who performed a software management task were given different levels of feedback – performance feedback alone, performance feedback with cognitive feedback, or performance feedback with feedforward feedback.[134,135] The combination of performance feedback with cognitive or feedforward feedback resulted in higher levels of performance than performance feedback alone.[134,135]

There are commonly two techniques used to deliver *feedforward feedback:* (a) providing an individual with guidelines from experts performing the task, or (b) allowing subjects to look into the future by observing effects of various present actions.[134,135,139] An example of the first approach can be seen in a study conducted by Gonzalez in which outcome and feedforward feedback was given to subjects completing a dynamic decision task. In this case, feedforward feedback was implemented by

providing subjects with the ability to compare their decisions with an expert by watching a video of an expert performing the task.[134] Results of the Gonzalez study showed that "participants who received feedforward feedback improved their performance considerably and continued to exhibit improved performance even after performing the task 24 hours later with a discontinuation of the feedback".[134] An example of the second approach can be seen in a study conducted by Lerch and Harter in which subjects received outcome feedback and feedforward feedback using the second approach.[140] In this study, the condition which used only feedforward feedback resulted in hampered performance and hindered learning, but when coupled with outcome feedback, there was a slight improvement in performance.[139,140]

5.2.2 Temporal Aspects of Feedback

Many empirical studies have been conducted to assess various aspects of feedback. One area of study has been assessing how feedback impacts a student when learning a new task. Several studies have shown the benefits of immediate feedback when using a cognitive tutoring system that provides guidance as a student performs a specific task. There is strong evidence that students reach mastery level quickly when using directive systems that provide immediate feedback.[147] Crowley, et al. assessed the effects of a computer-based medical intelligent tutoring system that provided immediate feedback on diagnostic performance gains in a complex medical domain. They determined that feature (evidence) identification performance improved significantly between pre-test and post-test assessments, and that the learning gains were entirely retained at a later retention test.[148] A second study showed similar effects when training residents to write diagnostic reports.[202] In another study, it was shown that immediate feedback in an intelligent tutoring system had a statistically significant positive effect on learning gains as well as metacognitive performance.[149] The removal of immediate feedback was associated with decreasing metacognitive performance, and other metacognitive scaffolds were not sufficient to replace immediate feedback.[149] Thus, recent research suggests that immediate feedback may impact metacognitive performance in addition to diagnostic accuracy.

5.2.3 Feedback and Diagnostic Errors

There is a growing recognition of the critical role and potential benefit of feedback in reducing diagnostic errors. Recent literature highlighting this trend includes empirical studies on diagnostic errors,[84,143] studies of clinical reasoning,[143] and field observations.[142] The use of feedback to address diagnostic errors has yet to be fully explored. 'How can we learn from errors if we don't know that an error has occurred?' is a viable question raised by Arthur S. Elstein in an article published in August, 2009 entitled "Thinking about diagnostic thinking: a 30-year perspective."[143] Elstein indicates "Improving feedback to clinical practitioners may be the most effective debiasing procedure available."[143] Strategies proposed to reduce diagnostic errors include increasing ones' knowledge, using clinical algorithms and guidelines, *reflecting on ones' decision-making process*, seeking second opinions and *debiasing an individual with improved feedback*.[143] Providing prompt feedback on decision-making has proven to be a reasonably effective debiasing procedure.[143] Ericsson states that expert performance is acquired by practice and feedback.[206] In describing feedback in the medical setting, Croskerry states

> "Feedback is the process by which clinicians find out about the veracity and efficacy of their clinical decisions that led to their diagnosis and management strategy, and inevitably, what happened to their patients over time. There is little justification for physicians to change their decision-making unless they know it to be wrong or otherwise suboptimal. In the absence of feedback about the outcome of their decisions, the assumption will be made that veracity and efficacy are optimal. There is no point in changing something that appears to be working. Thus, favorable or unfavorable feedback, appropriately given, can change the calibration of the decision maker."[142]

According to Schiff, a leader of the members of the inaugural conference on diagnostic errors in medicine, improving the provision of feedback and how it is used in the clinical setting is a priority task for reducing diagnostic errors.[144] When discussing the reduction of diagnostic errors by using cognitive interventions, Graber indicates "the pathway to improved calibration involves focused, timely and relevant feedback".[34] Not only is feedback a critical component to addressing diagnostic errors, in order for feedback to be effective it must be provided as quickly as possible after the a diagnosis has been confirmed.[145]

There is evidence that if a physician receives immediate feedback during the diagnostic reasoning process, over time, diagnostic errors will diminish.[207] A recent 2009 publication authored by Payne, Crowley, et al. provides details of a study conducted in which subjects diagnosed pathology

cases using an computer-based intelligent medical tutoring system. Using this system, subjects review virtual pathology slides; identify present and absent features and feature attribute values critical to their diagnosis; specify initial hypotheses; evaluate and refine hypotheses; and specify a final diagnosis. Based on each subject action, the intelligent tutoring system generates a dynamic solution graph which is a directed acyclic graph that models the current problem space and all valid-next-steps, including the *best-next-step*. The *best-next-step* represents the optimal action to be performed next in the diagnostic reasoning process. The system does not present the directed acyclic graph or the best-next-step to the subject – the system creates these in the background and uses them to provide the subject with feedback on possible reasoning strategies. In response to each action taken by the subject, the system provides immediate performance feedback indicating the accuracy of that action. For example, if the subject identifies a hypothesis that is not supported by the identified features, the system will alert the subject that the hypothesis is not supported by the evidence they identified. The subject can also request a hint as to what step to take next. When this occurs, the system will also provide hints regarding reasoning strategies and critical steps that should be taken to arrive at the proper diagnosis. One of several objectives of this research was to determine the impact immediate feedback has on errors that occur during the diagnostic reasoning process, and ultimately on diagnostic errors such as misdiagnosis. We found that as subjects continue to receive immediate feedback regarding reasoning strategies during the diagnostic reasoning process, over time, errors associated with critical aspects of diagnostic reasoning (identifying critical evidence within the case, generating initial hypotheses, refining hypotheses, and specifying a diagnosis) diminish.[207] Some categories we assessed diminished to nearly zero.[207] These findings are critical in the study of the reduction of diagnostic errors, errors that can occur during the diagnostic reasoning process, and how immediate feedback can significantly reduce such errors.

Empirical results and other studies suggest that timely and relevant feedback is the key. This research used timely and relevant feedback as part of the intervention and assessed the impact of diagnostic accuracy.

6 RESEARCH STATEMENT

Cognitive heuristics are used to reduce the complexity of large amounts of data to a manageable dimension and can be a valuable tool during the diagnostic reasoning process. However, critical diagnostic errors may occur if heuristics are used inappropriately. This research study will assess the use of heuristics and biases during diagnostic reasoning and determine how such use impacts diagnostic reasoning and medical errors.

6.1 RESEARCH OBJECTIVES

The objectives of this research study are to:

1. Reduce biased judgment and diagnostic errors using a metacognitive intervention by inducing physicians to think about how they think.
2. Reduce biased judgment and diagnostic errors by providing feedback regarding reasoning strategies in association with Mental Model Theory.

6.2 RESEARCH QUESTIONS

The questions that were investigated during this research are as follows:

1. What is the frequency of use of the cognitive heuristic *Anchoring and Adjustment* and the cognitive bias *Confirmation Bias* during diagnostic reasoning?
2. Does the use of *Anchoring and Adjustment* and/or *Confirmation Bias* impact diagnostic accuracy?

3. How does a feedback-based intervention, modeled after the Mental Model Theory, received during diagnostic reasoning, impact:
   a. The post-test use of *Anchoring and Adjustment* and *Confirmation Bias*?
   b. Diagnostic accuracy?
4. Does heuristic use and diagnostic accuracy differ in the Intuitive and Analytical modes of thinking?
   a. What is the frequency of use of the cognitive heuristic *Anchoring and Adjustment* and the cognitive bias *Confirmation Bias* in each mode of thinking?
   b. What is the frequency of diagnostic errors in each mode of thinking?

6.3 HEURISTIC AND BIAS UNDER STUDY

The cognitive heuristic and bias that will be assessed during this research study are:

Anchoring and Adjustment – the adjustment up or down from an original starting value, or anchor. In diagnosis, this is the tendency to lock onto salient features early in the diagnostic process, followed by adjusting this initial impression in the light of later information. This heuristic is commonly associated with the *Confirmation Bias*.

Confirmation Bias – a phenomenon wherein decision makers actively seek out and assign more weight to evidence that confirms their hypothesis, and ignore evidence that could disconfirm their hypothesis and/or lead to an alternative diagnosis.

These heuristics have been selected for study because evidence has shown that *Anchoring and Adjustment* and *Confirmation Bias* occur frequently during clinical reasoning. In a study performed by Graber et al. it was determined that premature closure was the most common cause of cognitive-based diagnostic errors.[35] Premature closure, which is the tendency to stop considering other possibilities after reaching a diagnosis, is often caused by the commission of *Anchoring* and/or *Confirmation Bias*. The occurrence of *Anchoring* and *Confirmation Bias* and *Premature Closure* in clinical reasoning has been corroborated in various empirical studies conducted by numerous research groups.[7, 9, 49, 65, 140, 141] There has been previous work showing that the commission of *Confirmation Bias* can be identified using think-aloud and eye-tracking techniques[9], which I propose to use in my dissertation research. A

published study conducted by Payne and Crowley demonstrated that the think-aloud technique was successful in identifying the use of Representativeness during diagnostic reasoning.[19]

6.4 RESEARCH METHODS

During this research, two experimental studies were performed. Study one will address research questions 1 and 2 by gathering descriptive statistics on the frequency of heuristic use and diagnostic accuracy. Study two will address research questions 3 and 4 by testing the impact of a metacognitive intervention and assessing heuristic use and diagnostic accuracy within mode of thinking. The sections within this chapter will address each research question, referencing the appropriate experimental study where applicable.

6.4.1 Subjects

6.4.1.1 Type and Number of Subjects

Fourth year medical students and resident physicians were the subjects used for this study. Residents within the Division of General Internal Medicine enrolled in the Internal Medicine Residency Program, Transitional Year Residency Program and the Family Medicine Residency Program were solicited for participation. In order to obtain the number of subjects required to meet power requirements, residents of all post-graduate years were solicited. Reference chapter 7 for a breakdown of the number fourth-year medical students and residents.

Based on the consultation with a Clinical and Translational Science Institute (CTSI) statistician, for the first study, the sample size was calculated with an alpha error probability of 0.05, a power of 0.80 and a medium effect size, using a Chi-square test assessing the variance for one study group. This resulted in a sample size of 74 subjects. For study two, a sample size of 35 subjects was calculated using an alpha error probability of 0.05, a power of 0.80 and a medium effect size (0.30), using a 2 sample t-test assessing the difference between means of repeated measures (pre- and post-test) between two study groups (control and intervention). During the data-analysis phase of the project, the biostatistician recommended using logistic regression. The power requirements for a Chi-

square test are more stringent than a logistic regression test; therefore recalculating the power (number of subjects) was not required.

6.4.1.2 Subject Approval, Recruitment and Consent

A letter explaining the research study, subject expectations, and payment information was sent to Dr. John F. Mahoney, MD, the Associate Dean for Medical Education. Dr. Mahoney chairs a committee that must approve the use of medical students for research purposes. Upon review of the research project overview, this committee granted permission to use the medical students as subjects. The approval letter from this committee is located in Appendix A.

The University of Pittsburgh Institutional Review Board (IRB) reviewed the research study protocol and granted approval of the use of human subjects under IRB # PRO09120344. The IRB approval letter is located in Appendix A.

Fourth year medical students and residents received an email solicitation from the student affairs office and residents received an email from their program coordinator requesting their participation in the study. Recruitment letters and consent forms are located in Appendix A. Subjects were paid $50.00 per hour for their participation in this research study; a rate approved by the University of Pittsburgh IRB.

6.4.2 Instrument

During this research study subjects assessed clinical scenarios developed by a group of researchers led by Charles Friedman, Arthur Elstein and Fredric Wolf. [150] The cases were based on patients presenting at three academic medical centers: the University of Illinois in Chicago, the University of Michigan in Ann Arbor and the University of North Carolina in Chapel Hill. The cases represented diseases clinicians would commonly be required to assess within daily clinical rounds and cases representing rare diseases and/or diseases presenting in an unusual manner. The authors of the cases provided a definitive correct diagnosis for each case which was the diagnosis that was used as the gold standard for the study. Figure 4 is an example of a case that the subjects assessed.

Patient #092

Chief Complaint: This 65 year old white male has had right rib pain for 1 year. The pain has been severe for the past one and a half weeks.

History of Present Illness: The patient developed back pain in his lower right rib region about 1 year ago. It became so severe 1.5 weeks ago that he had difficulty sleeping. He then required morphine for relief. The pain was worse with deep breathing, coughing, sneezing and movement of his chest. A bone scan prior to admission showed multiple abnormal foci in an asymmetric pattern in bones including the ribs on the right and the left, ischium and tibia. An X-ray of his right ribs showed a fracture. There was no history of trauma. He had smoked 1 pack per day for 35 years but quit 2 years ago. He denied other symptoms including any related to his bowels or urinary system. He had not seen blood in his stool or urine. He had gained 15 pounds over the past month or so. Nausea had occurred once a week. Prior laboratory studies were said to show a normal serum leucine aminopeptidase, but an elevated alkaline phosphatase.

Previous Medical/Surgical History: Degenerative joint disease for which he had a right hip arthroplasty 3 months previously. Nephrolithiasis for which he had a lithotripsy. Crohn's disease for which he has a resection of his transverse colon 12 years before and resection of his small bowel 5 years before. Transurethral resection of his prostate on two occasions; no malignancy was reported. Hypertension and hypothyroidism.

Medications: Morphine, Tylenol #4, Nortriptyline, Zantac, Torecan, Azulfidine, Synthroid, Lasix, KCl, Folic Acid.

Social History: The patient is a retired tool and dye maker.

Physical Examination: The patient appeared to be in marked pain. Vital signs: BP 121/73, pulse 103/min, respirations 20/min. There was marked tenderness in area of back pain over the right lower rib cage. Tenderness was also present over the left rib cage. There were no nodules in his prostate. There were no other significant findings.

Laboratory Data:

			At a Later Date	Normal
CBC	Hct	36.1	33.0	*42-52%*
	Hgb	12.5	11.3	*14.0-18.0g/dl.*
	MCV	90.2		*80-100 fl*
	MCHC	31.1		*32-36%*
	RBC	3.8	3.6	*4.2-6.2 X 103/mm3*
	WBC	8.1		*4-10 X 10^9/L*
	Neut	75		*50-75 %*
	lymph's	22		*20-50 %*
	mono	normal		*3-10 %*
	platelet count	314		*200-400 X 10^9/L*
Chemistries	sodium	138		*136-146 mmol/l*
	potassium	4.0		*3.5-5.0 mmol/l*
	chloride	111		*99-111 mmol/l*
	CO2	18		*24-34 mmol/l*
	creatinine	1.1		*.9-1.3 mg/dl*
	BUN	6.0		*8-20 mg/dl*
	calcium	8.5	8.7	*8.6-10.2 mg/dl*
	phosphorus	2.4	2.2	*2.5-4.9 mg/dl*
	magnesium	2.1		*1.5-2.3 mg/dl*
	protein, total	6.1		*6.0-8.3 g/dl*
	albumin	3.1		*3.5-4.9 g/dl*
	bilirubin, total	0.4		*0.1-1.1 mg/dl*
	AST (SGOT)	16		*2-35 U/L*
	ALT (SGPT)	13		*0-45 U/L*
	LDH	220	206	*60-200 U/L*
	ALP	353	321	*30-130 U/L*
Serum leucine-amino peptidase		14 (normal)		

		At a Later Date	Normal
PT	13.0		*10-13 sec*
TSH	0.6		*0.3-6.5 uU/ml*
PSA	<1		*Up to 4 ng/ml*

Serum protein electrophoresis: no monoclonal spike. **Urinalysis:** normal
Chest X-ray: non-specific interstitial markings in right upper lobe. Sputum negative for malignant cells.

Figure 4 Clinical Case Sample

The case authors provided a clinical difficulty level for each case. This level ranged from 1 representing the easiest level to 7, the hardest level. For this study, the cases were categorized as easier, medium or harder cases. The difficulty levels from 1 through 4 were classified as easy; levels 4.00 through 5.50 were classified as medium; levels greater than 5.50 were considered harder cases. The doctoral candidate consulted with three board-certified physicians to determine the appropriate case difficulty rating based on the manner that the disease was presented in the case, and the difficulty levels assigned by the case authors (who were also board-certified physicians). This process was performed during the case review by these experts as described in the next section (section 6.4.3). The division of cases, showing the clinical difficulty provided by the authors and the easy, medium and harder levels are provided in Appendix A.

6.4.3 Expert Case Models

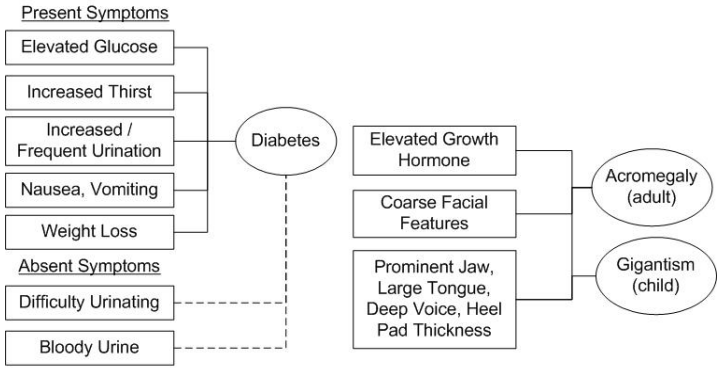
As part of this research project, three board certified physicians independently annotated the clinical cases identifying data corresponding to the gold-standard diagnosis provided by the case authors. Reviewer one annotated eighteen cases; reviewer two also annotated eighteen cases including thirteen of the same cases assessed by reviewer one. Reviewer three annotated ten cases, five assessed by reviewer one and five assessed by reviewer two. Table 10 details the cases analyzed by each reviewer.

Table 10 Expert Case Annotation

	Case	Case Diagnosis	Reviewer	Reviewer	Reviewer
Easier Cases	52	Colon Cancer	1	2	
	91	Guillain-Barre Syndrome	1	2	
	32	Ulcerative Colitis	1	2	
	42	Appendicitis	1	2	
	133	Pernicious Anemia	1	2	
	62	Cryptococcal Meningitis		2	3
	102	Pheochromocytoma	1		3
Medium Cases	53	Crohn's Disease	1		3
	21	Carcinoid Syndrome	1	2	
	93	Subarachnoid Hemorrhage	1	2	
	72	Hemolytic Uremic Syndrome		2	3
	12	Metastic Hepatic Adeno (liver) Cancer	1		3
	83	Aortic Dissection	1	2	
	181	Temporal Arteritis	1	2	
	122	Hypokalemic Periodic Paralysis		2	3
Hard Cases	11	Blastomycosis		2	3
	31	Cryoglobulinemia		2	3
	82	Miliary (disseminated) TB	1	2	
	103	Cardiac Amyloidosis	1	2	
	1	Acromegaly	1	2	
	123	Syphilitic Meningitis	1	2	
	112	Whipple's Disease	1		3
	291	Gaucher's Disease	1		3

The annotations from each independent reviewer were analyzed and the differences were noted. A meeting took place with all reviewers in attendance to resolve the differences. For the cases in which a consensus was not reached between the two reviewers who annotated the case, the third reviewer reviewed the case. Discussion among the three reviewers resulted in a consensus agreement being reached for all cases on the critical data that is associated with the gold standard diagnosis. Using this data, a mental model representing the data an expert would use to diagnose the case was developed. The derived mental models were verified by the board certified physicians for accuracy and medical soundness. An example of the mental models associated with case 001 which has a diagnosis of Acromegaly is provided in Figure 5. This case could easily be misdiagnosed as Diabetes without knowledge that critical features associated with Acromegaly is an elevated growth hormone level resulting in the enlargement of body tissues, commonly presenting as severe disfigurement of the facial features.

This case includes symptoms commonly associated with **Diabetes** including high levels of sugar in the blood (glucose) which can cause increased thirst and urination, nausea, vomiting and weight loss. The absence of fever, chills, excessive sweating, dysuria (painful urination and/or difficulty urinating) and hematuria (blood in the urine) are additional features that could lead to a diagnosis of Diabetes. If you focus on these symptoms, you will misdiagnose the patient with Diabetes. The figure below is a graphic depicting the symptoms relating to Diabetes.



In order to diagnose this case properly, you have to realize that critical findings of **Acromegaly** include an elevated growth hormone level resulting in the enlargement of body tissues, commonly presenting as severe disfigurement of the facial features, hands, ears, etc. The patient population is also significant in that excessive growth hormone in adults is Acromegaly; and in children it is **Gigantism**. The graphic above depicts the symptoms related to the correct diagnosis Acromegaly.

Figure 5 Case 001 Mental Model

6.4.4 Study Design

This research project consisted of two experimental studies. The purpose of the first research study was to determine how frequently fourth-year medical students and residents commit *Anchoring* and *Confirmation Bias* during diagnosis. Data from this study was used to answer research questions one and two. The purpose of the second research study was to (1) assess the impact of a feedback intervention on the use of *Anchoring and Adjustment*, *Confirmation Bias* and diagnostic errors; (2) determine how frequently this cognitive heuristic and bias are used in the Intuitive and Analytical modes of thinking; and (3) determine the frequency of diagnostic errors in each mode of thinking. Data collected from this study was used to answer research questions three and four.

Experimental study one consisted of subjects independently assessing twenty-four (24) clinical scenarios using an Internet-based system. This study was a **descriptive study** used to assess the frequency of events; this study did not consist of the comparison of a control and intervention group (reference Figure 6).

Obtain Informed Consent and Introduce Computer System	**Case Assessment**
	Each subject silently assessed 24 clinical cases

Figure 6 Experimental Study One Research Design

Experimental study two consisted of a **Between Subjects** (control and intervention)**, Repeated Measures** (pre-test and post-test analysis) study design (reference Figure 7). Using a computer-based system, subjects assessed fifteen (15) clinical scenarios during three (3) periods. Each period consists of subjects assessing five consecutive cases. The first period was used to establish a baseline of the frequency of diagnostic errors and cognitive heuristic / bias use. In the second period, subjects received feedback after each case. The third period was used to determine if the intervention was corrective.

Subjects were assigned to one of two method groups including (1) a **think-aloud group** where subjects were asked to think-aloud and verbalize their thought processes; and (2) an **eye-tracking group** where subjects' eye position and pupil size was captured by eye-tracking equipment. Within the control and intervention study group, there were ten subjects in the think-aloud group and ten subjects in the eye-tracking group. The entire study took between two and two and a half hour per subject.

	Study Group	Pre-Test Period	Feedback Period	Post-Test Period
Between Subjects Control Vs. Intervention	**Control** **10 Think-Aloud** **10 Eye-Tracking**	5 Cases	5 Cases Disease Info Only	5 Cases
	Intervention **10 Think-Aloud** **10 Eye-Tracking**	5 Cases	5 Cases Disease Info & Reasoning Strategies	5 Cases

Repeated Measure Pre vs. Post

Figure 7 Experimental Study Two Research Design

In accordance with IRB requirements, informed consent was obtained (consent forms are located in Appendix A). Subjects were given an overview of the computer system. Dr. Claudia Mello-Thoms explained the eye-tracking equipment to the subjects in the eye-tracking group. Think-aloud training was provided to the subjects in the think-aloud group. Subjects then assessed the clinical scenarios.

Experimental study two used a subset of the cases assessed in experimental study one (reference Appendix B for a list of the cases used in each study). Three criteria were used to determine the proper cases to be used in the second experimental study. **(1) Frequency of heuristic use**. Data from study one was assessed to determine the frequency of heuristic use. The cases that promoted *Anchoring and Adjustment* and *Confirmation Bias* were selected for use in study two. **(2) Case difficulty and diagnostic accuracy**. An element being studied in the second study is mode of thinking. The authors of the cases provided a 1 (easy) to 7 (hard) clinical difficulty level for each case. Cases of varying difficulty levels were selected for this study including cases on the lower and upper scale of the clinical difficulty level *in an attempt to promote* intuitive and analytical thinking. Medium level difficulty cases were also used, but special attention was paid to ensuring there were cases that may promote intuitive and analytical thought. In addition, diagnostic accuracy of the cases solved in the first experimental study was reviewed to determine which cases the study population seemed to easily diagnose, and those that seemed to be challenging. **(3) Presence of prominent disease differentiating clinical data**. During the case annotation process, the board-certified physicians were asked to identify cases that could commonly be misdiagnosed and/or contained specific data associated with the gold-standard diagnosis. For example, one of the cases could commonly be mis-diagnosed as Diabetes. This case had one specific data element that corresponds with the correct diagnosis of Acromegaly (coarse facial features).

6.4.5 Case Review Process

The clinical data for each case was presented to the subject over three screens (reference Figure 8). During experimental study one, the subject was not required to enter an initial diagnosis. Although, once all the clinical data had been revealed, they were required to enter a final diagnosis. This design was used to determine how frequently subjects formulated a hypothesis early in the diagnostic reasoning process. If an initial diagnosis was entered, and upon entry of the final diagnosis, the subject

was required to select the clinical data they used to arrive at the diagnosis by clicking the check-box adjacent to the data. It was conveyed to the subjects that they could enter multiple diseases at one time if all the clinical data selected were supportive of all the diseases entered. For example, entry of diseases A and B at the same time was permitted as long as the clinical data selected supported both diseases. There was no computer-based verification that data selected was supportive of all diseases entered. Subjects were also permitted to log more than one diagnosis on each screen. For example, the subject could enter disease A and B (associated with clinical data elements a, b and c) on screen 1, then enter disease C (associated with clinical disease d and e), also on screen one. Subjects were instructed to enter an initial diagnosis as soon as it came to their mind. If a disease entered on screen one was further supported by data displayed on screen two, subjects were instructed to reenter the disease and select the additional supportive data from screen two. Subjects had the ability to review previously displayed data.

The case review process was the same for experimental study two, with one additional requirement. If an initial diagnosis was not entered on screen one, the subject was required to enter a diagnosis on screen two. The 'Next Page' button (to move to page three) did not appear until an initial diagnosis was entered on page two. This forced the subject to *Anchor*, so that *Adjustment* and *Confirmation Bias* could be assessed. If the subject does not *Anchor*, these factors cannot be assessed.

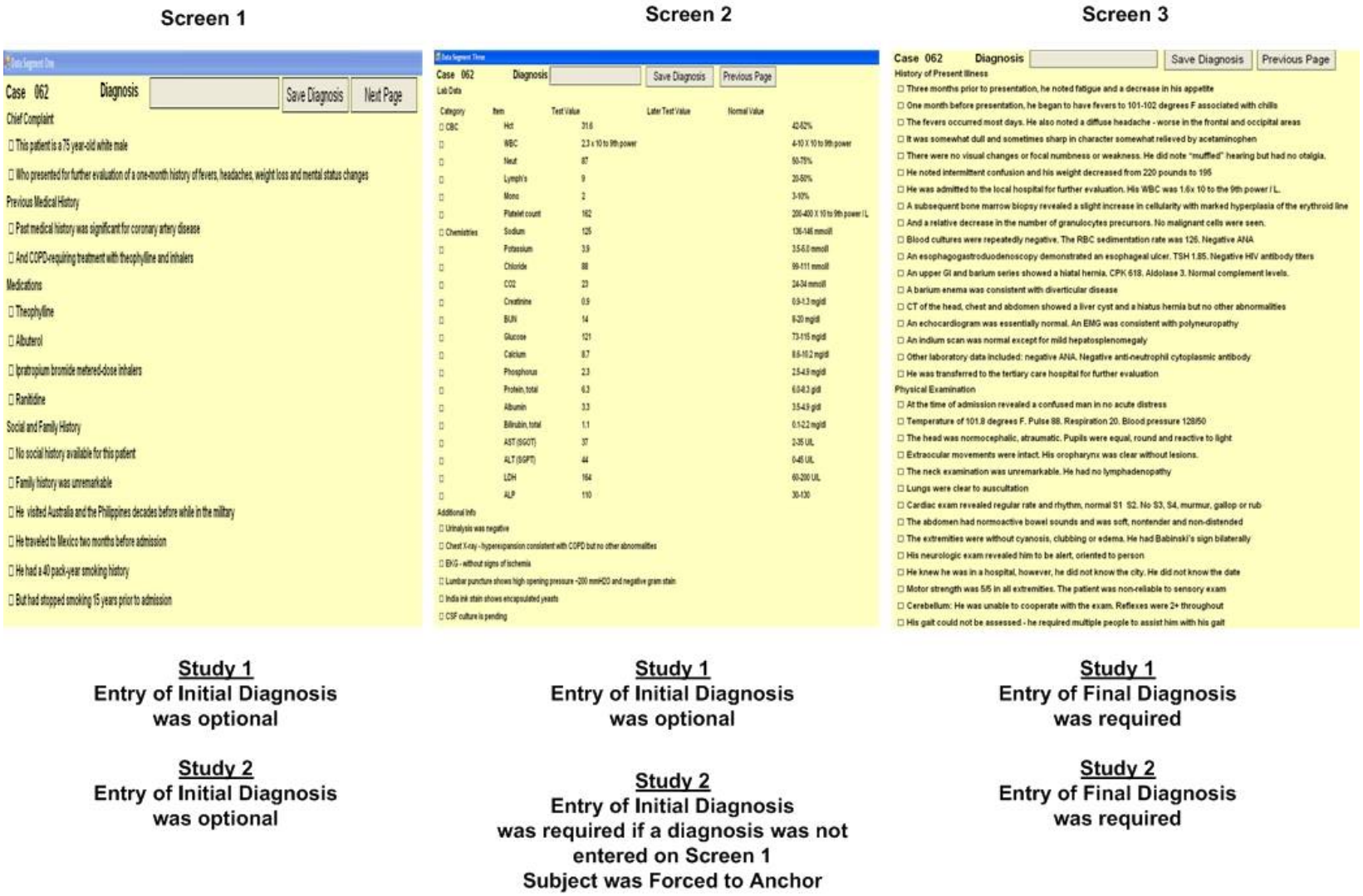


Figure 8 Clinical Review Process

6.4.6 Determining Subject's Mental Model

For each case, the subject was required to specify the data they used to arrive at their diagnosis. In the feedback period, the computer system used this data to derive the mental model the subject used while diagnosing the case. If the subject specified the same symptom for multiple diagnoses, it was assumed that the subject realized that symptom is associated with multiple diseases, i.e. it is an overlapping symptom for multiple diseases.

For the case in Figure 9, if the subject highlighted *back pain*, *x-ray of his right ribs showed a fracture*, and *degenerative joint disease*, and provided the diagnosis of Osteoporosis, the mental model shown in Figure 10 was automatically derived by the system. This mental model represents mapping a set of symptoms to a single disease.

Patient #092

Chief Complaint: This 65 year old white male has had right rib pain for 1 year. The pain has been severe for the past one and a half weeks.

History of Present Illness: The patient developed back pain in his lower right rib region about 1 year ago. It became so severe 1.5 weeks ago that he had difficulty maneuvering and sleeping. He then required morphine for relief. The pain was worse with deep breathing, coughing, sneezing and movement of his chest. A bone scan prior to admission showed multiple abnormal foci in an asymmetric pattern in bones including the ribs on the right and the left, ischium and tibia. An X-ray of his right ribs showed a fracture. There was no history of trauma. He had smoked 1 pack per day for 35 years but quit 2 years ago. He denied other symptoms including any related to his bowels or urinary system. He had not seen blood in his stool or urine. He had gained 15 pounds over the past month or so. Nausea had occurred once a week. Prior laboratory studies were said to show a normal serum leucine aminopeptidase, but an elevated alkaline phosphatase.

Previous Medical/Surgical History: Degenerative joint disease for which he had a right hip arthroplasty 3 months previously. Nephrolithiasis for which he had a lithotripsy. Crohn's disease for which he has a resection of his transverse colon 12 years before and resection of his small bowel 5 years before. Transurethral resection of his prostate on two occasions; no malignancy was reported. Hypertension and hypothyroidism.

Medications: Morphine, Tylenol #4, Nortriptyline, Zantac, Torecan, Azulfidine, Synthroid, Lasix, KCl, Folic Acid.

Social History: The patient is a retired tool and dye maker.

Physical Examination: The patient appeared to be in marked pain. Vital signs: BP 121/73, pulse 103/min, respirations 20/min. There was marked tenderness in area of back pain over the right lower rib cage. Tenderness was also present over the left rib cage. There were no nodules in his prostate. There were no other significant findings.

Figure 9 Sample of subject designating Data used to arrive at Diagnosis

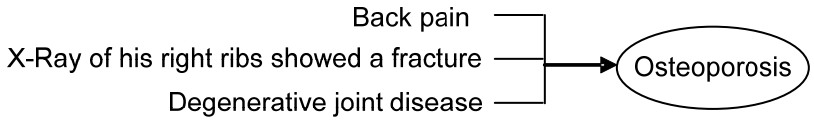


Figure 10 Symptom set maps to a single disease

For the same case, if the subject also entered a diagnosis of Osteomalacia and indicated they used the symptoms *back pain*, *x-ray of his right ribs showed a fracture*, *degenerative joint disease*, and *difficulty maneuvering*, since the symptoms of *bone fracture* and *inability to maneuver* are symptoms associated with both diseases, the subject mental model derived by the computer system is shown in Figure 11. This model corresponds to mapping a set of symptoms to multiple diseases.

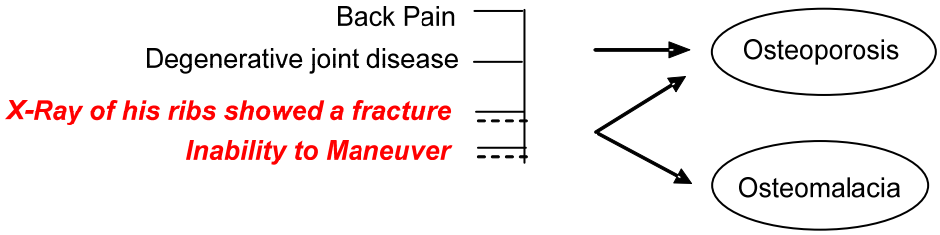


Figure 11 Symptom set maps to multiple diseases

6.4.7 Intervention

Feedback was used as the intervention technique for this research project. During the feedback period feedback was provided to the subject immediately after they logged the final diagnosis. Subjects in the intervention group received feedback regarding diagnostic reasoning strategies in association with the Mental Model Theory. This intervention was designed to induce the subject to "think about how they think", result in proper use of *Anchoring and Adjustment*, and reduce the commission of *Confirmation Bias* and diagnostic errors. In order to attempt to maintain a sense of equality between the control and intervention groups in terms of cognitive load and time-on-task, subjects in the control group were provided with general disease information regarding the correct diagnosis for each feedback cases. This information had no bearing on their diagnostic reasoning processes due to its association with the diagnosis for a case already assessed, and a diagnosis not associated with cases yet to be assessed.

6.4.8 Application of Mental Model Theory to Feedback

The principles of the Mental Model Theory used within this research focuses on mental model construction and reasoning with mental models. Specific MMT principles for each category are as follows:

Errors in Mental Model Construction:

1. Each mental model constructed represents a possibility compatible with the situation
2. Each model normally represents what is true but does not represent what is false

Errors in Reasoning with Mental Models:

3. When reasoning with mental models a conclusion is valid only if it holds in all models
4. A reasoning error occurs when one does not consider all applicable models
5. Reasoning errors occur because the models do not represent what is false

Feedback, dynamically generated by the computer-based case analysis system, contains information regarding:

1. The mental model constructed by the subject. The computer-based case analysis system, determined what mental model the subject constructed by analyzing the diagnosis the subject entered and the data they used to arrive at that diagnosis (reference section 6.4.5.1). Prior to conducting the experimental studies, three subjects were brought in to test the process and methodology. One subject thought aloud and wore eye-tracking equipment while using the computer system to assess the cases. A second subject was asked to only think-aloud while assessing the cases when using the computer-based case analysis system. A third subject did not think-aloud, but instead wore eye-tracking equipment while performing the task. Data from these sessions (pilot study) were carefully assessed to determine if (a) the data subjects indicated they use to formulate their diagnosis was consistent with the reasoning strategies they verbalized when thinking-aloud; and (b) if the subjects' mental model generated by the computer system was consistent with the data the subject used to arrive at the diagnosis.

2. Accurate representation of the evidence within the case and accurate mental model construction. The computer-based case analysis system assessed the subject's mental model to

determine if the model (a) accurately represents every possibility compatible with the situation (MMT Principle One); (b) represents what is true but does not represent what is false (MMT Principle Two).

3. Impact of reasoning strategies on diagnostic errors. Key principles of the MMT is that a conclusion is valid only when it holds true in all models (MMT Principle 3); a reasoning error occurs when reasoners do not consider all the applicable models (Prediction 2; section 4.3.1.2); and when falsity matters, fallacies occur (Prediction 3; section 4.3.1.2). The feedback addressed the principles associated with reasoning with mental models by including an explanation of how various reasoning strategies could result in a diagnostic error. Figure 5 depicts mental models associated with a case that has a diagnosis of *Acromegaly*. This case could be easily misdiagnosed as *Diabetes* if the subject does not realize that an elevated growth hormone is a critical finding associated with *Acromegaly*. Feedback provided for this case indicated that if one does not realize the 'elevated growth hormone' symptom is critical to correctly diagnosing the case with *Acromegaly*, misdiagnosing the case as *Diabetes* will occur.

4. Alternative mental models and reasoning strategies to consider Metacognition is a critical component of this dissertation research. Therefore, the feedback intervention was designed to cause physicians to think about how they think. The intervention not only provided information regarding mental model construction and current reasoning strategies, it also provided alternative mental models and reasoning strategies.

Some of the suggestions for alternative mental model construction were the inclusion of present and absent (denied) symptoms; consideration of symptom attributes (cough with mucus vs. dry hacking cough); and ensuring all applicable models were constructed. Some of the alternative reasoning strategies included ensuring all available information is considered before finalizing the diagnosis; not considering information refuting a stated diagnosis; considering all possible diagnoses rather than narrowing the diagnosis too quickly; considering not only what is true, but also what is false; considering not only present symptoms, but absent symptoms; thinking outside-the-box by considering non-medical information; and taking into account a situation where a disease may present in an unusual location within the human body.

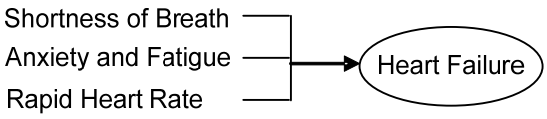
Figures 12 and 13 contain an example of feedback provided to subjects within the intervention group. Two screens were displayed - Figure 12 was displayed first to show the subject the reasoning strategies the computer system felt they were using. This was displayed to provide the subject with a

comparison of the reasoning strategies to consider that were provided on the second feedback screen (Figure 13). Only the last section (marked with an *) was provided to the control group subjects.

Assessment of Your Reasoning Strategies

Initial Diagnosis Strategy: You seem to be mapping a *Set of Findings* to a *Single Disease*

For an initial diagnosis, you provided a diagnosis of **Heat Failure**. You indicated that you based your diagnosis on **Shortness of Breath, Anxiety and Fatigue,** and **Rapid Heart Rate**. Your mental model representation is as follows



Final Diagnosis Strategy: You adjusted your strategy to mapping a *Set of Findings* to *Multiple Diseases*

Your final diagnosis was a differential of **Heart Failure** and **Pericardial Effusion**. You indicated that you based your diagnosis on **Shortness of Breath**, **Anxiety and Fatigue, Rapid Heart Rate**, **Difficulty Breathing** and **Chest Pain**. Your mental model representation is as follows

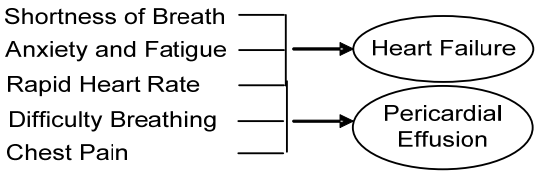


Figure 12 Feedback Example (Screen 1)

Alternative Model and Reasoning Strategy to Consider

Often times when diagnosing a patient, one has to *'think outside the box'*. This case is unique in that the disease is present in an unusual location within the human body. An Aortic Dissection commonly occurs in the renal area. In this case, the aortic tear is near the heart. Considering alternative locations of symptoms or a disease is essential to properly diagnosing this case. Failure to think outside the box and considering alternative locations may result in misdiagnosing this case.

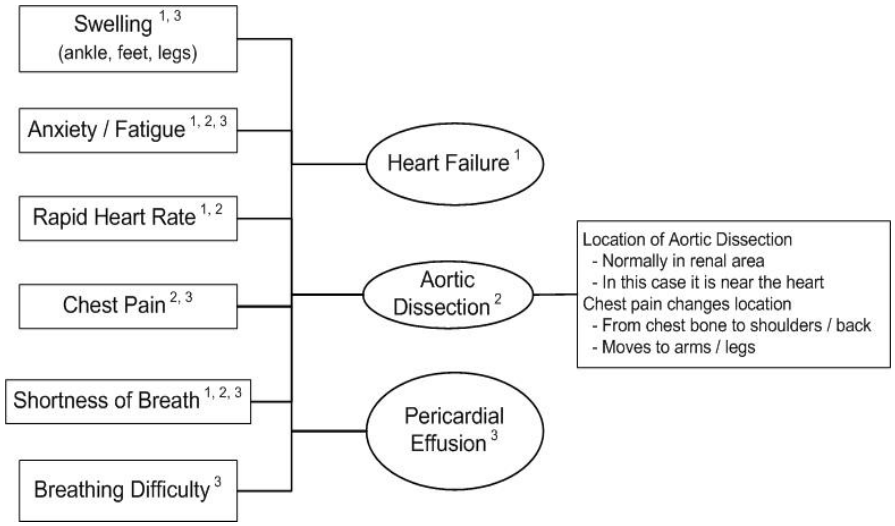
*** The correct diagnosis for this case is *Aortic Dissection***

An aortic dissection is a serious condition in which a tear develops in the inner layer of the aorta, the large blood vessel branching off the heart. Blood surges through this tear into the middle layer of the aorta, causing the inner and middle layers to separate (dissect). If the blood-filled channel ruptures through the outside aortic wall, aortic dissection is usually fatal.

Aortic dissection, also called dissecting aneurysm, is relatively uncommon. Anyone can develop the condition, but it most frequently occurs in men between 60 and 70 years of age. Symptoms of aortic dissection may mimic those of other diseases, often leading to delays in diagnosis. When an aortic dissection is detected early and treated promptly, the chance of survival greatly improves. Aortic dissection symptoms may be similar to those of other heart problems, such as a heart attack. Symptoms include Sudden severe chest or upper back pain, often described as a tearing, ripping or shearing sensation, that radiates to the neck or down the back; Loss of consciousness; Shortness of breath; Weakness or paralysis; Stroke; Sweating; High blood pressure; and Different pulse rates in each arm.

This case is easily misdiagnosed as it includes symptoms such as chest pain, cough and shortness of breath that are associated with many diseases. Some incorrect diagnoses may be Heart Failure and Pericarditis or

Pericardial Effusion. Both of theses diseases commonly present with lower extremity swelling, which is not present in this case. The figures below depict these scenarios.



The key to proper diagnosis of this case is to consider an uncommon or alternative location of a common diagnosis.

Figure 13 Feedback Example (Screen 2)

Table 11 demonstrates how the feedback incorporated each of the principles of the MMT.

Table 11 Use of MMT Principles within Feedback

Mental Model Theory Principle	Description of MMT to Feedback (Based on feedback section in Figures 12 and 13)
Principle 1: Each mental model constructed represents a possibility compatible with the situation	**Reasoning Strategy Assessment** – this feedback section designated the mental model used to arrive at the initial and final diagnosis. **Reasoning Strategy to Consider** - this section provided alternative reasoning strategies to consider and pointed out where the subject did not accurately represent the situation
Principle 2: Each model normally represents what is true but does not represent what is false **Principle 5:** Reasoning errors occur because the models do not represent what is false	**Reasoning Strategy Assessment** – this section addressed the representation of what is false primarily by addressing absent findings (the patient denies…) and/or test result values
Principle 3: When reasoning with mental models a conclusion is valid if it holds in all models **Principle 4:** A reasoning error occurs when reasoners do not consider all the applicable models	**Reasoning Strategies to Consider** – this feedback section included valid models applicable to the case that the subject did not consider. **Considering Absent Symptoms and Findings** – this feedback section will point out absent findings that the subject did not consider. The subject will not be able to arrive at the correct diagnosis unless they consider all models applicable to the case – a valid conclusion will not be reached if it does not hold in all the applicable models representing the situation.

6.4.9 Data Capture

6.4.9.1 Case Analysis Computer-Based Systems

For both experimental studies subjects used a computer-based system to assess the clinical cases. The system for experimental study one was an Internet-based system that consecutively displayed twenty-four (24) cases. The system used for experimental study two was installed locally on a desktop computer. These systems are described in section 6.4.5. The data captured by the computer-based systems includes the following data elements:

- Randomly assigned unique subject identification number
- Study group subject was assigned to – Control or Intervention group (study 2 only)
- Subject education level – fourth year medical student or post-graduate year (i.e., first, second, third or fourth year resident) (study 2 only)
- Method group subject was assigned to – eye-tracking or think-aloud group (study 2 only)
- Initial Diagnosis for each case (optional for study 1)
- Final Diagnosis for each case
- Unique data element identifier for each of the items the subject indicated they used to arrive at the diagnosis (initial and/or final) for each case
- Unique screen identifier for each screen display, in the order the screens were displayed to maintain an audit-trail of the screen flow for each case

6.4.9.2 Think-Aloud Protocols

In the second experimental study, think-aloud protocols were used as a secondary measure to assess the subjects' *Anchoring*, *Adjustment*, and *Confirmation Bias* behavior. Thinking-aloud is a technique used to elicit the inner thoughts or cognitive processes occurring during the performance of a task.[208] Elstein et al., Anderson, and Kuiper and Kassirer support the claim that speaking aloud during problem-solving is not significantly unlike problem-solving while remaining silent.[211-213] The think-aloud protocol technique has an advantage over simple observation as there is the potential to gain valuable insights into what the participant is thinking on the spot.[209] There are also limitations to the think-aloud protocol technique. Verbal protocol methods including thinking-aloud are designed to tap into certain types, but not all, of one's thought processes. Asking one to think-aloud may not result in

gathering sufficient information to analyze a problem without the use of probing.[210] In view of the limitations of thinking-aloud, probing is now commonly used to gather more information from participants; although probing may influence the reliability of the verbal protocol. Ericsson and Simon recommend that additional information be collected in the form of retrospective reports after the task to avoid any interruptions of task flow.[210]

"Does thinking-aloud have an effect on the cognitive load and mode of thinking?" is a question that has been debated within the literature. There is evidence indicating that when an individual verbalize their thoughts, they shift to a more deliberate mode of information processing.[130,218,219] The next logical question is does this more deliberate mode of information processing lead to better outcomes? There is evidence that has shown that when engaged in problem solving, performance is negatively affected when an individual verbalizes their thoughts. [130,218,219] However, this effect has been verified only with problems that require "insightful solutions" and not with more analytic problems.[130, 218, 219] Researchers who have studied this topic have indicated "verbalizing ones thoughts forces people to act in a deliberate mode and it tends to cut off access to tacit processes"; and "since recognition memory is highly dependent on the tacit system, problem solving requiring recognition memory can be less accurate if people are asked to make explicit use of the deliberate mode of information processing through verbalization".[218, 220, 221]

For this study the think-aloud protocol technique as described by Ericsson & Simon was used.[209, 210] During the second experimental study, a subset of subjects (N=20) were asked to think-aloud, verbalizing their thoughts as they assessed the cases in the pre- and post-test periods. Their words were audio-recorded, transcribed verbatim, and coded in accordance with Ericsson & Simon's think-aloud protocol coding technique. A total of two-hundred (200) cases were transcribed and coded. Two individuals were trained to perform the protocol coding. The coders consisted of a research assistant within the Department of Biomedical Informatics at the University of Pittsburgh, and an Assistant Vice-President at PNC Bank in Pittsburgh, PA who has served as a Software Development Manager within the Information Technology Division for 40 years. Both coders were provided with training on the proper technique of coding the transcripts that included providing the instructions found in Appendix C followed by a meeting where the trainer and coders worked through five cases together, discussing proper segmentation and coding techniques.

During the coding process the transcripts were segmented, dividing the subjects' words into phrases - each phrase representing a single idea. Once each case was segmented, the coders assigned operators and knowledge states at the point that the subject specified an initial diagnosis (Anchored),

extended or changed their initial diagnosis, mentioned the critical clinical data (if applicable), and specified the final diagnosis. The coders also noted the adjustment factor for each case (sufficient adjustment, insufficient adjustment, no adjustment necessary, adjusted away from the correct diagnosis). Table 12 contains the operators and knowledge states used in the coding process. Since the think-aloud protocols were used as a secondary measure to determine heuristic use, they were not exhaustively coded line-by-line.

Table 12 Think-Aloud Protocol Coding Schema

Item	Operator	Explanation
1	Specified Initial Diagnosis (Anchor)	The point at which the subject specified an initial diagnosis
2	Extended Initial Diagnosis	If the subject added a disease to their initial diagnosis
3	Changed Initial Diagnosis	If the subject removed a disease from their initial diagnosis
4	Used Critical Data	If the subject mentioned the case critical data
5	Specified Final Diagnosis	When the subject states the final diagnosis

Table 13 provides an example of a coded think-aloud protocol. This subject specifies an initial diagnosis at segment 9 where they say '*he could have diverticulitis*'. This is considered the point where the subject *Anchored*. At segment 32 the subject mentions the symptom of *pain in the right lower quadrant*, which is a piece of critical clinical data that should lead them to the correct diagnosis. At segment 90 the subject diagnosed the patient with Prostatitis, which is an incorrect diagnosis. The correct diagnosis of this case is Appendicitis. Since the subject anchored on a diagnosis of *Diverticulitis* and incorrectly diagnosed the case with Prostatitis, they *Insufficiently Adjusted*.

Table 13 Example of Think-Aloud Protocol Coding

Segment	Transcript Text	Operator	Knowledge State
1	48 year old white male		
2	with 2 days history of abdominal pain		
3	History of Bell's palsy		
4	Also has what seems like a long history of Microscopic hematuria		
5	Negative workup		
6	48 hours of abdominal pain		
7	So that could be a lot of different things		
8	Don't know if he is vomiting		
9	He could have diverticulitis	Specified Initial Diagnosis (Anchor Point)	Diverticulitis
29	Patient was previously healthy		

Segment	Transcript Text	Operator	Knowledge State
30	and then developed right and left lower pain		
31	which migrated to the right lower quadrant		
32	Hmmm…. Pain in RLQ – that's important	Used Critical Data	Right Lower Quad
90	I am going to say this patient has Prostatitis	Specified Final Diagnosis	Prostatitis
Insufficient Adjustment – correct diagnosis Appendicitis			

Once the transcripts were coded, every coded transcript was reviewed by the doctoral candidate. During the transcript analysis, cases were reviewed to determine where in the diagnostic process the subject specified an initial diagnosis (Anchored), where they specified a final diagnosis, how they adjusted, and if they committed Confirmation Bias. Since the think-aloud protocols were used as a secondary measure *only to determine heuristic use,* a review of diagnostic reasoning process did not extend beyond these items.

6.4.9.2.1 Think-Aloud Protocol Coding Reliability

In order to ensure the protocols were coded consistently, both coders independently coded the same sixty cases (30% of the two-hundred cases). After both individuals' coded ten (10) cases, these cases were reviewed for coding consistency, the differences were noted, an inter-rater reliability (Kappa) statistic was computed, and a meeting was held to discuss and resolve the coding differences. This process continued until a Kappa value of at least 0.75 was reached for each category assessed. Table 14 reflects the number of cases that were reviewed prior to reaching an acceptable Kappa value for each category. When both coders had completed coding of the initial 60 cases, the remaining 140 cases were divided between the coders who independently coded 70 of the remaining transcripts. Ten of these cases were assigned to both coders and reviewed to verify that the same inter-rater reliability level was maintained throughout the independent coding process.

Table 14 Inter-rater Reliability for Think-Aloud Protocol Coding

Category	Inter-rater Reliability (Kappa Value)	Number of Cases Coded To Reach Acceptable Kappa Value
Initial Coding of Sixty (60) Cases before Independent Coding		
Anchor Point / Initial Diagnosis	0.832	20
Final Diagnosis	0.757	30
Adjustment Value	0.841	30
Use of Critical Data	0.806	20

Category	Inter-rater Reliability (Kappa Value)	Number of Cases Coded To Reach Acceptable Kappa Value
Final Coding of Ten (10) Cases after Independent Coding		
Anchor Point / Initial Diagnosis	0.762	NA
Final Diagnosis	0.810	NA
Adjustment Value	0.789	NA
Use of Critical Data	0.837	NA

The coding completed by the protocol coders was compared to the data input by the subjects into the computer-based system. The percent of cases where the protocol coders identified the same initial (Anchor point) and final diagnosis as the data entered into the computer-based system was computed. The percent agreement was computed for all 200 cases, and then the percent agreement was computed after each fifty (50) cases to determine if the level of agreement increased as the protocol coders gained additional experience from coding more cases.

6.4.9.3 Think-Aloud Protocol Coding / Computer-System Correlation

Once all of the think-aloud protocols were coded, the level of agreement between the coding process and data entered into the computer-based system was performed. The initial and final diagnoses identified by the coder were compared to the diagnoses entered into the computer-system. The percentage of agreement for the two-hundred (200) cases was 20% for the initial diagnosis and 37% for the final diagnosis. The percent agreement was then calculated for the first 50 cases that the coders coded, and for each subsequent 50 cases to determine if there was a linear relationship between the coding experience level and the level of agreement. As the coders coded additional cases, the level of agreement between their coding and the data entered into the computer-based system increased to 86% agreement for Anchoring and 94% for designation of the final diagnosis (reference Table 15).

Table 15 Protocol Coding and Computer System Correlation

Coding / System Correlation – All Cases		
	Anchor Point	**Final Diagnosis**
	20%	37%
Coding / System Correlation Over Time		
% of Cases Coded	**Anchor Point**	**Final Diagnosis**

Coding / System Correlation – All Cases		
	Anchor Point	**Final Diagnosis**
	20%	37%
Coding / System Correlation Over Time		
% of Cases Coded	**Anchor Point**	**Final Diagnosis**
25% (50 cases)	20%	22%
50% (100 cases)	48%	46%
75% (150 cases)	72%	62%
100% (200 cases)	86%	94%

6.4.9.4 Eye-Tracking

Subjects participating in the second experimental study that were assigned to the eye-tracking group, wore eye-tracking equipment as they assessed cases to capture data associated with eye position and pupil size. Subjects' eye position was monitored using an ASL Model 6000 eye-tracking system (Applied Sciences Laboratories, Bedford, MA). This system uses an infrared beam to calculate the line of gaze by monitoring pupil and corneal reflection. The system also contains a magnetic head tracker that monitors the position of the head. The eye-tracking system determines the (x, y) coordinates of eye position in the display plane by integrating the collected eye and head positions. The system has an *accuracy* (measured by the difference between true eye position and computed eye position) of less than 1 degree of visual angle, and is capable of monitoring an image space covering 50 degrees of visual angle horizontally and 40 degrees vertically. In house-developed software from the laboratory of Dr. Mello-Thoms was used to calculate the (x, y) coordinates of all fixation clusters from the raw eye position data. A *fixation* is defined as a group of chronologically sequential raw eye position points clustered within a circle with a diameter of 0.5 degrees of visual angle.

For each subject, prior to assessing cases, the eye-tracking system was calibrated using a display showing a nine-point grid (referred to as the *calibration screen*). The subject was asked to hold their head as still as possible and look at each point on the grid when instructed by the experimenter. This calibration process allowed the eye-tracker to map the subject's eye position to the known coordinates of the nine (9) points on the screen. After the calibration period, the subject was able to move their heads freely (as long as they stay 20 inches away from the display) during case analysis.

For this research project, eye-tracking was used to determine if the subject (a) committed *Confirmation Bias*, and (b) operated in the Intuitive or Analytical mode of thinking in association with the Dual Process Theory. For these assessments, for each subject, each case, each case screen (each case consisted of three screens) and each line on the screen, the following data items were captured by the eye-tracking system:

- ***(x, y) Coordinate*** corresponding to the location on the screen where the subject is gazing
- ***Fixation Time*** defined as the total time the subject spends gazing at a particular location
- ***Pupil Size*** while gazing at each item, measured in millimeters

In order to carry out this analysis, the eye-tracking data needed to be synchronized with actions occurring during case analysis. As each new case was loaded, when an initial diagnosis entered and when the final diagnosis was entered by the subject when using the case analysis computer-based system, a date-time-stamp was recorded in a database. Date-time-stamps were also recorded as the subject switched screens to maintain an audit trail of the order the screens were reviewed. The values recorded by the computer-based system were compared to data logged by the eye-tracking system to determine the subjects' eye positions for each case.

6.4.10 Data Analysis

A consultation was held with a Biostatistician within the Department of Biomedical Informatics and from a Senior Statistician within the Clinical Translational Science Institute (CTSI) at the University of Pittsburgh to ensure that the statistical analysis plan for this research project was accurate and would provide the ability to explore the impact of each individual variable and multiple variable interactions on the outcomes. Data was analyzed using SPSS 15.0 for Windows; release 15.0.1 (22 November, 2006) and Microsoft ® Office excel 2003 (11.8328.8324) SP3.

6.4.10.1 Frequency of Heuristic Use (Research Question 1)

6.4.10.1.1 Anchoring and Adjustment

The frequency of *Anchoring and Adjustment* was determined by assessing the data collected from experimental study one. During this study, entering an initial diagnosis was optional; subjects were

not required to enter an initial diagnosis, they were only required to enter a final diagnosis. If the subject did enter an initial diagnosis, this was considered *Anchoring*. The number and percentage of cases that *Anchoring* occurred was computed to determine the frequency of use of this heuristic.

Adjustment was determined by comparing the initial diagnosis to the final diagnosis to determine the type of *Adjustment*. Various types of *Adjustment* were assessed including *Sufficient Adjustment* which is when the subject arrives at the correct final diagnosis after specifying an incorrect initial diagnosis; *Insufficient Adjustment* which is when the subject specifies an incorrect final diagnosis after specifying an incorrect initial diagnosis; and *No Adjustment Necessary* which is when the correct diagnosis is entered as the initial and final diagnosis. The number and percentage of cases where each type of *Adjustment* occurred were computed to determine the frequency of *Adjustment*.

The frequency of Anchoring and Adjustment was determined by case difficulty. A Kruskal-Wallis test was executed to determine if case difficulty statistically significantly impacted Anchoring and Adjustment.

6.4.10.1.2 Confirmation Bias

The frequency of *Confirmation Bias* was determined by assessing the data collected from experimental study one. For each subject a numerical *Confirmation Bias* score was derived by adding the number of elements supporting the diagnosis (from the data used to diagnose) to the number of items from the critical data feature set that was not used (ignored). The higher the *Confirmation Bias* score, the greater our belief that *Confirmation Bias* occurred. The number of critical items ignored, number of non-critical data items used, and the *Confirmation Bias* score was computed for each case. The mean score and the associated standard deviation were computed for each item for the cases within each difficulty category (easier, medium, and harder). A Kruskal-Wallis test was executed to determine if case difficulty statistically significantly impacted Confirmation Bias.

6.4.10.2 Impact of Heuristic Use on Diagnostic Accuracy (Research Question 2)

The data collected from the first experimental study was used to assess the impact of heuristic use on diagnostic accuracy.

6.4.10.2.1 Anchoring

The final diagnosis provided for each case was compared to the gold standard diagnosis provided by the case authors to determine if the case was diagnosed correctly or incorrectly. The number and percentage of cases diagnosed correctly and incorrectly when Anchoring occurred and did not occur was computed. In order to determine if Anchoring impacted diagnostic accuracy, a logistic regression statistical test was performed. A one-predictor logistic model was fit to the data to determine if Anchoring impacted diagnostic accuracy using the regression model **Diagnostic Accuracy ~ Anchoring**. Diagnostic accuracy (dichotomous variable; correct or incorrect) was the dependent variable and *Anchoring* (dichotomous variable; Anchor or Not Anchor) was the independent variable. The *p-value* (*p*) was assessed to determine if Anchoring was a statistically significant predictor of diagnostic accuracy.. The *Wald Chi-square Statistic* (Wald's χ^2) was assessed to determine the quantity each variable contributes to the outcome. The larger the statistic, the more substantially the variable contributed to the outcome. An *odds-ratio* (e^{β}) value which indicates the odds of the outcome occurring was also assessed to determine the odds of diagnostic accuracy occurring if *Anchoring* occurs.

6.4.10.2.2 Confirmation Bias

The frequency and percentage of diagnostic errors and the mean *Confirmation Bias* score was calculated for each case assessed. The impact that *Confirmation Bias* has on diagnostic accuracy was determined by performing a logistic regression test using the regression equation **Diagnostic Accuracy ~ *Confirmation Bias***. Diagnostic accuracy (dichotomous variable) was the dependent variable and Confirmation Bias (continuous variable) was the independent variable. As with Anchoring, the p-value, Wald Chi-square Statistic and odds-ratio was assessed to determine the impact *Confirmation Bias* has on diagnostic accuracy.

6.4.10.2.3 Additional Variables

The cases assessed during this research study were of varying difficulty levels. Case difficulty was assessed to determine if it had an impact on diagnostic accuracy. A logistic regression test was performed using the regression model **Diagnostic Accuracy ~ Case Difficulty**. Diagnostic accuracy (dichotomous variable) was the dependent variable and case difficulty was the independent variable. The p-value, Wald Chi-square Statistic and the odds-ratio were assessed to determine this variable's impact on diagnostic accuracy.

The variability of subjects' knowledge and skill level may have had an impact on their ability to correctly diagnose the cases. To determine if the subjects had an effect on diagnostic accuracy, a logistic regression test was performed using the regression model **Diagnostic Accuracy ~ Subject**. Diagnostic accuracy (dichotomous variable) was the dependent variable and the subject (categorical) variable was the independent variable. As with the other regression tests, p-value, Wald Chi-square Statistic and the odds-ratio were assessed to determine this variable's impact on diagnostic accuracy.

Assessing the impact of a combination of variables on diagnostic accuracy was also performed by running multi-predictor logistic regression tests using the **Diagnostic Accuracy ~ Anchoring * Case Difficulty * Subject** and **Confirmation Bias * Case Difficulty * Subject** models. Diagnostic accuracy is the dependent variable in both models; and the cognitive heuristic (Anchoring or Confirmation Bias), Case difficulty and Subject were the independent variables. The same variables were assessed to determine the combination of multiple variables on diagnostic accuracy.

6.4.10.3 Impact of Metacognitive Intervention (Research Question 3)

Data from the second experimental study was used to determine the impact of the metacognitive intervention on the post-test use of *Anchoring* and *Confirmation Bias*, and on diagnostic accuracy. The study consisted of a between subjects design where the differences between a control and intervention group were assessed. Subjects assessed cases over three periods including a pre-test, feedback period and post-test period. Data between the pre-test and post-test periods were compared to determine the impact of the intervention received during the second period.

6.4.10.3.1 Heuristic Use

From the data captured by the computer-based case analysis system, the frequency of *Adjustment* and *Confirmation Bias* was calculated during the pre-test and post-test periods. Data from the pre-test period was used as a baseline. Data from the post-test period was compared to the pre-test period to determine if there was a change in heuristic use after receiving the intervention. To support this quantitative assessment, think-aloud protocols were also assessed (qualitative assessment) to determine if the intervention altered diagnostic reasoning behaviors, and in turn impacted the use of *Anchoring and Adjustment* and *Confirmation Bias*. Since subjects were required to enter an initial diagnosis during the second experimental study, they were required to Anchor. Therefore, only the Adjustment factor was assessed by comparing the initial diagnosis to the final diagnosis to determine

if they *Sufficiently Adjusted*, *Insufficiently Adjusted*, *Adjusted Away from the Correct Diagnosis*, or there was no *Adjustment Necessary*. The mean frequency (and standard deviation) of each adjustment factor was determined for all subjects in the control and intervention groups during the pre-test and the post-test periods. The Confirmation Bias score (calculated as described in section 6.4.10.1.2 and 7.1.2) was calculated for each subject during the pre-test and post-test periods.

Confirmation Bias was also assessed using data captured by the eye-tracking system. The subjects' eye fixations were used to determine if they looked at data and/or ignored (i.e., did not fixate on) critical clinical data that should have led them to the correct diagnosis. In order for eye-tracking data to be properly collected, the clinical data reviewed by the subjects was displayed in bullet form. Each bullet consisted of one piece of evidence and was associated with an (x, y) coordinate. The bullets were placed at 0.5 degrees of visual angle apart (which corresponds to, in a typical 19", 1280 x 1024 pixels display, 25 pixels) to allow for accurate determination of where the subject is looking (reference Figure 14).

For each subject, for each clinical data screen they reviewed, the (x,y) coordinates captured by the eye-tracking system was assessed to determine the location (fixation) of the eyes and how long they dwelled at each item. Figure 15 is an example of a clinical data screen with a subject's fixations, dwell time and pupil size (used for mode of thinking analysis) while reviewing each line. The vertical lines were placed on the screen for the data analysis phase of the project; these lines were not present on the screens subjects reviewed during the study. Once it was determined where the subject gazed, and for how long they dwelled at each location, *Confirmation Bias* was determined by adding the number of lines containing non-critical data that were fixated and the number of lines containing critical data that were not fixated, and dividing that by the total number of lines on the screen. The following formula was used:

$$CB(ET) = \frac{\text{Number of non-critical lines fixated + Number of critical lines NOT fixated}}{\text{Total number of lines fixated and NOT fixated}}$$

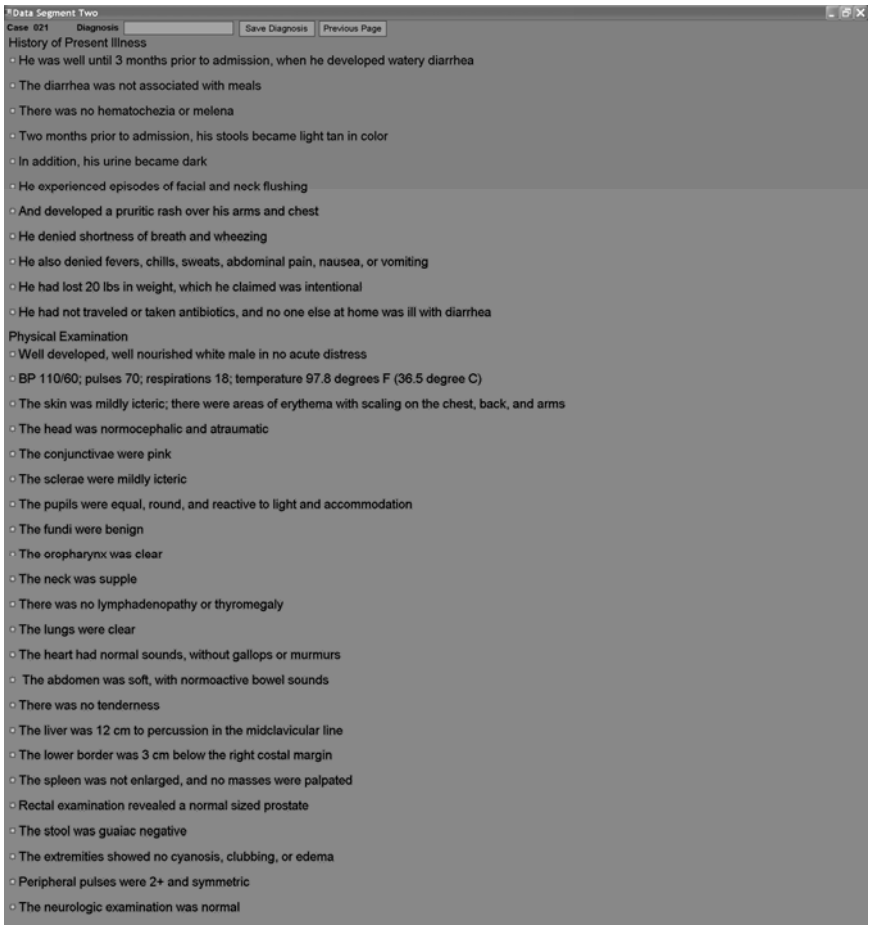


Figure 14 Case Analysis Screen

The percentage of cases where *Confirmation Bias* occurred was determined for the cases in which the subjects Sufficiently Adjusted, No Adjustment was Necessary, Insufficiently Adjusted, and when they Adjusted Away from the Correct Diagnosis. A Pearson Correlation test was ran to determine the correlation between the Confirmation Bias scores from the twenty think-aloud subjects and the twenty eye-tracking subjects.

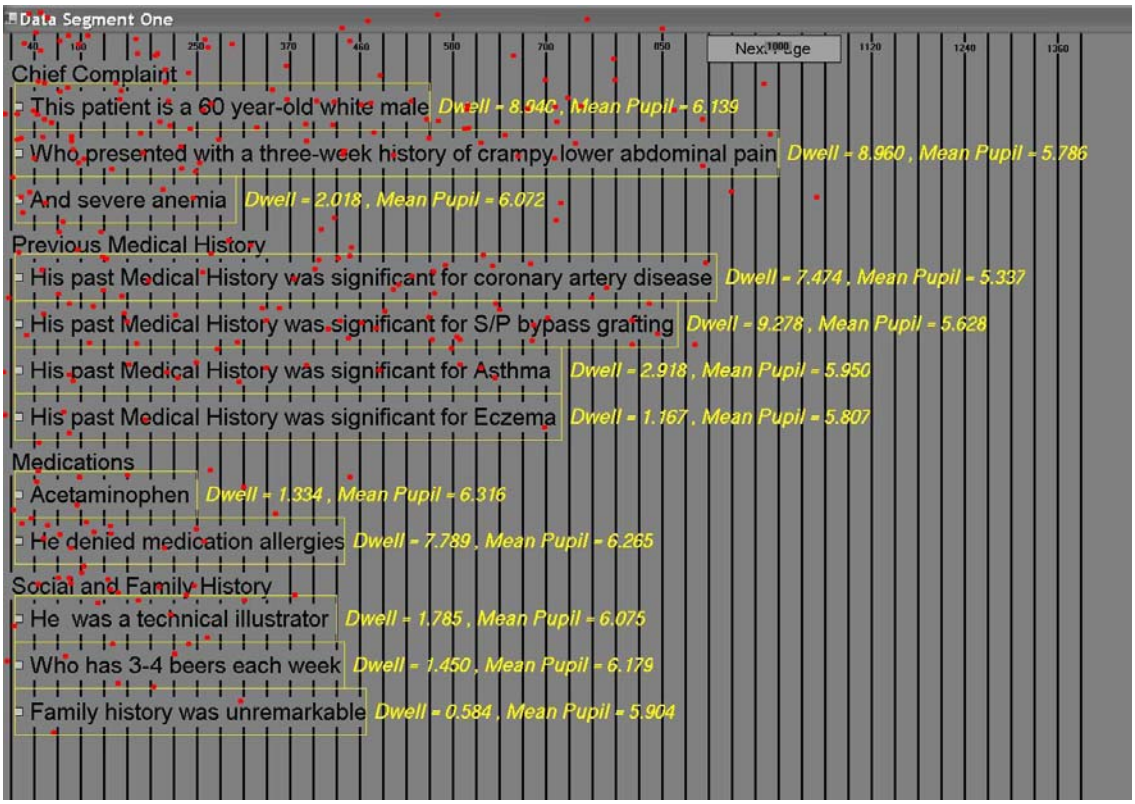


Figure 15 Subject Eye-Tracking Analysis

6.4.10.3.2 Diagnostic Accuracy

The impact of the intervention on diagnostic accuracy was also assessed by comparing the number of correct and incorrect diagnoses in the pre-test period to the post-test period for each study group (control and intervention).

The impact that *Study group, Feedback period, Subject, and Heuristic use* independently had on diagnostic accuracy was assessed by executing a single predictor logistic regression statistical test for each variable. The following models were used for this analysis:

Diagnostic Accuracy ~ Study Group

Diagnostic Accuracy ~ Feedback Period

Diagnostic Accuracy ~ Subject

Diagnostic Accuracy ~ Heuristic Use (Confirmation Bias only)

For each model, diagnostic accuracy (dichotomous variable) was the dependent variable, and study group (categorical variable), feedback period (categorical variable), subject (categorical variable) and

Confirmation Bias (continuous variable) was the independent variable for each test. For these single predictor regression tests, the regression coefficient, p-value, Wald Chi-square Statistic and the odds-ratio was assessed to determine if there was a statistically significant difference associated with the variables being tested.

A combination of the numerous variables within this research could have had an impact on diagnostic accuracy. A combination of the *Heuristic use*, *Subject* and *Feedback period* variables was tested using a multi-predictor logistic regression statistical test using the following model **Diagnostic Accuracy ~ Heuristic Use (Confirmation Bias) * Subject * Feedback Period**. As with the single predictor regression tests, the regression coefficient, p-value, Wald Chi-square Statistic and the odds-ratio was assessed to determine if there was a statistically significant difference associated with the variables being tested.

6.4.10.4 Mode of Thinking Analysis (Research Question 4)

Data from the second experimental study was used to determine if cognitive heuristic use and diagnostic accuracy differs within the Intuitive and Analytical modes of thinking. Eye-tracking data was assessed to determine if cognitive load and amount of time to arrive at a diagnosis (speed) could be used to determine the mode of thinking used as clinical scenarios were assessed. Once mode of thinking was established for each case, the frequency of diagnostic errors and heuristic use was assessed within mode of thinking.

6.4.10.4.1 Mode of Thinking Characteristics

The Dual Process Theory postulates that there are dual processes used in human information processing.[157,158] The dual modes of processing include processes that are *unconscious, rapid, automatic and high capacity* – commonly referred to the *Intuitive* mode of thinking; and processes that are *conscious, slow and deliberative*, commonly called the *Analytical* mode of thinking. The core components of these processes are consciousness, speed, and capacity (the number of processes that occur at the same time). Throughout decades of study by numerous researchers, additional attributes such as emotional attachment, reliability and error rate, have been used to describe the dual modes of processing.

Even though the Dual Process Theory has been extensively studied within disciplines including cognitive and social psychology, there is minimal empirical evidence of the application of this theory

within the domain of medicine. Editorial articles have been published detailing the manner in which it is *believed* physicians utilize the Dual Process Theory during clinical reasoning.[151,152,164] A topic of debate in such articles is the frequency of errors and heuristic use within each mode of thinking. To my knowledge, there is no empirical evidence regarding the frequency of heuristic use or diagnostic errors within each mode of thinking during diagnostic reasoning. This lack of empirical evidence was the primary reason the doctoral candidate chose to investigate application of this theory during diagnostic reasoning within her dissertation.

For this research it was **not** assumed that reasoning in the Intuitive mode of thinking leads to errors, and that reasoning in the Analytical mode of thinking leads to correct decisions. Instead, this evaluation was based on other characteristics of the modes of thinking including cognitive load and speed. For this analysis, it will be shown that cognitive load and speed **separately** are related to error rate (diagnostic accuracy); then that the **joint criteria** of cognitive load and speed are related to diagnostic accuracy. Showing this indicates that the cognitive load and speed together are an accurate predictor of mode of thinking.

6.4.10.4.2 Relationship between Speed and Diagnostic Accuracy

This purpose of this phase of analysis was to show that *speed* is related to *diagnostic accuracy* (error rate). The premises used for this analysis are (a) longer dwell time lead to slower processing; and (b) shorter dwell time lead to faster processing.

The eye-tracking data captured during this study includes the amount of time subjects dwelled at each line of clinical data while assessing clinical scenarios. The distribution of dwell time when subjects arrived at the correct and incorrect final diagnosis was assessed. For each diagnostic category, subjects' dwell time of elements within the critical data feature set, and dwell time of lines where critical data does not exist was determined. The frequency distribution of dwell time in each of the following categories was calculated:

- Correctly diagnosed case, median dwell time on items in the critical feature set
- Correctly diagnosed case, median dwell time on items *not* in the critical feature set
- Incorrectly diagnosed case, median dwell time on items in the critical feature set
- Incorrectly diagnosed case, median dwell time on items *not* in the critical feature set
- Total dwell time for all categories to arrive at the median dwell time

The p-value of a Mann-Whitney U-test was used to determine if there was a statistically significant difference between the median dwell time when cases were diagnosed quickly (correctly and incorrectly) versus when they were diagnosed in a slower manner (correctly and incorrectly).

6.4.10.4.3 Relationship between Cognitive Load and Diagnostic Accuracy

The second phase of the mode of thinking assessment included showing the relationship between cognitive load and diagnostic accuracy. Evidence exists indicating that "*pupil size* is an intrinsic measure of the *level of information processing being carried out by the brain*,"[203] and is directly correlated to *mental capacity utilization and overall memory load*.[203,204,205] Evidence also exists indicating *cognitive load* corresponds to the *mode of thinking* in association with the Dual Process Theory.[157,158] Since *pupil size* is an indicator of *cognitive load*, and *cognitive load* directly correlates to *mode of thinking*, we sought to show that *pupil size* can be used as a factor (but not the sole factor) to categorize (or classify) *mode of thinking*.

The behavior of the subjects' pupils was reviewed to determine if a mean pupil size could be associated with various speed / accuracy combinations. Mean pupil sizes were calculated in the following categories:

- Mean pupil size, case diagnosed correctly, within the fast dwell time / Intuitive mode
- Mean pupil size, case diagnosed correctly, within the slow dwell time / Analytical mode
- Mean pupil size, case diagnosed incorrectly, within the fast dwell time / Intuitive mode
- Mean pupil size, case diagnosed incorrectly, within the slow dwell time / Analytical mode
- Mean pupil size, over all categories

A Mann-Whitney U test was used to determine statistical significance of the mean pupil size over all categories (cases diagnosed correctly or incorrectly, within a fast or slow dwell time).

6.4.10.4.4 Relationship between Speed and Cognitive Load (Joint Criteria) and Diagnostic Accuracy

Once it was shown that speed and cognitive load each were related with diagnostic accuracy, an analysis was performed to determine if it could be confidently stated that the combination of speed and cognitive load (joint criteria) was related to diagnostic accuracy. During this analysis the mean pupil size was investigated for:

- Cases diagnosed incorrectly within a short period of time
- Cases diagnosed correctly within a longer period of time

Once the mean pupil size was assessed to determine if it is related to diagnostic accuracy for cases assessed quickly and slowly, the statistical significance of the pupil size was assessed. A one factor Analysis of Variance (ANOVA) was performed with mean pupil size as the dependent variable and diagnostic accuracy and dwell time as the independent variables.

6.4.10.4.5 Mode of Thinking Used During Case Analysis

Once it had been established that the joint criteria of *speed* and *cognitive load* can be used to categorize (or classify) mode of thinking, a mode of thinking was assigned to every line of clinical data within each case assessed. Only the lines with a median dwell time greater than zero were included in the analysis. Each line was characterized as one of the following:

- **Pure Intuitive Line**– when the median dwell time (speed) and mean pupil size (cognitive load) fell within the thresholds associated with the Intuitive mode of thinking. That is, median dwell time was $\leq$ 1.902 seconds (the speed threshold established in section 6.4.10.1.1) and the mean pupil size was $\leq$ 5.571 millimeters (the cognitive load threshold established in section 6.4.10.1.2).
- **Pure Analytical Line** – when the median dwell time (speed) and mean pupil size (cognitive load) fell within the thresholds associated with the Analytical mode of thinking. That is, median dwell time was > 1.902 seconds and the mean pupil size > 5.571 millimeters.
- **Mixed Mode Line** – when the median dwell time was $\leq$ 1.902 seconds **and** the mean pupil size was > 5.571 millimeters **OR** when the median dwell time was > 1.902 seconds **and** the mean pupil size was $\leq$ 5.571 millimeters

Once a mode of thinking was assigned to each line within the case, a mode of thinking was assigned at the case level by examining all the data elements assessed for the case. If all the data elements were assessed using the 'Pure Intuitive' mode, the case was classified as a 'Pure Intuitive' case. If all the data elements were assessed using the 'Pure Analytical' mode, the case was classified as a 'Pure Analytical' case. There were no cases where **all** the clinical data elements in the case were assessed in the 'Pure Intuitive' or 'Pure Analytical' mode. Therefore, no cases could be classified as 'Pure Intuitive' or 'Pure Analytical'. Subjects commonly switched from one mode to the other throughout the case assessment process. There were several cases where the clinical data elements were assessed using either the 'Pure Intuitive' or the 'Mixed Mode' (no lines in the case were assessed with the 'Pure Analytical' category) – these cases have been classified as '**Intuitive / Mixed**'. There were also

several cases where the clinical data elements were assessed using either the 'Pure Analytical' or the 'Mixed Mode' (no lines in the case were assessed with the 'Pure Intuitive' category) – these cases have been classified as '**Analytical / Mixed**'. The cases where all three modes were present – these cases have been classified as '**Intuitive / Analytical / Mixed**'. The following are definitions of the modes of thinking assigned at the case level.

- **Pure Intuitive Case** – every line of the case was assessed with a 'Pure Intuitive' mode of thinking
- **Pure Analytical Case** – every line of the case was assessed with a 'Pure Analytical' mode of thinking
- **Intuitive / Mixed Case** – lines of the case were assessed with either a 'Pure Intuitive' or 'Mixed Mode' thinking
- **Analytical / Mixed Case** – lines of the case were assessed with either a 'Pure Analytical' or 'Mixed Mode' of thinking
- **Mixed Mode Case** – lines of the case were assessed with either a 'Pure Intuitive', 'Pure Analytical' or 'Mixed Mode' of thinking (all three classifications were used)

6.4.10.4.6 Heuristic Use and Diagnostic Accuracy within Mode of Thinking

6.4.10.4.6.1 Clinical Data Assessment within Mode of Thinking

In order to provide a high-level view of mode of thinking statistics for the subject population used in this research, the following descriptive statistics were calculated for across all cases solved by all the subjects:

- **Average number of clinical data lines assessed** in each mode of thinking
- **Median dwell time** spent assessing clinical data using each mode of thinking
- **Mean pupil size** associated with clinical data assessment within each mode of thinking

6.4.10.4.6.2 Frequency of Heuristic Use within Mode of Thinking

The number of times each *Adjustment factor* occurred while processing cases within each mode of thinking was determined. The adjustment factors assessed include *Sufficient Adjustment*, *Insufficient Adjustment*, and *No Adjustment Necessary* (as defined in section 6.4.10.1.1). A Kruskal-Wallis statistical test was ran to determine if there are statistically significant differences in the frequencies of

each *Adjustment factor* within the modes of thinking (Pure Intuitive, Pure Analytical and Mixed Mode).

For each mode of thinking, the *Confirmation Bias* score was calculated using the formula listed in section 6.4.10.3.1. A Kruskal-Wallis statistical test was ran to determine if there are statistically significant differences in the *Confirmation Bias* scores across the modes of thinking (Pure Intuitive, Pure Analytical and Mixed Mode).

6.4.10.4.6.3 Frequency of Diagnostic Errors within Mode of Thinking

The number of times a case was diagnosed correctly and incorrectly was calculated for each mode of thinking by comparing the final diagnosis provided by the subject to the gold standard diagnosis. A Kruskal-Wallis statistical test was ran to determine if diagnostic accuracy was statistically significantly different for the modes of thinking.

7 STUDY RESULTS

This research project consisted of a pilot study and two experimental studies. The pilot study (described in section 6.4.8; item 1), consisting of 3 subjects, was conducted to test the methods to be used in the experimental studies. Once the methods were tested, the first experimental study took place, followed by the second study. A total of 107 subjects participated in the two experimental studies; 67 in the first study (sixty-two fourth-year medical students and five residents; the resident post-graduate year was not captured for study one); and 40 (thirty-eight fourth-year medical students and two third-year residents) in the second study. The computer-based systems used the studies, as well as several data analysis systems were designed and developed by the doctoral candidate.

7.1 FREQUENCY OF HEURISTIC USE

Research Question 1: What is the frequency of use of the cognitive heuristic *Anchoring and Adjustment* and the cognitive bias *Confirmation Bias* during diagnostic reasoning?

The data from the first experimental study was analyzed to answer this research question.

7.1.1 Anchoring and Adjustment

Using a computer-based case analysis system, clinical information was presented to the subject via three screens. When reviewing data on the first two screens, subjects had the opportunity to enter an initial diagnosis; however, they were not required to do so. On the third screen of data, the subject was required to enter a final diagnosis. When entering an initial and/or a final diagnosis, subjects were

required to select the items they used to arrive at their diagnosis by checking the check-box adjacent to the data item (referred to through-out this document as *data used to diagnose*).

If the subject entered an initial diagnosis, this was considered *Anchoring*. This diagnosis was compared to the final diagnosis to determine how the subject *Adjusted*. Several categories of adjustment were assessed including (a) *Sufficient Adjustment* which is when the subject arrives at the correct final diagnosis (regardless of the correctness of the initial diagnosis); (b) *Insufficient Adjustment* which is when the subject specifies an incorrect final diagnosis (regardless of the correctness of the initial diagnosis); and (c) *No Adjustment Necessary* which is when the correct diagnosis is entered as the initial and final diagnosis.

Using these measures, the overall frequency of *Anchoring* and *Adjustment* was calculated and is reported in Table 16. For a total of 1,577 cases, solved by 67 subjects (three subjects did not solve all 24 cases), subjects did not Anchor in 102 (6.47%) of the cases, and *Anchored* in 1475 (93.53%) cases. When Anchoring occurred, subjects *Sufficiently Adjusted* (arrived at the correct final diagnosis) in 199 (13.49%) cases. *Insufficient Adjustment* (a diagnostic error occurred) occurred in 1036 (70.24%) of the cases. *No Adjustment Necessary*, which is when the subject *Anchored* on the correct diagnosis, and specified that diagnosis as the final diagnosis on 240 (16.27%) cases. Table 14 and Figures 17 and 18 provide *Anchoring and Adjustment* behavior by subject. As can be seen, *Anchoring and Adjustment* behavior is consistent across all subjects.

Table 16 Frequency of Anchoring and Adjustment

Category	Number of Cases	Percentage of Cases
No Anchoring	102	6.47% *
Anchoring Occurred	1475	93.53% *
Sufficient Adjustment	199	13.49% **
Insufficient Adjustment	1036	70.24% **
No Adjustment Necessary	240	16.27% **

* Based on Number of Cases Solved (1,577)
** Based on Number of Cases where Anchoring Occurred (1,475)

From these descriptive statistics it is clear that *Anchoring* on an initial diagnosis prior to reviewing all the clinical data available is a common trait for this study population. Many subjects logged an initial diagnosis prior to moving past the initial screen which contained only the chief complaint and the history of present illness. The number of cases where *Anchoring* did not occur was minimal. For this group of subjects, once an *Anchor* point was established, for 70% of the cases, they did not reach the

correct final diagnosis (insufficient adjustment); the case was misdiagnosed. For this subject population, correctly diagnosing the case was infrequent.

Figure 16 shows a breakdown of *Anchoring and Adjustment* by case difficulty. The percentage of cases where *Anchoring* occurred is nearly evenly distributed across the easy, medium and hard cases; as is the cases where the subject *Sufficiently and Insufficiently Adjusted*. The percent of cases where *No Adjustment was Necessary* is noticeably higher for the easy cases than the medium and harder cases. The Kruskal-Wallis test designates Adjustment is significantly affected by case difficulty (H= 67.584, df=2, p<=0.01). However, Anchoring is not significantly affected by case difficulty (H=2.178, df=2, p=0.349).

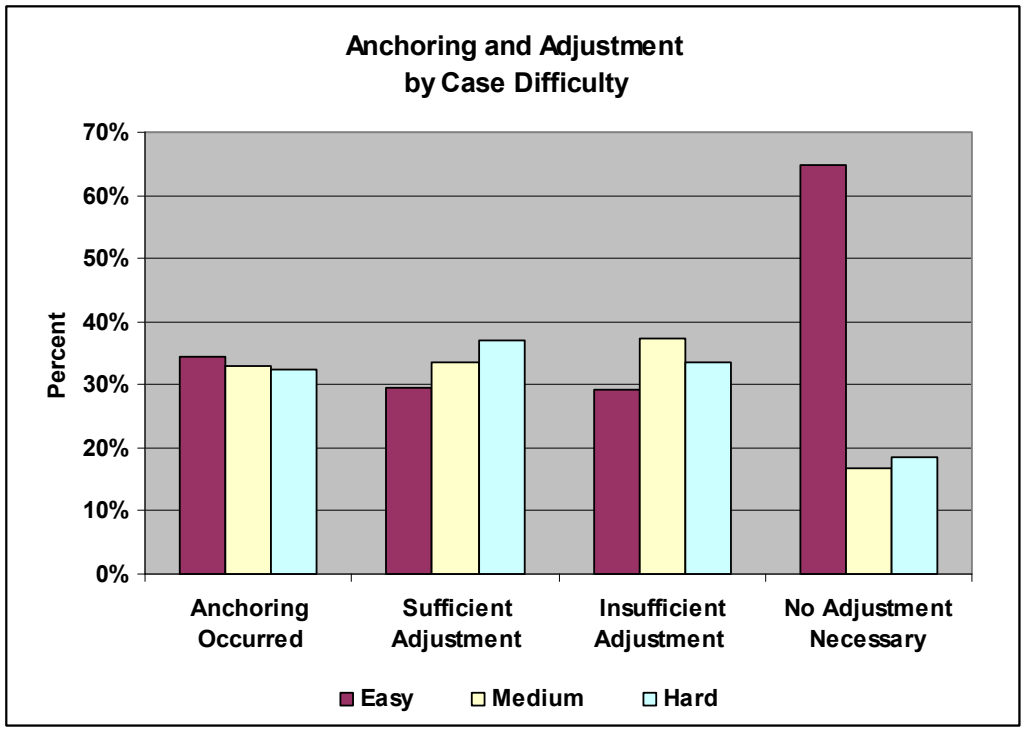


Figure 16 Anchoring and Adjustment by Case Difficulty

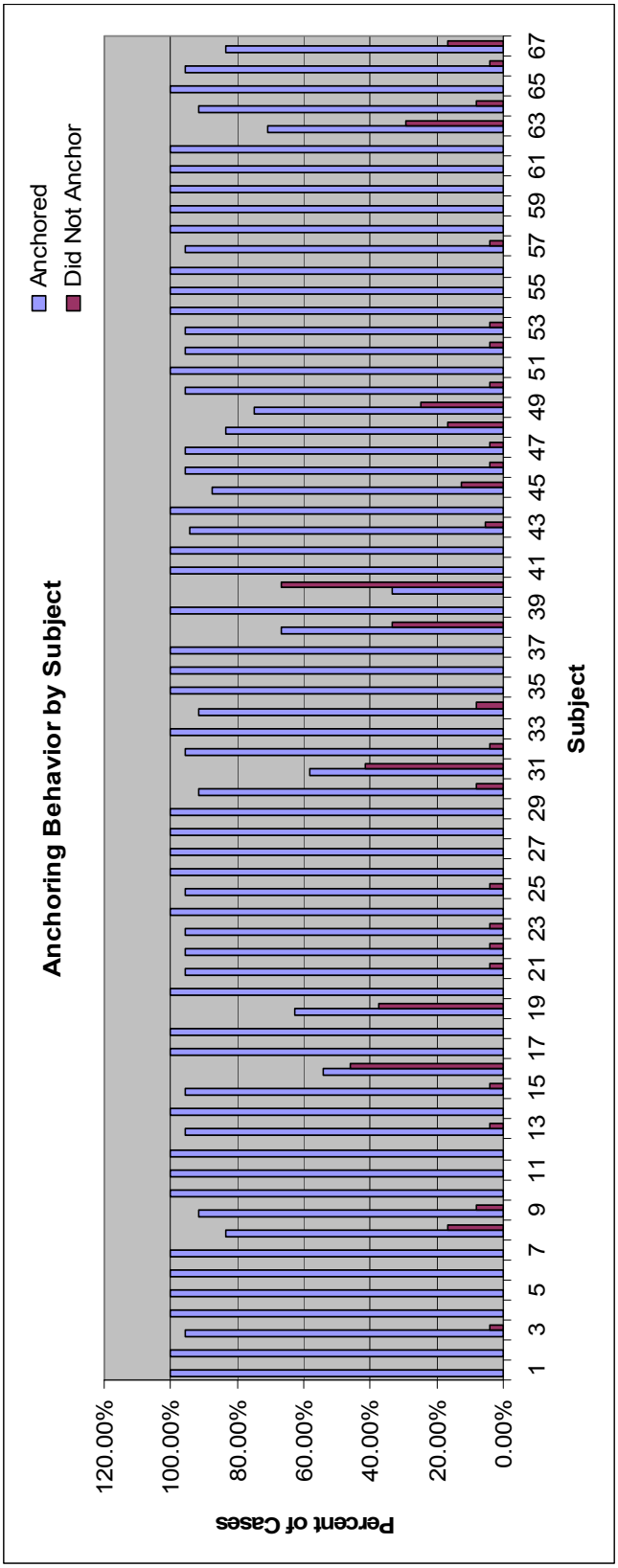


Figure 17 Anchoring Behavior by Subject

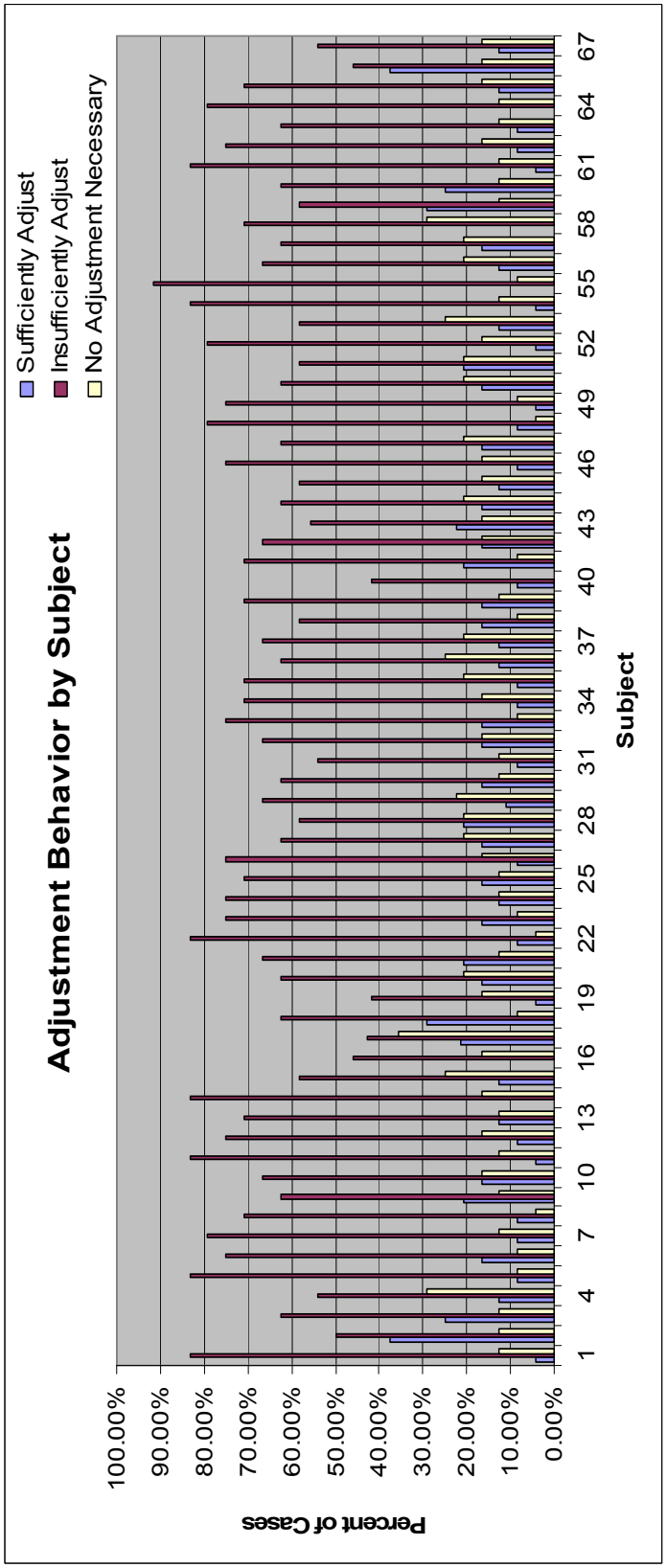


Figure 18 Adjustment Behavior by Subject

7.1.2 Confirmation Bias

Based on the definition of *Confirmation Bias* (reference section 6.3), commission of this bias is associated with two factors: (1) seeking evidence to confirm a diagnosis (hypothesis) and (2) ignoring evidence that could disconfirm a diagnosis and/or lead to an alternative diagnosis.

Once subjects entered a diagnosis, they designated the clinical evidence they used to arrive at that diagnosis (throughout this document this data is referred to as *data used to diagnose*). For this research study, it has been assumed that the *data used to diagnose* was deemed by the subject to support their diagnosis. Even though subjects were instructed to only select evidence that supports their diagnosis, this data may include evidence that was used to rule-out alternative diagnoses.

As discussed in section 6.4.3, for each of the twenty-four (24) cases board-certified physicians identified one or more critical data elements associated with the correct diagnosis (referred to as the *critical data feature set*). This data was used to determine if the subject ignored evidence that would refute their diagnosis and/or should lead them to an alternative diagnosis.

For each subject a numerical *Confirmation Bias score* was derived by (a) adding the number of elements the subject indicated they used to arrive at their diagnosis to (b) the number of items from the *critical data feature set* that was not used (ignored). Items in the *critical data feature set* were not included in part (a). For example, if the subject indicated that five items supported their diagnosis, two of which were part of the critical data element set, and they did not use (ignored) two of the five critical data items, the *Confirmation Bias score* was five. The higher the *Confirmation Bias score*, the greater our belief that *Confirmation Bias* occurred. Table 16 shows the average number of elements subjects used and did not use when diagnosing each case, along with the derived *Confirmation Bias score*. For each case, the scores were scaled by dividing the score by the number of clinical data elements that displayed on the computer screen. The adjusted *Confirmation Bias* score is displayed in the rightmost column of Table 17.

Table 17 Confirmation Bias Score
(Average per Case)

Case ID	Case Difficulty	Critical Evidence Items Ignored	Number of Critical Evidence Items in Case	Number of Items Used to Arrive at Diagnosis	Confirmation Bias Score	Number of Clinical Data Lines Within Case	Adjusted Confirmation Bias Score
52	Easier	1.97	2	8.13	10.10	59	0.17
91	Easier	3.31	4	5.19	8.51	50	0.17
42	Easier	0.60	2	5.89	6.40	55	0.12

Case ID	Case Difficulty	Critical Evidence Items Ignored	Number of Critical Evidence Items in Case	Number of Items Used to Arrive at Diagnosis	Confirmation Bias Score	Number of Clinical Data Lines Within Case	Adjusted Confirmation Bias Score
32	Easier	1.57	2	8.06	9.63	68	0.14
133	Easier	1.19	2	15.07	16.27	98	0.17
102	Easier	0.97	3	2.69	0.97	34	0.03
	Mean **SD**	**1.6** **1.0**	**2.5** **0.8**	**7.5** **4.2**	**8.6** **5.0**	**60.1** **21.5**	**0.13** **0.05**
62	Medium	0.25	2	7.10	7.36	99	0.07
21	Medium	0.18	1	11.40	11.58	82	0.14
93	Medium	1.40	2	1.85	3.25	62	0.05
122	Medium	1.85	3	3.36	5.21	55	0.09
83	Medium	0.15	1	17.37	17.50	83	0.21
181	Medium	0.72	2	11.21	11.93	77	0.15
72	Medium	1.79	4	12.29	13.90	59	0.24
92	Medium	1.64	3	7.39	8.70	71	0.12
53	Medium	1.21	3	7.66	8.52	72	0.12
12	Medium	2.30	3	5.23	7.30	78	0.09
	Mean **SD**	**1.1** **0.8**	**2.4** **1.0**	**8.5** **4.6**	**9.5** **4.2**	**73.8** **13.1**	**0.13** **0.06**
82	Harder	0.84	1	12.48	13.44	98	0.14
103	Harder	0.36	1	9.00	9.43	89	0.11
11	Harder	0.89	1	7.39	8.06	58	0.14
1	Harder	0.94	1	7.72	8.43	75	0.11
31	Harder	1.54	3	18.45	19.99	91	0.22
123	Harder	1.63	4	1.97	3.54	86	0.04
112	Harder	2.93	4	14.52	16.79	98	0.17
291	Harder	1.85	2	13.11	14.27	81	0.18
	Mean **SD**	**1.4** **0.8**	**2.1** **1.4**	**10.6** **5.1**	**11.7** **5.3**	**84.5** **13.3**	**0.13** **0.05**

Reviewing the means for each category across the three case difficulty levels, it is interesting to see that the (non-adjusted) *Confirmation Bias* scores increase as the case difficulty increases. This seemingly linear relationship implies that more challenging cases *may* induce commission of *Confirmation Bias* to a greater degree than easier cases. The normalized scores reveal a wide range of scores in each difficulty level; which makes it difficult to establish a relationship between the degree of *Confirmation Bias* and case difficulty. For all case difficulty levels, the mean adjusted *Confirmation Bias* scores are nearly equivalent.

As Table 18 shows, across the case difficulty levels, between 33% (harder difficulty) and 54% (medium difficulty) of the critical data items associated with the correct diagnosis were used to arrive

at the diagnosis; leaving between 46% (medium difficulty) and 67% (harder difficulty) of the critical data elements that were not used to arrive at the diagnosis. Of the total number of clinical data lines in the case, the percentage of non-critical data elements used to arrive at the diagnosis range between 11.5% (medium difficulty) and 12.5% (easy and hard difficulty); indicating that approximately 88% of the non-critical data elements in the case are being ignored (at least the subject did not select those items as data they used to arrive at their diagnosis). It is obvious that this subject population is using a small portion of the available clinical data to arrive at their diagnosis. The Kruskal-Wallis Test indicates that case difficulty did not have a statistically significant impact on *Confirmation Bias* (H=0.162, df=2, p=0.927).

Table 18 Percentage of Data Elements Used and Not Used

Case Difficulty Level	Critical Data		Non-Critical Data	
	Used	Not Used	Used	Not Used
Easier	45%	55%	12.5%	88%
Medium	54%	46%	11.5%	87%
Harder	33%	67%	12.5%	88%

7.2 IMPACT OF HEURISTIC USE ON DIAGNOSTIC ACCURACY

Research Question 2: Does the use of *Anchoring and Adjustment* and/or *Confirmation Bias* impact diagnostic accuracy?

For the cases assessed during the first experimental study, the final diagnosis provided by the subject was evaluated to determine if the case was diagnosed correctly or incorrectly. To determine diagnostic accuracy, the subjects' diagnosis was compared to the gold-standard diagnosis (the diagnosis provided by the case authors).

7.2.1 *Anchoring* and Diagnostic Accuracy

Table 19 details the frequency of diagnostic errors that occurred when the subject *Anchored* on an initial diagnosis. Of the 1,471 cases where *Anchoring* occurred, the subjects correctly diagnosed 439 (29.84%) cases, and incorrectly diagnosed 1,032 (70.16%) cases. When *Anchoring* did not occur, 22 (22.75%) of the cases were diagnosed correctly; and 84 (79.25%) were diagnosed incorrectly.

Table 19 Diagnostic Accuracy - Anchoring vs. No Anchoring

Diagnostic Accuracy			
	Correct	**Incorrect**	**Total**
Anchor	439 (29.84%)	1032 (70.16%)	1471 (93.28%)
No Anchor	22 (20.75%)	84 (79.25%)	106 (6.72%)
Total	461 (29.23%)	1116 (70.77%)	1577 (100%)

* Percentages are based on row totals

In order to determine if Anchoring has a statistically significant impact on diagnostic accuracy, a logistic regression was performed. A independent-predictor logistic model was fit to the data to test the research hypothesis regarding the relationship between diagnostic accuracy and *Anchoring*. The regression equation of the result is as follows

Diagnostic Accuracy ~ Anchoring

Predicted logit of (Diagnostic Accuracy) = 0.370 + (0.485) * *Anchoring*

For the *Anchoring* factor, the Wald statistic was 3.882 (p=0.049), indicating that *Anchoring* was a significant predictor of Diagnostic Accuracy (Table 20). Anchoring has a positive effect on Diagnostic Accuracy (Odds Ratio = 0.616), indicating that subjects who *Anchor* are more likely to get the diagnosis correct, when compared to subjects who do not *Anchor*.

Table 20 Impact of Anchoring on Diagnostic Accuracy

Predictor Statistical Tests						
Predictor Variable	**β**	**SE β**	**Wald's χ^2**	***Df***	***p***	**e^{β} (Odds Ratio)**
Intercept Only	1.340	0.239	31.294	1	0.000	3.818
Anchoring	0.485	0.246	3.882	1	0.049	0.616

7.2.2 Confirmation Bias and Diagnostic Accuracy

An analysis was performed to determine the frequency of diagnostic error when *Confirmation Bias* occurs. Diagnostic accuracy and the average *Confirmation Bias* score (across all subjects) for each case is shown in Table 21.

The percentage of cases diagnosed correctly and incorrectly varies across the cases, as does the *Confirmation Bias score*. There are times when a higher *Confirmation Bias* score coincides with a large percent of the cases being diagnosed incorrectly. However, there are times, such as case 031, where a higher score coincides with a nearly 50-50 split in terms of diagnostic accuracy.

Table 21 Mean Confirmation Bias Score and Diagnostic Accuracy

Case ID	Percent Cases Diagnosed Correctly	Percent Cases Diagnosed Incorrectly	Average Confirmation Bias Score	Standard Deviation	Number of Clinical Data Lines in Case	Adjusted Confirmation Bias Score
052	93.9%	6.1%	10.10	4.57	59	0.17
091	58.2%	41.8%	8.51	3.53	50	0.17
042	50.7%	49.3%	6.40	6.26	55	0.12
032	4.5%	95.5%	9.63	6.41	68	0.14
133	2.2%	97.8%	16.27	7.78	98	0.17
062	64.3%	35.8%	7.36	4.92	99	0.07
021	38.8%	61.2%	11.58	6.81	82	0.14
093	0%	100%	3.25	3.20	62	0.05
122	44.8%	55.2%	5.21	2.42	55	0.09
083	0%	100%	17.50	11.11	83	0.21
181	4.5%	95.5%	11.93	7.38	77	0.15
072	3.0%	97%	13.90	7.13	59	0.24
082	4.5%	95.5%	13.44	7.86	71	0.12
103	34.8%	65.2%	9.43	6.87	89	0.11
011	3.1%	96.9%	8.06	5.29	58	0.14
001	78.5%	21.5%	8.43	4.68	75	0.11
031	53.8%	46.2%	19.99	10.46	91	0.22
123	27.7%	72.3%	3.54	2.75	86	0.04
092	0%	100%	8.70	6.65	71	0.12
102	4.7%	95.3%	0.97	0.83	34	0.03
053	84.4%	15.6%	8.52	4.93	72	0.12
012	10.9%	89.1%	7.30	4.96	78	0.09
112	7.8%	92.2%	16.79	10.25	98	0.17
291	12.5%	87.5%	14.27	8.91	81	0.18

A logistic regression using ordinal data was performed to determine if *Confirmation Bias* had a statistically significant impact on diagnostic accuracy. The regression equation of the result is as follows

Diagnostic Accuracy ~ *Confirmation Bias*

Predicted logit of (Diagnostic Accuracy) = -0.858 + (-0.003) * *Confirmation Bias*

For the *Confirmation Bias* factor, the Wald statistic was 0.134 (p=0.715), indicating that *Confirmation Bias* was not a significant predictor of Diagnostic Accuracy (Table 22).

Table 22 Impact of Confirmation Bias on Diagnostic Accuracy

Predictor Statistical Tests						
Predictor Variable	**B**	**SE β**	**Wald's χ^2**	***Df***	***P***	**e^{β} (Odds Ratio)**
Intercept Only	-0.858	0.090	91.839	1	0.000	---
Confirmation Bias Score	-0.003	0.007	0.134	1	0.715	0.997

7.2.3 Other Variables' Impact on Diagnostic Accuracy

Since this experimental study includes numerous variables, an assessment was performed to determine what other variables and/or combination of variables may have impacted diagnostic accuracy.

7.2.3.1 Impact of Case Difficulty on Diagnostic Accuracy

Heuristic use is not the only variable used in this experiment that may have an impact on diagnostic accuracy. A logistic regression using ordinal data was ran to determine if the difficulty level of the case had an impact on diagnostic accuracy. The regression equation of the result is as follows

Diagnostic Accuracy ~ Case Difficulty

Predicted logit of (Diagnostic Accuracy) = 0.950 + (-0.230) Case Difficulty

For the Case Difficulty factor, the Wald statistic was 9.948 (p=0.002), indicating that Case Difficulty was a significant predictor of Diagnostic Accuracy (Table 23). Case Difficulty has a positive effect on Diagnostic Accuracy (Odd Ratio = 0.794), indicating that the more difficult the case, the more likely an incorrect diagnosis.

Table 23 Impact of Case Difficulty on Diagnostic Accuracy

Predictor Statistical Tests						
Predictor Variable	**β**	**SE β**	**Wald's χ^2**	***df***	***P***	**e^{β} Odds Ratio**
Intercept Only	0.950	0.098	94.410	1	0.000	2.586
Case Difficulty	-0.230	0.073	9.948	1	0.002	0.794

7.2.3.2 Impact of Subject on Diagnostic Accuracy

Since some subjects may be more skillful at correctly diagnosing a case than other subjects, a logistic regression was run to determine the impact of subject variability on diagnostic accuracy. The regression equation of the result is as follows

Diagnostic Accuracy ~ Subject

Predicted logit of (Diagnostic Accuracy) = 0.433 + (-0.981) * Subject ➔ Highest Neg Impact

Predicted logit of (Diagnostic Accuracy) = 0.433 + 1.70 * Subject ➔ Highest Pos Impact

For the Subject factor, the Wald statistic was 74.162 (p=0.229), indicating that Subject was not a significant predictor of Diagnostic Accuracy (Table 24).

Table 24 Impact of Subject on Diagnostic Accuracy

Predictor Statistical Tests						
Predictor Variable	**B**	**SE β**	**Wald's χ^2**	***df***	***P***	**e^{β} (Odds Ratio)**
Intercept Only	0.693	0.433	2.562	1	0.109	2.000
Subject	-0.981 to 1.70	0.692 to 0.850	74.162	66	0.229	0.097 to 1.900

It is not surprising that diagnostic performance varies across subjects. Nor is it completely surprising that this single predictor variable is not statistically significant since the study population consisted of individuals within the same level of education and medical training. The results may have been different had the subjects been from different levels of expertise.

7.2.3.3 Impact of Anchoring, Case Difficulty and Subject on Diagnostic Accuracy

There were multiple aspects of this research study, including diagnosing several cases of varying difficulty levels, representing varying diseases / organ systems (liver, heart, bone, nervous system, etc.), by sixty-seven different individuals who may approach the diagnostic reasoning process differently. Due to these multiple factors, a multi-predictor logistic regression was executed to determine the impact of the combination of *Anchoring*, *Case Difficulty* and *Subject* on diagnostic accuracy. The regression equation of the result is as follows:

Diagnostic Accuracy ~ *Anchoring* * Case Difficulty * Subject

Predicted logit of (Diagnostic Accuracy) = 1.033 + (-0.466) * *Anchoring*

+ (-0.227) Case Difficulty + (-0.911 to 1.777) Subject

For the *Anchoring* factor, the Wald statistic was 3.559 (p=0.059), indicating that Anchoring was not a significant predictor of Diagnostic Accuracy (Table 25). For the Subject factor, the Wald statistic was 72.452 (p=0.274), indicating that Subject was not a significant predictor of Diagnostic Accuracy. For the Case Difficulty factor, the Wald statistic was 10.842 (p=0.002), indicating that Case Difficulty was a significant predictor of Diagnostic Accuracy. Case Difficulty has a positive effect on Diagnostic Accuracy (Odds Ratio = 1.0), indicating that indicating that the more difficult the case, the more likely an incorrect diagnosis.

Table 25 Impact of Anchoring, Case Difficulty, Subject on Diagnostic Accuracy

Predictor Statistical Tests						
Predictor Variable	**β**	**SE β**	**Wald's χ^2**	***df***	***p***	**e^{β} (Odds Ratio)**
Intercept Only	1.033	0.509	4.116	1	0.042	2.809
Anchoring	-0.466	0.247	3.559	1	0.059	0.628
Case Difficulty	-0.227	0.073	10.842	1	0.002	1.000
Subject	-0.911 to 1.777	0.702 to 0.862	72.452	66	0.274	0.402 to 5.910

7.2.3.4 Impact of Confirmation Bias, Case Difficulty and Subject on Diagnostic Accuracy

A multi-predictor logistic regression was executed to determine the impact of the combination of *Confirmation Bias*, *Case Difficulty* and *Subject* on diagnostic accuracy. This model was tested in order to take into account the numerous variables within the study that may impact diagnostic accuracy. The regression equation of the result is as follows:

Diagnostic Accuracy ~ Confirmation Bias * Case Difficulty * Subject

Predicted logit of (Diagnostic Accuracy) = -0.429 + 0.001 * Confirmation Bias + (-0.232) Case Difficulty + (-0.911 to 1.777) Subject

For the *Confirmation Bias* factor, the Wald statistic was 0.019 (p=0.892), indicating that *Confirmation Bias* was not a significant predictor of Diagnostic Accuracy (Table 26). For the Subject factor, the Wald statistic was 72.452 (p=0.274), indicating that Subject was not a significant predictor of Diagnostic Accuracy. For the Case Difficulty factor, the Wald statistic was 9.883 (p=0.002), indicating that Case Difficulty was a significant predictor of Diagnostic Accuracy. Case Difficulty has a positive effect on Diagnostic Accuracy (Odds Ratio = 0.793), indicating that the more difficult the case, the more likely an incorrect diagnosis.

Table 26 Impact of Confirmation Bias, Case Difficulty, Subject on Diagnostic Accuracy

Predictor Statistical Tests						
Predictor Variable	**β**	**SE β**	**Wald's χ^2**	***Df***	***p***	**e^{β} (Odds Ratio)**
Intercept Only	-0.429	0.188	5.199	1	0.023	---
Confirmation Bias	0.001	0.007	0.019	1	0.892	1.001
Case Difficulty	-.232	0.074	9.833	1	0.002	0.793
Subject	-0.911 to 1.777	0.702 to 0.862	72.452	66	0.274	0.402 to 5.910

7.3 IMPACT OF METACOGNITIVE INTERVENTION

Research Question 3: How does a feedback-based intervention, modeled after decision-making strategies received during diagnostic reasoning, impact:

a. **The post-test use of *Anchoring and Adjustment* and *Confirmation Bias*?**
b. **Diagnostic accuracy?**

Data from the second experimental study was used to answer this research question. This study consisted of a between-subject (control, intervention) repeated measures design (pre-test, post-test period). Within the feedback period, subjects in the intervention group received feedback regarding the

correct diagnosis, information regarding their mental model and reasoning strategies, and alternative models and reasoning strategies to consider. Subjects in the control group received the correct diagnosis and information about the disease; no feedback was provided regarding mental models and reasoning strategies.

7.3.1 Time on Task - Amount of Time Spent Solving Cases

For each subject Figure 19 shows the total time spent solving cases within all three periods of the study (the pre-test, feedback and post-test periods). Subjects commonly took between 1.5 and 3 hours to complete the study; times vary across the subject population.

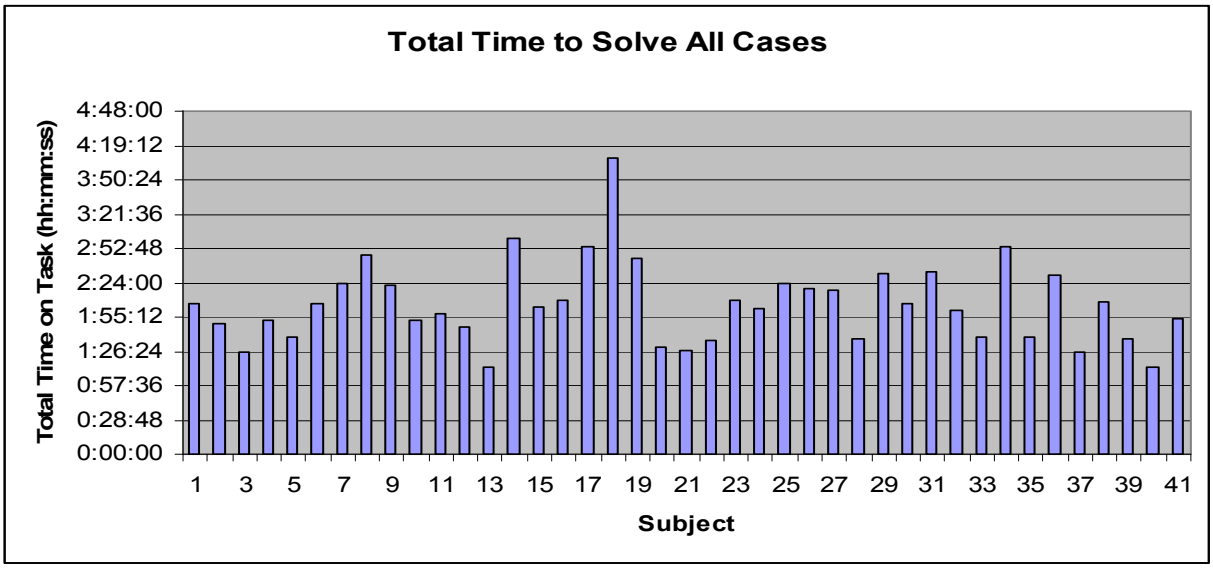


Figure 19 Average Time on Task per Subject

7.3.2 Impact of Intervention on Heuristic Use

The frequency of *Adjustment* and *Confirmation Bias* was calculated during the pre-test and post-test periods. Data from the pre-test period was used as a baseline. Data from the post-test period was compared to the pre-test period to determine if there was a change in heuristic use after receiving the feedback.

It is evident by reviewing the think-aloud protocols that this subject population commonly specified an initial diagnosis early in the diagnostic process (normally when reviewing the first screen

of data). Based on the data from the computer-based system, the adjustment from the initial to the final diagnosis is as specified in section 7.3.2.1, and the commission of Confirmation Bias is as discussed in section 7.3.2.2. The analysis of the think-aloud protocols revealed that subjects commonly verbalized additional initial diagnoses than they entered into the computer-based system. The initial diagnostic hypothesis they verbalized was normally a differential diagnosis consisting of several diseases. As the subject reasoned through the case, they narrowed the differential diagnosis to a differential with fewer diseases than the initial hypothesis, or to a single disease. By the time all the clinical data had been revealed and the subject entered the final diagnosis, the diagnosis they verbalized was generally the same as the diagnosis(es) entered into the computer-based system. The correlation of the think-aloud protocol data and the data entered into the computer-based system for the *data used to diagnose* was poor. Typically there were more items entered into the computer-based system than was verbalized. Beyond these differences, there were no additional distinguishable differences between the think-aloud protocols and the computer-based system.

7.3.2.1 Adjustment

Subjects were required to enter an initial diagnosis (*Anchor*) prior to receiving all the clinical data. Once all the clinical data had been displayed, the subject was required to enter a final diagnosis. The final diagnosis was compared to the initial diagnosis to determine the subjects' *Adjustment factor*. Four types of Adjustment were assessed including: (a) *Sufficient Adjustment* - arriving at the correct final diagnosis when an incorrect initial diagnosis was specified; (b) *Insufficient Adjustment* - arriving at an incorrect final diagnosis after specifying an incorrect initial diagnosis; (c) *No Adjustment Necessary* - entry of a correct initial and final diagnosis (the same diagnosis); and (d) *Adjust Away From Correct Diagnosis* - when the subject specified a correct initial diagnosis and an incorrect final diagnosis.

A breakdown of *Adjustment* is shown in Table 27. For the pre- and post-test periods, and for each study group, this table shows the mean number of cases within each category. During the pre-test period, subjects in the control group *Sufficiently Adjusted* in 0.72 cases (increasing to 1.11 in the post-test period) and in 1.40 cases in the intervention group (dropping to 1.13 during the post-test period). *Insufficient Adjustment* occurred in 1.83 cases for the control group and 1.87 cases for the intervention group during the pre-test period; these figures increased in the post-test period (to 2.61 for the control group and 2.67 in the intervention group). In the pre-test period, *No Adjustment was Necessary* in 2.39

and 1.73 cases in the control and intervention groups, respectively. In the post-test period these figures dropped to 0.50 and 0.40 cases for the control and intervention groups, respectively. Subjects *Adjusted Away from the Correct Diagnosis* in less than one case during the pre- and post-test periods for both study groups (0.39 for the control group and 0.60 for the intervention group during the pre-test period; and 0.39 and 0.73 for the control and intervention group, respectively, during the post-test period).

Table 27 Frequency of Adjustment

		Pre-Test Period		Post-Test Period	
Adjustment Category	**Study Group**	**Mean (# of Cases)**	**Standard Deviation**	**Mean (# of Cases)**	**Standard Deviation**
Sufficient Adjustment	**Control** **Intervention**	0.72 1.40	0.895 0.986	1.11 ↑ 1.13 ↓	0.900 0.640
Insufficient Adjustment	**Control** **Intervention**	1.83 1.87	0.786 0.990	2.61 ↑ 2.67 ↑	1.037 0.617
No Adjustment Necessary	**Control** **Intervention**	2.39 1.73	1.037 1.033	0.50 ↓ 0.40 ↓	0.707 0.632
Adjust Away From Correct Diagnosis	**Control** **Intervention**	0.39 0.60	0.608 0.632	0.39 -- 0.73 ↑	0.502 0.594

For the most part, for both study groups, performance degraded in the post-test period. Even though the number of cases the subjects in the Control group correctly diagnosed (sufficient adjustment) increased in the post-test period, so did the number of cases incorrectly diagnosed (insufficient adjustment). For both study groups, the number of cases where the subject *Anchored* on the correct diagnosis (no adjustment necessary) decreased in the post-test period.

7.3.2.2 Confirmation Bias

Confirmation Bias was assessed by analyzing data collected from two methods – data collected via the computer-based case analysis system, and data from the eye-tracking system. Results of the analysis of data from both systems are discussed.

Section 7.1.2 describes the procedure used to derive a *Confirmation Bias* score on data collected from the computer-based case analysis system. Table 28 details the average *Confirmation Bias* score for all subjects / cases for the pre-test and post-test periods. These calculations are based on forty subjects (N=40) since data was captured by the computer system for all subjects in the second

experimental study. In the *pre-test period*, the mean *Confirmation Bias* score was 0.13 and a standard deviation of 0.05 for the control group; and a mean of 0.09 with a standard deviation of 0.05 for the intervention group. In the *post-test period*, both values decreased; the mean for the control group was 0.10 and the standard-deviation was 0.05; for the intervention group, the mean was 0.06 and the standard deviation was 0.04.

Table 28 Confirmation Bias Score Pre-Test vs. Post-Test (Computer System)

	Pre-Test Period		Post-Test Period	
Study Group	**Mean CB Score**	**Standard Deviation**	**Mean CB Score**	**Standard Deviation**
Control	0.13	0.05	0.10 ↓	0.05
Intervention	0.09	0.05	0.06 ↓	0.04

To determine if **the intervention** impacted the frequency of *Confirmation Bias*, the eye-tracking data was also assessed. The technique to determine the *Confirmation Bias* score based on the eye-tracking data is described in section 6.4.10.3.1. Using this metric, the mean and standard deviation *Confirmation Bias* scores were calculated for the pre-test and post-test periods for the control and intervention groups. The figures shown in Table 29 are based on twenty subjects (N=20) since eye-tracking data was only available for the subjects within the eye-tracking group. In the pre-test period, the mean *Confirmation Bias* score was 0.91 and a standard deviation of 0.04 for the control group, and a mean of 0.91 and a standard deviation of 0.04 for the intervention group. In the post-test period, the mean value for the control group slightly decreased to 0.89 and a standard-deviation of 0.06; for the intervention group, the mean value decreased to 0.90 with a standard deviation of 0.06.

Table 29 Confirmation Bias Score Pre-Test vs. Post-Test (Eye-Tracking Data)

	Pre-Test Period		Post-Test Period	
Study Group	**Mean CB Score**	**Standard Deviation**	**Mean CB Score**	**Standard Deviation**
Control	0.91	0.04	0.89 ↓	0.06
Intervention	0.91	0.04	0.90 ↓	0.06

When assessing *Confirmation Bias* by taking into account the subjects' adjustment behavior (over both the pre-test and post-test periods / not comparing the pre-test to the post-test), *Confirmation Bias* occurred on 44.35% of the total number of cases. From these, on average, subjects committed

Confirmation Bias and Sufficiently Adjusted in 47.0% of the cases; Insufficiently Adjusted in 10.3% of the cases; Adjusted Away from the Correct Diagnosis in 23.9% of the cases; and did not need to adjust (anchored on correct diagnosis) in 18.9% of the cases (reference Table 30).

Table 30 Confirmation Bias and Diagnostic Accuracy based on Eye-Tracking Data

Adjustment Category	Percent of Cases Where Confirmation Bias was Committed
Sufficiently Adjusted	47.0%
No Adjustment Necessary	18.9%
Insufficient Adjustment	10.3%
Adjusted Away from Correct Diagnosis	23.9%

This outcome is a bit surprising in that for a large proportion of the cases that were diagnosed correctly (Sufficiently Adjusted / No Adjustment Necessary), the subjects committed *Confirmation Bias*. For the cases that were diagnosed incorrectly (Insufficient Adjustment / Adjusted Away From Correct Diagnosis), the rate of *Confirmation Bias* was minimal at 10.3% and 23.9%, respectively. Since these figures are based on the analysis of where the subjects' eyes fixated on the screen, these data reflect the fact that for the cases correctly diagnosed, the subjects did not fixate on a large number of either non-critical or critical pieces of information. This resulted in the numerator of the CB score being a mix of the total number of non-critical data elements the subject fixated on, and a large number of critical pieces of data ignored. On the other hand, for the cases diagnosed incorrectly, the subjects fixated on quite a few of non-critical and the critical pieces of data, so the number of non-fixated pieces of critical data was small, thus reducing the numerator of the *Confirmation Bias* scores.

The *Confirmation Bias* score calculated from data logged by the computer-based system was based on the subjects identifying the data they used to arrive at their diagnosis by placing a check-box adjacent to the item. The Confirmation Bias score calculated from data obtained by the eye-tracking system was based on the location of the subjects' eyes, data they fixated on, and data they did not fixate on. Due to the difference in the number of subjects and the difference in the data used to compute the scores for each method (computer system versus eye-tracking) it is difficult to draw a comparison between the results of the two data capture methods. Beyond stating that the mean Confirmation Bias score decreased in the post-test period for both data capture methods, no additional

conclusions can be drawn. Even though both data capture methods were used to derive the *Confirmation Bias* scores, the statistical analysis of the impact of the intervention on *Confirmation Bias* was based on the data from the computer system since the number data was available for forty (40) subjects; whereas the eye-tracking data was only available for twenty (20) subjects. Determining the outcome for a larger number of subjects was felt to be more representative of an accurate measure of the impact of the intervention.

A Pearson Correlation comparing the *Confirmation Bias* score between the twenty think-aloud subjects (data captured by the computer-based system) to the twenty eye-tracking subjects (data captured by the eye-tracking system) revealed a low correlation between the scores (r = 0.29541).

7.3.3 Impact of Intervention on Diagnostic Accuracy

7.3.3.1 Diagnostic Accuracy – Overall Statistics

Table 31 reflects the diagnostic accuracy in the pre- and post-test period for each study group. For both groups, the percentage of cases that were diagnosed correctly decreased in the post-test period. Subjects in the control group, on average, diagnosed 62% of the cases correctly in the pre-test period. This figure dropped to 30% in the post-test period. The same trend occurred for the intervention group who diagnosed, on average, 64% cases correctly in the pre-test group and 31% in the post-test group.

Table 31 Percentage of Correct Diagnosis

	Pre-Test Period		**Post-Test Period**	
	% of Correct Diagnosis		**% of Correct Diagnosis**	
Study Group	**Mean (# of Cases)**	**Standard Deviation**	**Mean (# of Cases)**	**Standard Deviation**
Control	62.2%	15.0%	30.0%	19.7%
Intervention	64.0%	20.0%	30.6%	12.8%

For this study, we see poor diagnostic performance in the post-test period. In order to determine if the use of cognitive heuristics and biases, and/or other variables impact diagnostic accuracy, several independent variables were tested, as discussed in the following sections.

7.3.3.2 Impact of Study Group on Diagnostic Accuracy

A single predictor logistic regression was performed to determine if only the *intervention* (study group) impacted diagnostic accuracy. The regression equation is as follows:

Diagnostic Accuracy ~ Study Group

Predicted logit of (Diagnostic Accuracy) = 0.203 + (-0.061) * Study Group

For Study Group, the Wald statistic was 0.075 (p=0.784), indicating that Study Group was not a significant predictor of Diagnostic Accuracy (Table 32).

Table 32 Impact of Intervention of Diagnostic Accuracy

Predictor Statistical Tests						
Predictor Variable	**B**	**SE β**	**Wald's χ^2**	***df***	***P***	**e^{β} (Odds Ratio)**
Intercept Only	0.203	0.348	0.339	1	0.560	1.225
Study Group (Control vs. Intervention)	-0.061	0.224	0.075	1	0.784	0.940

These statistics indicate that the intervention did not result in a reduction in the number of diagnostic errors. Further investigation is required to determine why this outcome occurred. Were the cases too difficult for the subjects; were some subjects better at correctly diagnosing the cases than others; was ignoring critical clinical data within the case a major contributor to the rate of diagnostic errors? Additional analysis is required to tease out the exact cause for the poor diagnostic accuracy between the periods of this study.

7.3.3.3 Impact of Period on Diagnostic Accuracy

In the second experimental study, subjects assessed cases over three periods (reference section 6.4.4), assessing five (5) cases per period. The periods included a pre-test period, an period, and a post-test period. Determining the impact of the period on diagnostic accuracy was assessed by executing a single predictor logistic regression. The regression equation of the result is as follows:

Diagnostic Accuracy ~ Period

Predicted logit of (Diagnostic Accuracy) = -1.121 + 0.620 * Period

For the Period factor, the Wald statistic was 27.908 (p<=0.000), indicating that Period was a significant predictor of Diagnostic Accuracy (Table 33). Period has a positive effect on Diagnostic

Accuracy (Odds Ratio = 0.289), indicating that subjects are less likely to get the diagnosis correct on cases in the post-test period, compared to cases in the pre-test period.

Table 33 Impact of Period on Diagnostic Accuracy

Predictor Statistical Tests						
Predictor Variable	**B**	**SE β**	**Wald's χ^2**	***df***	***P***	**e^{β} (Odds Ratio)**
Intercept Only	0.740	0.169	19.284	1	0.000	2.096
Period	-1.241	0.235	27.908	1	0.000	0.289

7.3.3.4 Impact of Subject Variability on Diagnostic Accuracy

A logistic regression was also performed to determine if the single variable *Subject* has an impact on diagnostic accuracy. Some subjects may have performed better at diagnosing than others. The regression equation of the result is as follows

Diagnostic Accuracy ~ Subject

Predicted logit of (Diagnostic Accuracy) = 0.000 + (-0.847) * Subject ➔ Highest Neg Impact

Predicted logit of (Diagnostic Accuracy) = 0.000 + 1.386 * Subject ➔ Highest Pos Impact

For the Subject factor, the Wald statistic was 21.221 (p=0.906), indicating that Subject was not a significant predictor of Diagnostic Accuracy (Table 34).

Table 34 Impact of Subject on Diagnostic Accuracy

Predictor Statistical Tests						
Predictor Variable	**B**	**SE β**	**Wald's χ^2**	***df***	***P***	**e^{β} (Odds Ratio)**
Intercept Only	0.000	0.632	0.000	1	1.000	1.000
Subject	-0.847 to 1.386	0.936 to 1.012	21.221	31	0.906	0.429 to 4.000

This finding is not surprising, as the subject population used in this study was confined to a narrow group of clinicians – fourth year medical students and residents. The majority of subjects were fourth year medical students, and a few residents. Since the study population was so controlled, the variability of the skill level of the subject was minimized.

7.3.3.5 Impact of Confirmation Bias on Diagnostic Accuracy

A logistic regression using ordinal data was performed to determine the impact *Confirmation Bias* had on diagnostic accuracy. The regression equation of the result is as follows

Diagnostic Accuracy ~ *Confirmation Bias*

Predicted logit of (Diagnostic Accuracy) = -0.158 + (-0.006) * *Confirmation Bias*

For the *Confirmation Bias* factor, the Wald statistic was 0.075 (p=0.784), indicating that *Confirmation Bias* was not a significant predictor of Diagnostic Accuracy (Table 35).

Table 35 Impact of *Confirmation Bias* on Diagnostic Accuracy

Predictor Statistical Tests						
Predictor Variable	**B**	**SE β**	**Wald's χ^2**	***df***	***p***	**e^{β} (Odds Ratio)**
Intercept Only	-0.006	0.199	0.626	1	0.429	---
***Confirmation Bias* Score**	-0.158	0.022	0.075	1	0.784	1.006

Commission of *Confirmation Bias*, by definition includes ignoring critical information associated with the correct diagnosis and/or only seeking information that supports an inaccurate diagnosis. It is not surprising that a reasoning strategy that ignores critical data associated with the correct diagnosis contributes to diagnostic errors.

7.3.3.6 Impact of Confirmation Bias, Subject, Period on Diagnostic Accuracy

Due to many variables being used in this study, a multi-predictor logistic regression was executed to test the impact of *Confirmation Bias*, *Subject* and *Period* on diagnostic accuracy. The regression equation of the result is as follows

Diagnostic Accuracy ~ Confirmation Bias * Subject * Period

Predicted logit of (Diagnostic Accuracy) = -1.237 + (-0.017) * Subject + 0.676 * Period + (-0.021) * Confirmation Bias ➔ Highest Neg Impact

Predicted logit of (Diagnostic Accuracy) = -1.237 + 1.573 * Subject + 0.676 * Period + (-0.021) * Confirmation Bias ➔ Highest Pos Impact

For the *Confirmation Bias* factor, the Wald statistic was 0.511 (p=0.094), indicating that *Confirmation Bias* was not a significant predictor of Diagnostic Accuracy (Table 36). For the Subject factor, the Wald statistic was 24.009 (p=0.810), indicating that Subject was not a significant predictor of

Diagnostic Accuracy. For the Period factor, the Wald statistic was 28.962 (p<=0.000), indicating that Period was a significant predictor of Diagnostic Accuracy. Period has a positive effect on Diagnostic Accuracy (Odds Ratio = 1.966), indicating that subjects are less likely to get the diagnosis correct on cases solved in the post-test period, compared to cases solved in the pre-test period.

Table 36 Impact of Confirmation Bias on Diagnostic Accuracy

Predictor Statistical Tests						
Predictor Variable	**B**	**SE β**	**Wald's χ^2**	***df***	***p***	**e^{β} (Odds Ratio)**
Intercept Only	-1.237	0.738	2.809	1	0.979	---
Period	0.676	0.126	28.962	1	0.000	1.966
Subject	-0.017 to 1.573	0.947 to 1.061	24.009	31	0.810	0.983 to 4.821
Confirmation Bias Score	-0.021	0.029	0.511	1	0.094	0.290

This multiple predictor model produces the result similar to the individual single predictor regression models. *Subject* variability, *case difficulty* and *heuristic use* were not statistically significant contributors to diagnostic accuracy. The only factor that is statistically significant is the *period* variable. It appears that the longer the subject assessed cases, the more diagnostic errors that occurred. In order to determine if this is a valid assessment, conducting a study in which the time-on-task is reduced, and comparing the outcome of the two studies would be necessary.

7.4 MODE OF THINKING ANALYSIS

Research Question 4: Does heuristic use and diagnostic accuracy differ in the Intuitive and/or Analytical modes of thinking?

a. What is the frequency of use of the cognitive heuristic Anchoring and Adjustment and the cognitive bias Confirmation Bias in each mode of thinking?

b. What is the frequency of diagnostic errors in each mode of thinking?

Data from the second experimental study was used to determine if heuristic use and diagnostic accuracy differs in the two modes of thinking in accordance with the Dual Process Theory.

Specifically, eye-tracking data was assessed to determine if the joint criteria of cognitive load and the amount of time to arrive at a diagnosis (speed) could be used to determine the mode of thinking subjects used as they assessed the clinical scenarios. Once mode of thinking has been established, diagnostic accuracy and heuristic use was assessed to determine the frequency of each event within mode of thinking.

7.4.1 Relationship between Speed and Diagnostic Accuracy

In order to establish the relationship between the *speed* and *diagnostic accuracy* (*error rate)* characteristics of the modes of thinking, I assessed the amount of time (represented by *median dwell time*) spent reviewing clinical data within the critical data feature set and data not considered critical data, for cases diagnosed correctly and incorrectly. As can be seen in Table 37, a median dwell time of 1.902 seconds differentiates *fast* versus *slow* processing for cases diagnosed correctly and incorrectly.

Table 37 Clinical Data Dwell Time

Category	Mean	Std Dev	Std Error	Count	Minimum	Maximum	Median
Dwell time for Correct Diagnosis while reviewing non-critical data	2.568	2.465	0.035	4963	0.083	20.555	1.818
Dwell time for Correct Diagnosis while reviewing critical data	3.284	3.118	0.301	107	0.117	17.551	2.353
Dwell time for Incorrect Diagnosis while reviewing non-critical data	2.816	2.891	0.031	8647	0.083	32.520	1.935
Dwell time for Incorrect Diagnosis while reviewing critical data	3.589	3.635	0.298	149	0.083	20.253	2.520
Total dwell time	2.739	2.762	0.023	13866	0.083	32.520	**1.902**

A Mann-Whitney U-test determined that this value is statistically significant ($p<0.0001$) when processing clinical data fast versus slow when cases are diagnosed correctly and incorrectly (Table 38). This indicates there is a relation between the speed and diagnostic accuracy components of mode of thinking.

Table 38 Mann-Whitney U for Dwell Time

Diagnostic Accuracy	Z-Value	P-Value
Correct	-61.642	< 0.0001
Incorrect	-81.206	< 0.0001

7.4.2 Relationship between Cognitive Load and Diagnostic Accuracy

To determine if there is a relationship between *cognitive load* and *diagnostic accuracy*, the *mean pupil size* was assessed when clinical data was being reviewed in an expedient manner (fast dwell time), and in a more methodical manner (elongated dwell time) for cases diagnosed correctly and incorrectly. Table 39 shows a mean pupil size of 5.571 millimeters is associated with correct and incorrect diagnosis, separated by speed time.

Table 39 Mean Pupil Size by Diagnostic Accuracy and Dwell Speed

Category	Mean	Std Dev	Std Error	Count	Minimum	Maximum
Pupil size for Correct Diagnosis with fast dwell time	5.571	0.720	0.006	13866	3.129	8.669
Pupil size for Correct Diagnosis with slow dwell time	5.527	0.726	0.015	2477	3.666	8.355
Pupil size for Incorrect Diagnosis with fast dwell time	5.472	0.704	0.014	2593	3.129	8.250
Pupil size for Incorrect Diagnosis with slow dwell time	5.666	0.737	0.011	4471	3.986	8.594
Total pupil size	**5.571**	0.696	0.011	4325	3.609	8.669

A Mann-Whitney U test ($p<0.0001$) indicates the difference in pupil sizes are statistically significant between correctly and incorrectly diagnosed cases when speed is considered a factor in the diagnosis (i.e., when the case is diagnosed quickly or slowly) (Table 40). This indicates there is a relation between the cognitive load and diagnostic accuracy components of mode of thinking.

Table 40 Mann-Whitney U for Pupil Size

Mode of Thinking	Z-Value	P-Value
Analytical	-2.196	0.0281
Intuitive	-.6.218	< 0.0001

7.4.3 Relationship between Speed and Cognitive Load (Joint Criteria) and Diagnostic Accuracy

An Analysis of Variance (ANOVA) test indicates that for the cases assessed slowly or quickly, pupil sizes were statistically significantly different between correctly and incorrectly diagnosed cases. The following values are for cases diagnosed *fast*:

	Df	Sum of Square	Mean Square	F-Value	P-Value	Lambda	Power
Final Diagnosis	1	30.743	30.743	57.248	<0.0001	57.248	1.000
Residual	6946	3730.036	0.537				

These figures suggest a statistically significant difference between correctly and incorrectly diagnosed cases, as further shown by Scheffe's post-hoc test (Table 41).

Table 41 Scheffe Test for Pupil Size
Effect: Final Diagnosis Significance Level: 5%

	Mean Difference	Critical Difference	p Value
Correct, Incorrect Diagnosis	-0.139	0.036	<0.0001

The following figures are for cases diagnosed in a *slower* manner.

	df	Sum of Square	Mean Square	F-Value	P-Value	Lambda	Power
Final Diagnosis	1	11.738	11.738	24.023	<0.0001	24.023	1.000
Residual	6916	3379.185	0.489				

Again, these figures suggest statistically significant differences for mean pupil size between cases diagnosed correctly or incorrectly, as further shown by Scheffe's post-hoc test (Table 42).

Table 42 Scheffe Test for Pupil Size
Effect: Final Diagnosis Significance Level: 5%

	Mean Difference	Critical Difference	p Value
Correct, Incorrect Diagnosis	-0.085	0.034	<0.0001

This confirms that the joint criteria of *speed* and *cognitive load* are related to diagnostic accuracy. Therefore, using median dwell time to assess speed, and using mean pupil size to assess cognitive

load, is a viable technique that can be used to categorize (or classify) mode of thinking in relation to the Dual Process Theory.

7.4.4 Heuristic Use and Diagnostic Errors within Mode of Thinking

Using the joint criteria of speed and cognitive load, a mode of thinking was assigned to each line of clinical data for each case diagnosed by the eye-tracking subjects. Once a mode of thinking was assigned to each line of clinical data, these values were reviewed to determine if a single mode of thinking could be assigned to the case. If all the clinical data lines were assessed in the 'Pure Intuitive' mode, the case would be considered a '**Pure Intuitive**' case. If all the clinical data lines were assessed in the 'Pure Analytical' mode, the case would be considered a '**Pure Analytical**' case. If all the clinical data lines were assessed in the 'Mixed Mode', the case would be considered a 'Mixed Mode' case. Subjects commonly switched from one mode to the other throughout the case assessment process. There were no cases where **all** the clinical data elements in the case were assessed in the 'Pure Intuitive' or 'Pure Analytical' mode. Therefore, no cases could be classified as 'Pure Intuitive' or 'Pure Analytical'. There were several cases where the clinical data elements were assessed using either the 'Pure Intuitive' or the 'Mixed Mode' (no lines in the case were assessed with the 'Pure Analytical' category) – these cases have been classified as '**Intuitive / Mixed**'. There were also several cases where the clinical data elements were assessed using either the 'Pure Analytical' or the 'Mixed Mode' (no lines in the case were assessed with the 'Pure Intuitive' category) – these cases have been classified as '**Analytical / Mixed**'. The cases where all three modes were present – these cases have been classified as '**Intuitive / Analytical / Mixed**'.

7.4.4.1 Clinical Data Assessment within Mode of Thinking

Table 43 contains descriptive statistics associated with the assessment of clinical data elements (lines), computed for all cases diagnosed by all the eye-tracking subjects. For the cases solved by the twenty eye-tracking subjects, the average number of clinical data elements analyzed in the 'Pure Intuitive' mode is 27.2% of the total number of elements (11.1% for correctly diagnosed cases and 16.1% for incorrectly diagnosed cases). 23.1% of the clinical data elements were assessed using a 'Pure Analytical' mode of thinking (9.2% for correctly diagnosed cases and 13.9% for incorrectly diagnosed

cases). 49.8% of the clinical data elements were assessed using a 'Mixed Mode' (21.0% for correctly diagnosed cases and 28.8% for the incorrectly diagnosed cases). The largest percentage of lines was analyzed in the mixed mode, and the fewest percentage of lines was analyzed in the pure intuitive mode.

The median dwell time spent assessing clinical data elements using a 'Pure Intuitive' mode of thinking for correctly diagnosed cases was 76.95 seconds, and 92.49 for incorrectly diagnosed cases. The median dwell time spent assessing clinical data elements using a 'Pure Analytical' mode of thinking is 198.29 seconds for cases diagnosed correctly, and 332.75 for cases diagnosed incorrectly. A medium dwell time of 356.75 seconds was spent assessing clinical data elements in the 'Mixed Mode' for correctly diagnosed cases, and 530.10 seconds for incorrectly diagnosed cases. The greatest amount of time was spent analyzing data in the mixed mode, and the least amount of time was spent in the pure intuitive mode. Across all three modes, more time was spent for incorrectly diagnosed cases.

The mean pupil size assessing clinical data elements using a 'Pure Intuitive' mode of thinking for correctly diagnosed cases was 4.851 millimeters, and 5.427 millimeters for incorrectly diagnosed cases. The mean pupil size assessing clinical data elements using a 'Pure Analytical' mode of thinking was 5.683 millimeters for cases diagnosed correctly, and 7.072 millimeters for cases diagnosed incorrectly. The mean pupil size assessing clinical data elements using the 'Mixed Mode' was 5.142 millimeters for correctly diagnosed cases, and 6.020 millimeters for incorrectly diagnosed cases.

Table 43 Clinical Data Element Assessment Statistics

Description	**Pure Intuitive**		**Pure Analytical**		**Mixed Mode**	
	Correct	**Incorrect**	**Correct**	**Incorrect**	**Correct**	**Incorrect**
Mean Percent of Data Elements Analyzed	11.1%	16.1%	9.2%	13.9%	21.0%	28.8%
Median Dwell Time Spent Assessing Clinical Data (seconds)	76.95	92.49	198.29	332.75	356.75	530.10
Mean Pupil Size When Assessing Clinical Data (millimeters)	4.851	5.427	5.683	7.072	5.142	6.020

7.4.4.2 Frequency of Heuristic Use within Mode of Thinking

7.4.4.2.1 Adjustment

The frequency of each adjustment factor was computed for mode of thinking by assessing data for at the case level using the above described categories. Table 44 includes the percent of cases solved in each category. For the cases in the *Sufficient Adjustment* category, very few cases were solved using the Analytical and Mixed modes; the largest percentage of cases were solved using all three modes. For the cases in the *Insufficient Adjustment* category, the largest percentage of cases were solved using all three modes of thinking, with the Analytical / Mixed category as the least used. In the *No Adjustment Necessary* category, the Intuitive and Mixed modes were used least; the largest percentage of cases in this adjustment category was solved using all three modes. In the *Adjust Away from Correct Diagnosis* category, the Analytical and Mixed modes were used the least; with all three modes being used the most. A Kruskal Wallis statistical test indicates that the adjustment factor is not statistically impacted by the mode of thinking (H=0.540, df=2, p=0.766).

Table 44 Adjustment Category within Mode of Thinking

Adjustment Category	Intuitive / Mixed	Analytical / Mixed	Intuitive / Analytical / Mixed
Sufficient Adjustment	5.5%	1.5%	14.0%
Insufficient Adjustment	10.0%	5.00%	31.5%
No Adjustment Necessary	3.0%	3.5%	16.5%
Adjust Away From Correct Diagnosis	2.0%	1.5%	6.0%

7.4.4.2.2 Confirmation Bias

Table 45 shows the mean, standard deviation, median, minimum and maximum *Confirmation Bias* score for each mode of thinking categories. Obviously, the values for the 'Intuitive / Analytical / Mixed' category are greater than for the 'Intuitive / Mixed' and 'Analytical / Mixed' modes since subjects used all modes of thinking for a large proportion of the cases assessed. The mean and standard deviation for the 'Intuitive / Mixed' and 'Analytical / Mixed' categories are similar, with little variation. A Kruskal Wallis statistical test indicates that Confirmation Bias is statistically significantly impacted by the mode of thinking (H=9.325, df=2, p<=0.009).

Table 45 Confirmation Bias Score within Mode of Thinking

Description	Intuitive / Mixed	Analytical / Mixed	Intuitive / Analytical / Mixed
Mean	0.488	0.325	0.857
SD	0.455	0.456	0.206
Median	0.802	0.000	0.906
Min	0.000	0.000	0.000
Max	0.971	0.966	0.975

7.4.4.3 Frequency of Diagnostic Errors within Mode of Thinking

The frequency of diagnostic errors was determined at the case level for each of the mode of thinking categories including 'Intuitive / Mixed', 'Analytical / Mixed' and 'Intuitive / Analytical / Mixed'. Table 46 shows the number of cases diagnosed correctly and incorrectly for each mode. For each of the mode of thinking categories, the largest percentage of cases were incorrectly diagnosed. A Kruskal Wallis statistical test indicates that diagnostic accuracy is not statistically significantly impacted by the mode of thinking (H=0.90, df=2, p=0.975).

Table 46 Diagnostic Accuracy within Mode of Thinking

Diagnostic Accuracy	Intuitive / Mixed	Analytical / Mixed	Intuitive / Analytical / Mixed
Correct	8.50%	5.0%	30.0%
Incorrect	12.0%	6.5%	38.0%

8 DISCUSSION

A goal of this research was to enhance the understanding of the cognitive processes that occur during the diagnostic reasoning process. Specifically, to provide insight on the frequency of cognitive heuristic and bias use during diagnostic reasoning; determine how heuristic use impacts diagnostic accuracy; reduce biased judgment and diagnostic errors using a metacognitive intervention designed to induce physicians to think about how they think by suggesting alternative diagnostic reasoning strategie; and investigate the frequency of diagnostic errors and use of heuristics when processing cases in an intuitive and analytical manner. This chapter will begin with a discussion of the approach used to pursue these goals, including a discussion of the novel aspects of this research, followed by a discussion of the research findings, a discussion on how approaching the problem differently may have resulted in alternative outcomes, the limitations of the study, ending with a conclusion of this research.

8.1 APPROACHES USED TO PURSUE RESEARCH GOALS

The foundations of human judgment and decision theory are concepts that have been studied for decades by a number of researchers within various domains. An area that has garnered much attention is decision-making under conditions of uncertainty. Medicine is a domain where such decisions are made. There are aspects of medicine that require clinicians to reduce complex tasks to simpler judgmental operations, including conditions of uncertainty; the pace at which critical decisions must be made; and incorrect decisions, some of which lead to severe outcomes. Studies have shown that individuals often deal with such situations by using cognitive heuristics. Even though use of heuristics, or mental shortcuts, can lead to appropriate judgments, inappropriate heuristic use can result in severe and systematic errors. In medicine, such errors could include incorrect diagnosis, delayed diagnosis or inappropriate treatment; all of which can result in adverse medical events and patient harm. Due to the

severe consequences of some medical errors, it is imperative to find a way to minimize inappropriate use of cognitive heuristics.

An objective of this research was to reduce biased judgment and diagnostic errors using a metacognitive intervention designed to induce physicians to think about how they think by suggesting alternative diagnostic reasoning strategies. This research extends prior research through the incorporation of several novel techniques. Many attempts have been made, using a variety of methods, to debias individuals from using heuristics inappropriately.[9,22-25,48,54,56,79,118-132] Most attempts have been based on normative decision theories such as Bayes' Theorem and Expected Utility Theory[123,130]and used techniques such as training subjects on heuristics and biases and providing examples of how heuristic-based errors can occur.[9, 118, 123,126,127,129] These attempts have approached the problem by instructing subjects on how they *should* think. This research extends prior debiasing attempts by using *descriptive theories of decision-making* which are based on how people *actually* think. This research approached the problem of flawed judgment by applying principles of the Mental Model Theory, a well-described theoretical foundation of decision-making. The principles of this theory used during this research included assessing the construction of, and reasoning with, mental models constructed while subjects assessed clinical data and diagnosed simulated patients. A computer-based system developed for this project determined the mental model constructed by the subject and assessed it to determine if it accurately represented the case. Based on the subjects' model, suggestions were offered on how to more accurately construct the mental model. Use of the Mental Model Theory, and approaching flawed judgment from how the subject *actually* thinks, are novel components of this research.

In addition to addressing the problem using descriptive theories of judgment and decision-making, this research was unique in that it attempted to induce physicians to think about how they think. It was my hope that the subjects would step outside of the diagnostic reasoning process in order to assess how they reasoned through diagnosing a patient. It was my goal to get the subjects to think about how they think in order to help them gain a better understanding of techniques leading to flawed reasoning. If they were aware of specific actions resulting in flawed reasoning, it was my hope that they may be receptive to receiving suggestions of alternative reasoning strategies. Improving reasoning and judgment through metacognition, or thinking about how you think, is a component that has been investigated in other domains, but that has had limited investigation within medicine and the diagnostic process.

Another novel aspect of this research involved the use of another descriptive theory of decision-making, the Dual Process Theory. A statement made by researchers that have extensively studied heuristics and biases is "there is little direct evidence of the extent to which cognitive biases are leading to diagnostic errors".[151] This research investigated the frequency of cognitive heuristic use and diagnostic accuracy within the two modes of thinking as defined by the Dual Process Theory. The mode of thinking utilized when assessing the clinical scenario was determined by using data obtained when tracking subjects' eye positions including attributes such as pupil size. Assessing mode of thinking is using eye-tracking data is something, to my knowledge, that has never been attempted.

This research not only utilized a novel approach to determining mode of thinking, it investigated if attributes commonly associated with each mode of thinking held true within the realm of diagnostic reasoning and decision-making. Even though the Intuitive mode of thinking is commonly referred to as *heuristic-based* because intuition often involves the recognition of patterns[152,158,163-167,174], to my knowledge, there is no empirical evidence designating the frequency of heuristic use during diagnostic reasoning while operating in this (or either) mode of thinking. Another attribute commonly associated with mode of thinking is error rate. It is *believed* that more errors occur in the Intuitive mode of thinking than in the Analytical mode. Again, to my knowledge, there is no empirical evidence designating the frequency of diagnostic errors that occur in each mode of thinking. This research investigated these two attributes of the modes of thinking defined by the Dual Process Theory, another well-described theoretical foundation of decision-making.

8.2 FREQUENCY OF DIAGNOSTIC ERRORS

The discussion of the results of this study will begin with a review of the frequency of diagnostic errors, as it is believed that this outcome significantly altered the ability to draw conclusions on many other aspects of this research. During both experimental studies, there were a large number of diagnostic errors. Various logistic regression models were applied to the data in an attempt to determine if a single or multiple factors identified the cause of the large percentage of diagnostic errors. *Anchoring,* case difficulty and period were shown to be statistically significant predictors of diagnostic accuracy. *Confirmation Bias,* subject variability and study group were not found to be statistically significant predictors of diagnostic accuracy. The multiple predictors of 'heuristic use,

subject variability and case difficulty' were also found to not be a statistically significant predictor of diagnostic accuracy.

There does not appear to be one distinguishing factor that explains the high rate of diagnostic errors. The case difficulty rating was shown to be a statistically statistic predictor of diagnostic accuracy. Reviewing the diagnostic errors by case difficulty show that case difficulty only had an impact for the cases where the subject specified the correct diagnosis at the point of *Anchoring* (when they specified the initial diagnosis; no adjustment was necessary). The percent of cases diagnosed correctly in the easy difficulty level when no adjustment was necessary is noticeably higher than those in the medium and hard categories. However, the trend of diagnosing easy cases correctly and harder cases incorrectly did not carry over to the other adjustment categories. For example, for the number of cases diagnosed incorrectly (*Insufficient Adjustment*) was somewhat higher in the medium and hard categories than in the easy category, but not significantly higher. There is a "reverse linear relationship" between the difficulty level and the percentage of cases diagnosed correctly in the *Sufficient Adjustment* category. That is, the largest percent of cases diagnosed *correctly* were at the *hardest* difficulty level; the lowest percent of cases diagnosed correctly are in the easy level difficulty.

The percentage of cases misdiagnosed incorrectly in the first experimental study was sizeable at seventy percent. However, in the second experimental study the percentage of cases diagnosed incorrectly during the first period (prior to the feedback period) was much lower at forty percent even though the study population and task between the two studies were comparable. There are a number of factors that could have contributed to this outcome. Even though the process of diagnosing a case when using the computer-based diagnostic system was the same in both studies, the study conditions were considerably different. During the first study, a training lab containing twenty computers was used as the study site. At any time during the hours of 8:00 a.m. and 8:00 p.m. subjects came to the study site, worked silently at a workstation while diagnosing twenty-four cases. Subjects were coming and going throughout the day. A research associate was available at the back of the room to answer questions, but was not in close proximity to the subject as they assessed cases. During the second experimental study, only one subject at a time was present at the study location, which was a small laboratory with two computers. The subject sat at a workstation, with a research associate in very close proximity behind the subject operating the eye-tracking or audio recording equipment. The methods used in this study were considerably different than in the first study. Subjects in the think-aloud group thought aloud as they assessed the first and last five cases; subjects in the eye-tracking group were wearing eye-tracking equipment during the entire session. Based on the data available, it is not

possible to determine if, and how, the study conditions, or other factors, could have impacted the performance differences between the first study and the initial period of the second study.

Based on the data available, it appears that the knowledge base of the subject population used in this study may not have been the level required to correctly diagnose the cases used in this study. Cases of varying level difficulty were purposely selected for use in the study so as to have some cases that the subjects could properly diagnose with ease, along with cases that would be challenging. However, it appears a considerable proportion of the cases were a significantly challenging to this study population.

8.3 FREQUENCY OF HEURISTIC USE

A component of this research was to obtain descriptive statistics on the use of *Anchoring and Adjustment* and *Confirmation Bias* and the frequency of diagnostic errors within the population of fourth-year medical students and residents as they diagnose clinical scenarios.

8.3.1 Frequency of Anchoring

One of the first steps that occur early in the diagnostic process is generation of one or more diagnostic hypotheses.[49] This step, which corresponds to *Anchoring* on an initial diagnosis, was evident in this research. During the first experimental study, even though subjects were not required to specify an initial diagnosis, for the majority of the cases subjects *Anchored* on an initial diagnosis after reviewing the first of three screens of clinical data. Subjects rarely deferred specifying a diagnostic hypothesis until all the clinical data had been reviewed. Clinical data was presented to the subjects over three screens, each screen revealing additional details of the case. *Anchoring* commonly occurred after review of the first screens which only revealed the chief complaint and history of present illness. The second screen revealed the physical examination data along with the past medical history, medications and allergies as well as the family history. In most cases subjects did not wait to review the physical examination data; they had *Anchored* on an initial diagnostic hypothesis after review of only the chief complaint and history of present illness information. The difficulty level of the cases did not alter the

Anchoring behavior, in that the percentage of cases from the easier, medium and harder difficulty levels where *Anchoring* occurred was nearly equivalent. Regardless of the difficulty of the case, subjects continued to Anchor.

8.3.2 Frequency of Confirmation Bias

Once an initial or differential diagnosis is specified, the hypothesis refinement process occurs. This process normally includes the elimination of (ruling-out) diseases that are no longer plausible and/or specifying (ruling-in) additional diseases once additional data is reviewed.[49] Several strategies are used to narrow down an initial diagnosis, including seeking information that enhances a highly likely hypothesis and/or reduces the likelihood of an unlikely hypothesis.[49] When seeking information that confirms a particular diagnosis, clinicians may ignore information that refutes that diagnosis. When this occurs, the clinician's actions are biased toward confirming their diagnosis; an action referred to as *Confirmation Bias*.

Descriptive statistics derived from data collected in the first experimental study (based on data subjects indicated they used to arrive at their diagnosis) indicate subjects use a small proportion of the clinical data available to arrive at a diagnosis. A large proportion of both the critical data elements that directly correspond to the correct diagnosis as well as the less significant data elements are not utilized to arrive at a diagnosis. The *Confirmation Bias* scores reflect a measurement of *Confirmation Bias*. These scores vary across the cases within each difficulty level. The data indicate that across all three case difficulty levels, commission of *Confirmation Bias* seemed to be common for this subject population.

Data from the second experimental study captured from the eye-tracking system indicate that a large percentage of all the data available (both the critical and non-critical data) was reviewed by the subject; only a small proportion of the data did not seem to draw the attention of the subjects. This result was expected from medical students and those with limited clinical experience. According to the eye-tracking data, commission of *Confirmation Bias* occurred on slightly less than half of the cases assessed. Relative to the number of clinical data lines in each category (critical and non-critical data elements), the amount of time spent reviewing data in both categories was comparable. Therefore, it is hard to determine if seeking data that supports a diagnosis or avoiding data that refutes that diagnosis was a greater contributor to the occurrence of *Confirmation Bias*.

The difference between the number of data elements used to arrive at the diagnosis, based on data from the first experimental study, and the data elements reviewed based on eye-tracking data from the second experimental study is quite interesting. One can only speculate as to why the differences occurred. Perhaps the subjects in the first experimental study failed to identify all the elements they used (or at least reviewed) to arrive at the diagnosis, and/or the behavior of an individuals' eye indicate they look at more than they think they look at.

8.3.3 Frequency of Adjustment

Once the refinement process is complete, a final diagnosis(es) is specified. If the final diagnosis is different than the initial diagnosis (Anchor point), an *adjustment* has occurred. Of the various types of adjustment factors assessed during this research, *Insufficient Adjustment* in which the subject misdiagnosed the case occurred more frequently than cases where *No Adjustment was Necessary* and *Sufficient Adjustment* occurred. Since diagnostic accuracy and the adjustment factor are directly correlated, I will not repeat the discussion on adjustment since it was discussed above when discussing the frequency of diagnostic errors.

8.4 IMPACT OF HEURISTIC USE ON DIAGNOSTIC ACCURACY

The subject population used in this research commonly arrived at an incorrect final diagnosis regardless of heuristic use. In the first experimental study, a large proportion of the cases were diagnosed incorrectly when the subjects *Anchored* and when they did *not Anchor*. In the second experimental study even though subjects were forced to Anchor, diagnostic errors were still prevalent. Even though the outcome of the single predictor logistic regression test indicates *Anchoring* is a statistically significant predictor of diagnostic accuracy, this subject population committed diagnostic errors regardless of their Anchoring behavior. Obviously there are additional factors that caused the large proportion of diagnostic errors by this particular study population when diagnosing this specific set of clinical scenarios.

As with Anchoring, diagnostic errors were prevalent when Confirmation Bias occurred and when it did not occur. Use of critical data within the case did not decrease the number of diagnostic errors in that a large proportion of cases in both experimental studies were diagnosed incorrectly even when the subjects indicated they used the critical data to arrive at their diagnosis, and/or when the eye-tracking data revealed subjects spent a considerable amount of time reviewing the critical data elements. In addition, ignoring the critical data within the case did not necessarily reduce the number of cases diagnosed correctly. The eye-tracking data revealed that even in cases where the critical data was ignored, the case was still diagnosed correctly. These findings are consistent with several studies that have shown that lay people, medical students and physicians often do not select optimal diagnostic data; they often select pseudo-diagnostic information which can lead to premature selection of a diagnosis (often referred to as premature closure).[9,132,215-217]

Further investigation with this subject population and/or an alternative set of clinical scenarios is required to determine the impact *Anchoring* and *Confirmation Bias* has on diagnostic accuracy.

8.5 IMPACT OF METACOGNITIVE INTERVENTION

This study is one of the only empirical studies that have used a metacognitive intervention of feedback designed to induce physicians to *think about how they think* as they reason through diagnosing patients. Given the innovativeness of this project, the ability to predict the outcome prior to conducting the study and/or to compare the outcome to other studies is impossible. There are several aspects of the intervention that have the potential to provide valuable insight and information to enhance future research.

The efficacy of this intervention was determined by assessing the impact of variables including study group, period, subject variability, commission of Confirmation Bias, and the combination of these variables on heuristic use and diagnostic accuracy. Of these variables, the only one that had a statistically significant impact on heuristic use and diagnostic accuracy was period. The other variables (study group, subject variability, and commission of Confirmation Bias) did not significantly impact the outcomes measures in this study. Therefore, there was no significant effect of the intervention on diagnostic accuracy. An in-depth discussion of the results and trends associated with *diagnostic*

accuracy is found earlier in this chapter. All that can be stated with confidence is that diagnostic accuracy was significantly impacted by period.

The *outcome of the metacognitive intervention* was that **the intervention had no effect.** The intervention of feedback was designed to induce physicians to think about how they think by providing alternative reasoning strategies to consider, with the hopes of improving diagnostic reasoning and accuracy. Unfortunately, the desired outcome did not occur. There are a number of reasons that the intervention may not have had a positive effect, all of which are speculative and subjective in nature. Given the data available, one can only speculate as to the exact cause of this outcome. Some possible reasons could include the time-on-task, the number of cases diagnosed, or the amount of clinical data comprising each case. The amount of time subjects were at the study location ranged from slightly under one hour to nearly four hours, averaging slightly over two hours across all the subjects. Subjects in the think-aloud group thought-aloud during approximately three-fourths of the total time-on-task (during assessment of ten of the fifteen cases). Subjects in the eye-tracking group wore eye-tracking equipment for the entire session. For the fifteen cases each subject reviewed, the number of clinical data elements per case ranged between fifty-seven to one hundred and ten data lines, averaging seventy-six elements per case. All of these factors could have played a role in the performance degradation. The subjects could have experienced fatigue and/or data overload. However, since the performance degraded for both study groups, this does not explain why the intervention failed. If the intervention had had an impact, the performance decline would most likely have been more dramatic for the control group than it was for the intervention group.

The outcome of this study does not correlate with the positive results that have been shown from empirical studies that have assessed the impact of feedback on performance.[148,149,202] Aspects of feedback tested during prior studies included assessing the impact of temporal aspects of feedback. There is empirical evidence indicating receipt of immediate feedback while using a computer-based medical intelligent tutoring system resulted in significant improved performance between a pre-test and post-test.[148, 202] There is also empirical evidence that receiving immediate feedback while using an intelligent tutoring system had a statistically significant positive effect metacognitive performance.[149] This dissertation research study used feedback techniques similar to these studies, however performance enhancements were not comparable. Reasons of the differences in performance outcomes may be partially due to the prior studies providing feedback after each and every action performed versus this study providing feedback after the final diagnosis of the case being specified. The feedback provided during this study may be considered *delayed feedback* instead of the *immediate feedback*.

Delayed feedback has been shown to be associated with decreased performance.[149] Providing feedback on each and every action taken by the subjects of this study as they diagnosed cases would have defeated the purpose of assessing the manner in which this population naturally utilized cognitive heuristics during the diagnostic process. There are also studies that have shown limited effectiveness, and even negative effects, have occurred when one receives only performance feedback versus receiving performance feedback plus additional information that enables a decision maker to examine the future effects of their actions.[134, 135, 137-139] Even though this study included a form of feedforward feedback by providing reasoning strategies to consider to enhance future diagnostic reasoning, a difference between this study and prior studies where improved performance was the outcome is that in those studies implemented feedforward feedback by providing subjects with the ability to compare their decisions with an expert by watching a video of an expert performing the task and/or by providing what-if tool that revealed the correctness of an outcome if certain actions were taken.[134,135,139, 140] The outcome of this study may (or may not) have been different had some of these techniques been implemented.

So why did the intervention fail? Did the study that was executed permit the best possible measure of the metacognitive intervention? Obviously, the study design selected, including the task, instrument, methods and study population, were all thought to be a combination that would produce favorable results.. There are aspects of this study that, if implemented differently, different outcomes *may* have been more favorable. For example, changing the order in which the feedback components were presented to those in the intervention group may or may not have resulted in different outcomes. Once the subject entered a final diagnosis for each of the cases in the feedback period, the subject was presented with a summary of the reasoning strategy and mental model that the computer-based system determined they used as they diagnosed the case. This was not a complex message; it included phrases such as "You appear to be mapping a set of symptoms to a single disease" or "You appear to be mapping a set of features to multiple diseases, recognizing that some symptoms are associated with more than one diagnosis". This message was followed by a summary of the symptoms used by the subject to arrive at the diagnosis, along with the initial and final diagnosis. Once the subject reviewed this information, they pressed a button which then displayed a screen with four components including (1) alternative reasoning strategies to consider, (2) the correct diagnosis for the case, (3) a short description of the disease, common symptoms associated with the disease, the population (age, gender, sex, etc.) that the disease normally presents in, and a short description of specific ways attributes in the case reinforce the reasoning strategies to consider; followed by (4) a graphical representation of a

mental-model that properly depicts the clinical data presented in the case along with the correct diagnosis and common diseases one might select if the case is misdiagnosed. In reviewing the eye-tracking data collected when subjects reviewed the feedback data presented during the intervention (a task performed to determine if the subject even looked at critical components of the feedback) it was found that subjects spent a considerable amount of time reviewing the graphical representation of the mental-model (forth / last component presented) and very little time reviewing the reasoning strategies to consider. On average, the amount of time spent reviewing the graphical mental-model component was three times greater than the amount of time spent reviewing the alternative reasoning strategies to consider. The least amount of time was spent reviewing the alternative reasoning strategies to consider. One could say that the "you can lead a horse to water but can't make him drink" saying may apply in this situation. The subjects did not spend a great deal of time reviewing a component of the feedback intervention that was meant to be a critical component associated with inducing the subject to think about how they think. Without consciously considering the alternative reasoning strategies, one will receive little benefit from such information. Alternative techniques may have been used to instill the alternative reasoning strategies. Had techniques designed to reinforce the message been utilized, the outcome may have been more favorable.

Some might feel the metacognitive intervention had a *negative* effect on the subjects, and that the intervention is what resulted in the suboptimal outcome. Data collected during the study does not provide evidence that this is a valid conclusion. Had the metacognitive intervention had a negative effect, performance would not have equally degraded for both study groups; it would have only had a negative effect on the intervention group. This is confirmed by the outcome of the multivariate logistical regression models that were ran for this study. These models showed that there were multiple variables that impacted diagnostic performance; variables such as case difficulty which had a statistically significant impact on diagnostic accuracy. There is no evidence to support the conclusion that the intervention caused the negative outcome; actually, there is evidence to dispute this claim.

8.6 MODE OF THINKING ANALYSIS

During this study a novel approach of using eye-tracking data was used to determine if this subject population assessed cases using the Intuitive or Analytical mode of thinking. Not only does this research provide a technique of assessing mode of thinking within the domain of eye-tracking, it also provides empirical evidence of the frequency of heuristic use and diagnostic errors within the modes of thinking as defined by a well-described theoretical foundation, the Dual Process Theory.

8.6.1 Critical Findings Relating to Aspects of Deriving Mode of Thinking

This research extends prior research performed within the domain of medicine that has established a relationship between pupil size and the level of information processing being carried out by the brain,[203] an individual's mental capacity and overall memory load.[203,204,205] This research also extends prior research in non-medical domains which has shown that cognitive load directly correlates to mode of thinking.[157,158] This research has not only *utilized findings* of research conducted within different domains, but it has *connected the findings of research conducted separately* and arrived at a novel technique of assessing critical aspects of judgment, clinical reasoning, and medical decision-making.

A noteworthy outcome of this research was the identification of statistically significant factors of speed (represented within the domain of eye-tracking by median dwell time), and cognitive load (represented by mean pupil size) that can be used to categorize (or classify) mode of thinking when diagnosing cases correctly and incorrectly. Identification of a significant median dwell time and mean pupil size is a critical finding of this research; and is a significant contribution to the study of human decision-making, as using aspects of an individual's eyes to determine mode of thinking (as defined by the Dual Process Theory) has not been reported prior to this study.

It should be noted that we classified a large percentage of the clinical data elements (nearly 50%) as 'Mixed Mode'. This finding indicates that future research is required to verify that the statistically significant dwell time and pupil size we identified are appropriate for this study population assessing these clinical scenarios. The fact that the range between the mean pupil sizes in the 'Pure Intuitive' and 'Pure Analytical' categories are not vastly different for cases diagnosed correctly and incorrectly is promising. This conveys that perhaps the statistically significant mean pupil size we

derived is in close proximity to a pupil size that may be a good predictor of the cognitive load associated with the joint criteria of speed and cognitive load that defines mode of thinking.

Beyond establishing a connection between eye-tracking and mode of thinking, this research has also established a relationship between the *speed in which clinical data is processed* and *diagnostic accuracy*. In addition, we established a relationship between *cognitive load* and *diagnostic accuracy* using aspects of eye-tracking. Identification of specific pupil sizes and information processing speeds that have the potential to move an individual to the point of flawed judgment that could lead to a medical error is a striking finding. The relationships identified from this research have the potential to provide valuable insight into the aspects of faulty clinical reasoning and diagnostic errors that could impact the safety of patients.

8.6.2 Frequency of Heuristic Use and Diagnostic Errors within Mode of Thinking

The literature indicates that when in the Intuitive mode of thinking we typically use heuristics, or mental shortcuts, and judgments are often made by relying on our instinctive first impressions.[50,152,167] From these statements one could imply that heuristics are not commonly used in the Analytical mode of thinking. The findings of this research do not totally support these claims. Our research indicates that for this study population, adjustment behavior is not statistically impacted by mode of thinking. This indicates that one's adjustment behavior does not differ between the modes of thinking. Our findings indicate that commission of Confirmation Bias is statistically significantly impacted by mode of thinking. This finding is not surprising since the formula used to calculate the Confirmation Bias score was based on the number of clinical data elements within the case, and the mode of thinking assigned to the case is also based on the assigning a mode of thinking to each clinical data element in the case.

The literature also states that the Analytical mode of thinking "epitomizes the kind of thinking that is usually associated with effective problem-solving".[152] Our research found that diagnostic accuracy is not statistically significantly impacted by mode of thinking. We found that the percentage of cases diagnosed correctly and incorrectly are similar in each mode of thinking (Intuitive mode: correct diagnosis – 8.50%; incorrect diagnosis – 12.0%; Analytical mode: correct diagnosis – 5.0%; incorrect diagnosis – 6.5%). This finding does not support claims in the literature that there is a low

error rate in the Analytical mode of thinking, and that the Intuitive mode of thinking is associated with a higher error rate than the Analytical mode.

The differences between the findings of this research and the claims made by practicing clinicians, and those who have spent decades studying clinical reasoning, convey that additional research is warranted to fully understand how the Dual Process Theory applies to the process of diagnostic reasoning. Even though additional empirical research is imperative, it is apparent that use of eye-tracking data to determine mode of thinking has great potential to help us understand the cognitive aspects of diagnostic reasoning, contributors of flawed reasoning, and factors that may lead to medical errors.

8.7 STUDY LIMITATIONS

Every research study has limitations, as does this study. This study was conducted in a laboratory setting with artificial conditions; therefore the results of the study may not transfer to real world tasks. As with many experimental studies, the ecological validity of this experiment did not represent the environment clinicians face on a day-to-day basis. The methods used in this experiment were not consistent with those that occur in the clinical setting in that when an individual assesses a case in order to diagnose a patient, they may not think-aloud while reasoning through each clinical data element; they definitely do not wear eye-tracking equipment while diagnosing a patient. Even though the task of diagnosing patients was the same task that physicians perform on a daily basis, the instrument used in the study was not commensurate with manner in which a clinical case is normally presented in the real-world clinical environment. In the real-world clinical data is normally not presented in list form and may not be organized in categories such as 'history of present illness', 'past medical history', etc. While related data may be found in close proximity, the data may be scattered throughout a paper chart requiring one to hunt for the data associated with a particular category. In the real-world setting clinicians may not think-aloud and/or to specify what data they use to arrive at their diagnosis as was required during this study. Use of computer-based systems to review clinical data may be more prevalent in today's society than it was a decade ago with the incorporation of electronic medical records; however, diagnosing a patient strictly by assessing electronic data and never seeing the patient is not something that commonly occurs in medicine except for situations such as a

consultation. If a computer-based system is used in the real-world environment, the clinician more than likely is not required to specify a diagnosis after reviewing a portion of the data, as was required in the second experimental study of this research. The setting of both experimental studies of this research was much different than would be found in a hospital or the private practice of a clinician. In general, the diagnostic behaviors observed and recorded in this study may somewhat reflect the behaviors that actually occur in natural setting, but the external factors that may impact an individual's diagnostic process was not present in this study. In the clinical environment there commonly are many distractions and interruptions that can impede one's thought process. Additional factors that have the potential to impact diagnostic reasoning include the context of the clinical setting and the situation being dealt with, the interaction of the team of clinicians dealing with the patient being diagnosed, one's own individual personality and preferences, the amount of things on one's mind or the cognitive load caused not only by the current situation, but by others things that may be going on in the physician's mind, having to deal with indolence in those who need to complete tasks that are required to properly treat a patient, etc.[142, 152] Duplicating the real-world clinical setting in a laboratory is nearly impossible; one can only hope to create an environment where the goals of the research are not impeded to the point where valid conclusions cannot be drawn.

Other limitations of this study include the ability to generalize the findings to practicing physicians. The study population used in this study consisted primarily of fourth year medical students and a few residents in the first, second and third years of residency. The knowledge base and reasoning strategies of this population is much different than those of physicians that have been practicing for a number of years. Generalizing the results of the study to clinicians beyond this study population may not be possible. The use of the computer-based system to diagnose patients may have impacted the outcome of the study in several ways. Even though in the first experimental study subjects were only required to enter a final diagnosis, and were not required to enter an initial diagnosis, the presence of a text-box on each data screen may have made them feel an entry of an initial diagnosis was expected. Evidence exists that clinicians commonly arrive at an initial diagnosis very early in the diagnostic process, so this bias toward entering an initial diagnosis may not have been a result of the computer-based system, but more a result of the manner is which medical students (and clinicians in general) commonly perform when reasoning through a clinical case. Specific instructions were given to each subject that entry of an initial diagnosis was not necessary; in addition subjects were instructed that they could revisit previously seen data at any time, therefore the design and/or use of the computer-based system may not have had an impact on the outcome. An area that the computer system may have

had an impact on the outcome was during the review of the feedback for the subjects in the intervention group. The inability to utilize color when presenting the feedback may have impacted the ability of the subject to link components of the feedback. The feedback was designed and presented in a manner where the information presented in the "reasoning strategies to consider" section was linked to the mental model graphic and the text that described the mental model. The feedback was designed in this manner so as to reinforce the reasoning strategy to consider both via text and visually in graphical form so as to maximize the potential of the feedback having a positive impact. The inability to use colored text and graphics due to the restrictions of the monitors used during the second experimental study made it difficult to connect critical components of the feedback. Using black and white text and various line formats such as dashed and dotted lines was required instead of using color that could have been used to visually correlate aspects of the feedback. Another aspect of the feedback that may have contributed to the intervention having no effect was that the information design of the feedback may not have been optimal. The order in which the feedback components were sequentially presented may have resulted in the feedback having no effect. Ordering the components of the feedback differently, such as physically locating the reasoning strategies to consider closer to the mental model graphic may have resulted in the subject being able to more clearly recognize the connection between the two components. Another component of the computer-based system that may have skewed the results and/or caused the intervention to have had no effect was the fidelity of the structured data captured by the system. In comparing the think-aloud protocols to the data logged by the computer system (based on the subject identifying the data they used to arrive at the diagnosis) there were differences between what was verbally spoken and what was selected on the computer screen as data used to arrive at a specific diagnosis. Even though subjects were given instruction to only place a check-mark adjacent to those items they used to arrive at the diagnosis being entered into the system (by entering the disease(s) name in a text-box), there seemed to be items verbally spoken that did not have a check-mark adjacent to them, as well as items checked that verbally the subject indicated they used to rule out other diseases. The items that the subject placed a check-mark adjacent to (identified as 'data used to diagnose') was used by the computer system to dynamically build the component of the feedback that indicated what mental model and reasoning strategy the subject seemed to be using as they diagnosed the case. If the items selected via the computer system was not representative of what actually was used, or did not represent the manner in which the data was used (ruling in versus ruling out a disease) the feedback may have been confusing to the subject, resulting in an inability to comprehend and/or apply the feedback to their personal reasoning strategies.

Within areas such as cognitive science and psychology, there are general definitions of cognitive heuristics and biases. However, when applying heuristics and biases to the domain of medicine, diagnostic reasoning in particular, there are no agreed upon definitions of some of the heuristics and biases. In addition, the application of heuristics and biases to specific actions occurring during diagnostic reasoning is subjective. Also, a particular action could include the use of several heuristics and/or biases. For example, one might be using the heuristic of *Availability* when arriving at an initial diagnosis, but yet be biased by their personal belief that a particular complaint is psychological in nature rather than a 'real medical problem'. Therefore, the metrics used to identify heuristics and biases may not necessarily be measuring what is intended to be measured. Identifying the occurrence of a single heuristic and/or bias is extremely difficult. Research in the development of metrics to identify heuristics and biases has been ongoing for many years. Identifying and measuring heuristics and biases continues to be an area in which continued research is needed.

The final limitation which I will discuss is the lack of a gold standard when assessing the eye-tracking data; specifically in terms of the ability to use the data to determine mode of thinking. The use of eye-tracking data to determine mode of thinking has never been attempted; therefore determining the validity of the dwell time and pupil size threshold values was difficult. The threshold values derived were based on a single dataset. The use of the method developed needs to be applied to additional datasets to determine its' validity.

8.8 CONCLUSION

The identifiable positive outcomes of this dissertation research are numerous. Even though the metacognitive intervention did not have a positive effect and did not seem to reduce diagnostic errors and reduce the inappropriate use of cognitive heuristics and biases, there were many lessons learned during the design, development, and implementation and testing of several novel methods and techniques utilized in this research. The potential of the novel aspects of this research are easily recognizable. Aspects of this research such as inducing physicians to think about how they think could be viewed by some as 'going against the grain' or 'out in left field' since during their medical training clinicians are not taught to think about how they think; they are taught diseases, symptoms, and the underlying pathophysiological processes of the human body. This dissertation project attempted to

take a unique approach to debiasing individuals on the inappropriate use of cognitive heuristics and biases. Even though the outcome was not as hoped and expected, the doctoral candidate still believes that this approach has a great deal of potential if implemented within an experiment utilizing subjects that have a knowledge base more commensurate of the clinical scenarios.

Not only did this dissertation attempt to deal with the well-known problem of the inappropriate use of cognitive heuristics and biases in a unique manner, this dissertation research also *pushed the envelope* by incorporating the use of eye pupil-size to determine an individual's level of cognitive load and mode of thinking during diagnosis. Use of such an overt physiological marker to quantify mode of thinking has not previously been reported. The results of this component of this research are impressive and have the potential to significantly contribute to the understanding of cognitive aspects of medical decision making, how physicians think, and the overall impact on medical error.

APPENDIX A

MEDICAL SCHOOL APPROVAL FOR USE OF MEDICAL STUDENTS

University of Pittsburgh

School of Medicine
Office of Medical Education

M211 Scaife Hall
3550 Terrace Street
Pittsburgh, PA 15261
412-648-8714
Fax: 412-383-7477

John F. Mahoney, MD
Associate Dean

April 22, 2010

Velma Payne
University of Pittsburgh School of Medicine
Department of Biomedical Informatics
Parkvale Building M-183
200 Meyran Ave
Pittsburgh, PA 15260

Subject: Proposed Research Project Involving Medical Students

Dear Ms. Payne,

Thank you again for submitting your proposed project, "Analysis of the effect of feedback on use of cognitive heuristics during diagnostic reasoning", for review.

We have reviewed your protocol and IRB submission documents. Based on our review of these documents, we approve of your stated plans.

Please feel free to forward this letter to the IRB to assist with your submission.

Sincerely,

John F. Mahoney, MD
Associate Dean for Medical Education

INSTITUTIONAL REVIEW BOARD APPROVAL FOR USE OF SUBJECTS

PI Notification: Your research study received approval under expedited review
irb@pitt.edu [irb@pitt.edu]

Sent Wednesday, April 07, 2010 9:59 AM
To: Payne, Velma

University of Pittsburgh
Institutional Review Board

3500 Fifth Avenue
Pittsburgh, PA 15213
(412) 383-1480
(412) 383-1508 (fax)
http://www.irb.pitt.edu

Memorandum

To: Velma Payne

From: Sue Beers PHD, Vice Chair

Date: 4/7/2010

IRB#: PRO09120344

Subject: Analysis of the effect of feedback on use of cognitive heuristics during diagnostic reasoning

The University of Pittsburgh Institutional Review Board reviewed and approved the above referenced study by the expedited review procedure authorized under 45 CFR 46.110 and 21 CFR 56.110. Your research study was approved under:
45 CFR 46.110.(6) Recordings made for data/research
45 CFR 46.110.(7) characteristics/behaviors

Approval Date: 4/5/2010

Expiration Date: 4/4/2011

For studies being conducted in UPMC facilities, no clinical activities can be undertaken by investigators until they have received approval from the UPMC Fiscal Review Office.

Please note that it is the investigator's responsibility to report to the IRB any unanticipated problems involving risks to subjects or others [see 45 CFR 46.103(b)(5) and 21 CFR 56.108(b)]. The IRB Reference Manual (Chapter 3, Section 3.3) describes the reporting requirements for unanticipated problems which include, but are not limited to, adverse events. If you have any questions about this process, please contact the Adverse Events Coordinator at 412-383-1480.

The protocol and consent forms, along with a brief progress report must be resubmitted at least one month prior to the renewal date noted above as required by FWA00006790 (University of Pittsburgh), FWA00006735 (University of Pittsburgh Medical Center), FWA00000600 (Children's Hospital of Pittsburgh), FWA00003567 (Magee-Womens Health Corporation), FWA00003338 (University of Pittsburgh Medical Center Cancer Institute).

Please be advised that your research study may be audited periodically by the University of Pittsburgh Research Conduct and Compliance Office.

STUDY ONE SUBJECT SOLICITATION

I am writing to ask you to participate in a research study that will be conducted at the University Of Pittsburgh, Department of Biomedical Informatics, within the School of Medicine. The purpose of this research study is to determine how cognitive heuristics are used during diagnostic reasoning; if use of heuristics impacts diagnostic accuracy; and if receiving feedback regarding diagnostic reasoning strategies contributes to the appropriate use of heuristics and impacts diagnostic accuracy. The results of this study will be used to develop new instructional tools for medical students and residents that will be designed to enhance physicians' diagnostic reasoning processes. I am writing to invite your participation in this study.

During the study, you will be asked to assess and diagnose clinical scenarios using a computer-based system. You will assess 24 cases, providing an initial diagnosis if you can, designate what data you used to arrive at that diagnosis, then provide a final diagnosis and designate the data you used to arrive at the final diagnosis.

As a token of our appreciation, each participant will receive $25 per hour. The study will last approximately 2 hours. Compensation will be based on a quarter hour basis, rounded up to the next quarter hour. For example, if the study takes between 1 hour and 15 minutes and 1.5 hours, you will be paid $37.50. If you complete the study in 1 hour and 40 minutes, you will be paid for 1.75 hours, receiving $43.75. The maximum amount you will be paid is for 2 hours, totaling $50.00. If you do not complete the entire study, you will be paid on a quarterly hour basis, rounded up to the next quarter hour, for the time they spent at the study location. For example, if you decide after 1.5 hours to discontinue, you will be paid $37.50. If you decide after 40 minutes to discontinue the study, you will be paid $18.75 for 45 minutes.

If you are interested in participating, or have further questions, please contact me at (412) 977-3978, or send an e-mail to vlp2@pitt.edu

Sincerely,

Velma L. Payne, MS, MBA, MS
University of Pittsburgh School of Medicine
Department of Biomedical Informatics

STUDY TWO SUBJECT SOLICITATION

I am writing to ask you to participate in a research study that will be conducted at the University Of Pittsburgh, Department of Biomedical Informatics, within the School of Medicine. The purpose of this research study is to determine how cognitive heuristics are used during diagnostic reasoning; if use of heuristics impacts diagnostic accuracy; and if receiving feedback regarding diagnostic reasoning strategies contributes to the appropriate use of heuristics and impacts diagnostic accuracy. The results of this study will be used to develop new instructional tools for medical students and residents that will be designed to enhance physicians' diagnostic reasoning processes. I am writing to invite your participation in this study.

During the study, you will be asked to assess and diagnose clinical scenarios using a computer-based system. As you assess the scenarios, you will be asked to verbalize all thoughts that come to your mind. Your thoughts will be audio-recorded in order to allow us to attempt to determine what cognitive processes you use during the diagnostic process. You will also wear eye-tracking gear that will track the movement of your eyes, as well as the size of your pupils.

As a token of our appreciation, each participant will receive $25 per hour. The study will last a maximum of 2 hours. Compensation will be based on a quarter hour basis, rounded up to the next quarter hour. For example, if the study takes between 1 hour and 15 minutes and 1.5 hours, you will be paid $37.50. If you complete the study in 1 hour and 40 minutes, you will be paid for 1.75 hours, receiving $43.75. The maximum amount you will be paid is for 2 hours, totaling $50.00. If you do not complete the entire study, you will be paid on a quarterly hour basis, rounded up to the next quarter hour, for the time they spent at the study location. For

example, if you decide after 1.5 hours to discontinue, you will be paid $37.50. If you decide after 40 minutes to discontinue the study, you will be paid $18.75 for 45 minutes.

If you are interested in participating, or have further questions, please contact me at (412) 977-3978, or send an e-mail to vlp2@pitt.edu

Sincerely,

Velma L. Payne, MS, MBA, MS
University of Pittsburgh School of Medicine
Department of Biomedical Informatics

STUDY ONE CONSENT FORM

Department of Biomedical Informatics
University of Pittsburgh School of Medicine

Title: Assessment of the use of Cognitive Heuristics during Diagnostic Reasoning

PI: Velma L Payne
Department of Biomedical Informatics
University of Pittsburgh School of Medicine
Parkvale Bldg M183
200 Meyran Ave Pittsburgh, PA 15260

Informational Script:

The purpose of this research study is to determine (1) if and how cognitive heuristics are used during diagnostic reasoning; (2) if inappropriate use of cognitive heuristics during diagnostic reasoning results in diagnostic errors. The results of this study will be used to develop new instructional tools for medical students and residents that will hopefully enhance physicians' diagnostic reasoning processes.

During the study, you will be asked to assess clinical scenarios using a computer-based system. You will see a portion of the clinical data and be asked to provide an initial diagnosis based on this data if you can. If you provide the initial diagnosis, you will be asked to specify what data you used to arrive at your diagnosis. The computer system will then display additional clinical data. You will have the opportunity to provide another initial diagnosis, and specify the data you used to arrive at that diagnosis. The computer system will then provide the remaining clinical data. At this time you will be asked to provide a final diagnosis and to specify the data you used to arrive at the final diagnosis.

The only risks associated with this project may be a breach of confidentiality, fatigue and/or frustration. In order to minimize the risk of breach of confidentiality, you will not be asked to provide any information that could lead to your identification. Participants will be randomly assigned a subject identification number by the computer-based system. Only this random ID number will appear on any research materials. The results of the study will be de-identified so that you will remain anonymous, i.e., you will not be identifiable in any way. All responses are confidential and will be kept under lock and key. In order to minimize the risk of fatigue and/or frustration, you can work at your own pace and take a break at any time. It is anticipated the study will last 2 hours. You can withdraw from the study at any time you feel unduly fatigued and/or frustrated.

As a token of our appreciation, each participant will receive $25 per hour. The study will last approximately 2 hours. Compensation will be based on a quarter hour basis, rounded up to the next quarter hour. For example, if the study takes between 1 hour and 15 minutes and 1.5 hours, you will be paid $37.50. If you complete the study in 1 hour and 40 minutes, you will be paid for 1.75 hours, receiving $43.75. If you do not complete the entire study, you will be paid on a quarterly hour basis, rounded up to the next quarter hour, for the time they

spent at the study location. For example, if you decide after 1.5 hours to discontinue, you will be paid $37.50. If you decide after 40 minutes to discontinue the study, you will be paid $18.75 for 45 minutes.

This study is being conducted by Velma L. Payne, a PhD student in the Department of Biomedical Informatics, University of Pittsburgh School of Medicine, who can be reached at (412) 977-3978 if you have any questions.

RIGHT TO WITHDRAW

Your participation is voluntary and you may withdraw from the project at any time. Your current and future status with the University or any other benefits to which you are entitled will be the same whether you participate in this study or not.

****

VOLUNTARY CONSENT

All of the above has been explained to me and all of my questions have been answered. I understand that, if not already done, I may request that my questions be answered by an investigator(s) involved in the research study. I also understand that any future questions I have about this research will be answered by the investigator(s) listed on the first page of this consent document at the telephone number(s) listed. Any questions I have about my rights as a research subject will be answered by the Human Subject Protection Advocate of the IRB Office, University of Pittsburgh (1-866-212-2668). By signing this form, I agree to continue to participate in this research study.

_____ _____

Subject Signature **Date**

INVESTIGATOR'S CERTIFICATION

I certify that I have explained this new information and its significance to the above individual and that any questions about this information have been answered.

_____ _____

Investigator's Signature **Date**

STUDY TWO CONSENT FORM

Department of Biomedical Informatics
University of Pittsburgh School of Medicine

Title: Assessment of the Effect of Feedback on use of Cognitive Heuristics during Diagnostic Reasoning

PI: Velma L Payne
Department of Biomedical Informatics
University of Pittsburgh School of Medicine Parkvale Bldg M183
200 Meyran Ave Pittsburgh, PA 15260

Informational Script:

The purpose of this research study is to determine (1) if and how cognitive heuristics are used during diagnostic reasoning; (2) if inappropriate use of cognitive heuristics during diagnostic reasoning results in diagnostic errors; and (3) if receiving feedback regarding your diagnostic reasoning strategies improves the appropriate use of cognitive heuristics and decreases medical errors. The results of this study will be used to develop new instructional tools for medical students and residents that will hopefully enhance physicians' diagnostic reasoning processes.

During the study, you will be asked to assess clinical scenarios using a computer-based system. You will see three data segments of clinical data. During the first two segments, you will be able to provide an initial diagnosis and to specify what data you used to arrive at your diagnosis. The computer system will then display the remainder of the clinical data. At this time you will be asked to provide a final diagnosis and to specify the data you used to arrive at the final diagnosis. While reviewing the scenarios, I would like you to verbalize all thoughts that come to your mind. This session will be recorded. The purpose of the recording is to allow us to attempt to determine what cognitive processes you use during diagnosis.

The only risks associated with this project are fatigue and/or frustration and a breach of confidentiality. In order to minimize the risk of fatigue and/or frustration, you can work at your own pace, take a break at any time, will be asked to participate for approximately 2 hours, and can withdraw from the study at any time you feel unduly fatigued and/or frustrated. In order to minimize the risk of breach of confidentiality, you will be randomly assigned a subject identification number by the computer-based system. Only this random ID number will appear on any research materials. The results of the study will be de-identified so that you will remain anonymous, i.e., you will not be identifiable in any way. All responses are confidential and will be kept under lock and key.

As a token of our appreciation, each participant will receive $50 per hour. The study will last approximately 2 hours. Compensation will be based on a quarter hour basis, rounded up to the next quarter hour and whole dollar amount. For example, if the study takes between 1 hour and 15 minutes and 1.5 hours, you will be paid $75.00. If you complete the study in 1 hour and 40 minutes, you will be paid for 1.75 hours, receiving $88.00. If you do not complete the entire study, you will be paid on a quarterly hour basis, rounded up to the next quarter hour, for the time they spent at the study location.

This study is being conducted by Velma L. Payne, a PhD student in the Department of Biomedical Informatics, University of Pittsburgh School of Medicine, who can be reached at (412) 977-3978 if you have any questions.

RIGHT TO WITHDRAW

Your participation is voluntary and you may withdraw from the project at any time. Your current and future status with the University or any other benefits to which you are entitled will be the same whether you participate in this study or not.

****

VOLUNTARY CONSENT

All of the above has been explained to me and all of my questions have been answered. I understand that, if not already done, I may request that my questions be answered by an investigator(s) involved in the research study. I also understand that any future questions I have about this research will be answered by the investigator(s) listed on the first page of this consent document at the telephone number(s) listed. Any questions I have about my rights as a research subject will be answered by the Human Subject Protection Advocate of the IRB Office, University of Pittsburgh (1-866-212-2668). By signing this form, I agree to continue to participate in this research study.

_____ _____

Subject Signature **Date**

INVESTIGATOR'S CERTIFICATION

I certify that I have explained this new information and its significance to the above individual and that any questions about this information have been answered.

_____ _____

Investigator's Signature **Date**

APPENDIX B

RESEARCH STUDY INSTRUMENT – SET OF CLINICAL SCENARIOS

Below is a list of the diagnosis and difficulty rating for each clinical scenario assessed by the subjects during both research studies.

Patient ID	Case Diagnosis	Organ System	Clinical Difficulty Scale: 1 - 7
052	Colon Cancer	GI/sm--lg Intestine	3.17
121	Myasthenia Gravis	Nervous System	3.17
091	Guillain-Barre Syndrome	Nervous System	3.33
032	Ulcerative Colitis	GI/sm--lg Intestine	3.50
033	Thrombotic Thrombocytopenic Purpura	Nervous System, Hem/Bone Marrow/Lymphoretic	3.67
042	Appendicitis	GI/sm--lg Intestine	3.67
043	Amoebic Liver Abscess	GI/Liver	3.67
113	Hemachromatosis	GI/Liver	3.67
133	Pernicious Anemia	Hem/Bone Marrow/Lymphoretic	3.67
171	Silicosis	Lung	3.67
062	Cryptococcal Meningitis	Nervous System	3.83
102	Pheochromocytoma	Endo/Adrenal	4.00
111	Mucormycosis	Head/Neck & Nervous System	4.00
143	Polymyalgia Rheumatica	Bones/joints	4.00
053	Crohn's Disease	GI/sm--lg Intestine	4.17
161	Porphyria (cutanea tarda)	Skin	4.17
021	Carcinoid Syndrome	GI/sm--lg Intestine	4.33
093	Subarachnoid Hemorrhage	Nervous System	4.50
022	Non-Hodgkins Lymphoma	Hem/Bone Marrow/Lymphoretic	4.67
122	Hypokalemic Periodic Paralysis	Nervous System	4.67
251	Amyloidosis (renal)	Renal	4.83
012	Metastic Hepatic Adeno (liver) Cancer	GI/Liver	5.00
083	Aortic Dissection	Heart/Blood vessels	5.17
181	Temporal Arteritis	Bones/joints	5.17
072	Hemolytic Uremic Syndrome	Renal	5.50

Patient ID	Case Diagnosis	Organ System	Clinical Difficulty Scale: 1 - 7
092	Osteomalacia	Bones/joints	5.50
063	Brucellosis	Hem/Bone Marrow/Lymphoretic	5.67
082	Miliary (disseminated) TB	Hem/Bone Marrow/Lymphoretic	5.67
103	Cardiac Amyloidosis	Heart/Blood vessels	5.67
011	Blastomycosis	Lung & Skin	5.83
001	Acromegaly	Endo/Pituitary	6.00
023	Hairy Cell Leukemia	Hem/Bone Marrow/Lymphoretic	6.00
052	Cryoglobulinemia	Renal	6.00
121	Syphilitic Meningitis	Nervous System	6.00
091	Whipple's Disease	GI/sm--lg Intestine	6.17
032	Gaucher's Disease	Hem/Bone Marrow/Lymphoretic	6.17

Division of Cases into Easy, Medium and Hard

	Patient ID	Case Diagnosis	Organ System	Clinical Difficulty
Easier Cases	52	Colon Cancer	GI	3.17
	121	Myasthenia Gravis	Nervous	3.17
	91	Guillain-Barre Syndrome	Nervous	3.33
	32	Ulcerative Colitis	GI	3.50
	33	Thrombotic Thrombocytopenic Purpura	Hem / Bone /Lymphoretic	3.67
	42	Appendicitis	GI	3.67
	43	Amoebic Liver Abscess	GI / Liver	3.67
	113	Hemachromatosis	GI / Liver	3.67
	133	Pernicious Anemia	Hem / Bone / Lymphoretic	3.67
	171	Silicosis	Lung	3.67
	62	Cryptococcal Meningitis	Nervous	3.83
	102	Pheochromocytoma	Endo / Adrenal	4.00
	111	Mucormycosis	Head / Neck/ Nervous	4.00
	143	Polymyalgia Rheumatica	Bones / Joints	4.00
Medium Cases	53	Crohn's Disease	GI	4.17
	161	Porphyria (cutanea tarda)	Skin	4.17
	21	Carcinoid Syndrome	GI	4.33
	93	Subarachnoid Hemorrhage	Nervous	4.50
	22	Non-Hodgkins Lymphoma	Hem / Bone / Lymphoretic	4.67
	122	Hypokalemic Periodic Paralysis	Nervous	4.67
	251	Amyloidosis (renal)	Renal	4.83
	12	Metastic Hepatic Adeno (liver) Cancer	GI / Liver	5.00
	83	Aortic Dissection	Heart / Blood	5.17
	181	Temporal Arteritis	Bones / Joints	5.17
	72	Hemolytic Uremic Syndrome	Renal	5.50
	92	Osteomalacia	Bones / Joints	5.50
Harder Cases	63	Brucellosis	Hem / Bone / Lymphoretic	5.67
	82	Miliary (disseminated) TB	Hem / Bone / Lymphoretic	5.67
	103	Cardiac Amyloidosis	Heart / Blood	5.67
	11	Blastomycosis	Lung / Skin	5.83
	1	Acromegaly	Endo / Pituitary	6.00
	23	Hairy Cell Leukemia	Hem / Bone / Lymphoretic	6.00
	31	Cryoglobulinemia	Renal	6.00
	123	Syphilitic Meningitis	Nervous	6.00
	112	Whipple's Disease	GI	6.17
	291	Gaucher's Disease	Hem / Bone / Lymphoretic	6.17

Experimental Study One Cases
In Order of Presentation

Patient ID	Case Diagnosis	Organ System	Clinical Difficulty Scale: 1-7
52	Colon Cancer	GI	3.17
91	Guillain-Barre Syndrome	Nervous	3.33
42	Appendicitis	GI	3.67
32	Ulcerative Colitis	GI	3.50
133	Pernicious Anemia	Hem / Bone / Lymphoretic	3.67
62	Cryptococcal Meningitis	Nervous	3.83
21	Carcinoid Syndrome	GI	4.33
93	Subarachnoid Hemorrhage	Nervous	4.50
122	Hypokalemic Periodic Paralysis	Nervous	4.67
83	Aortic Dissection	Heart / Blood	5.17
181	Temporal Arteritis	Bones / Joints	5.17
72	Hemolytic Uremic Syndrome	Renal	5.50
82	Miliary (disseminated) TB	Hem / Bone / Lymphoretic	5.67
103	Cardiac Amyloidosis	Heart / Blood	5.67
11	Blastomycosis	Lung / Skin	5.83
1	Acromegaly	Endo / Pituitary	6.00
31	Cryoglobulinemia	Renal	6.00
123	Syphilitic Meningitis	Nervous	6.00
92	Osteomalacia	Bones / Joints	5.50
102	Pheochromocytoma	Endo / Adrenal	4.00
53	Crohn's Disease	GI	4.17
12	Metastic Hepatic Adeno (liver) Cancer	GI / Liver	5.00
112	Whipple's Disease	GI	6.17
291	Gaucher's Disease	Hem / Bone / Lymphoretic	6.17

Experimental Study Two Cases
In Order of Presentation

Period	Patient ID	Case Diagnosis	Organ System	Clinical Difficulty (out of 7)
Pre-Test	52	Colon Cancer	GI	3.17
	21	Carcinoid Syndrome	GI	4.33
	42	Appendicitis	GI	3.67
	31	Cryoglobulinemia	Renal	6.00
	122	Hypokalemic Periodic Paralysis	Nervous	4.67
Feedback	11	Blastomycosis	Lung / Skin	5.83
	93	Subarachnoid Hemorrhage	Nervous	4.50
	32	Ulcerative Colitis	GI	3.50
	83	Aortic Dissection	Heart / Blood	5.17

Period	Patient ID	Case Diagnosis	Organ System	Clinical Difficulty (out of 7)
	1	Acromegaly	Endo / Pituitary	6.00
Post-Test	133	Pernicious Anemia	Hem / Bone / Lymphoretic	3.67
	181	Temporal Arteritis	Bones / Joints	5.17
	82	Miliary (disseminated) TB	Hem / Bone / Lymphoretic	5.67
	62	Cryptococcal Meningitis	Nervous	3.83
	103	Cardiac Amyloidosis	Heart / Blood	5.67

APPENDIX C

THINK-ALOUD CODING TRAINING INSTRUCTIONS

Background

For my dissertation, medical students and residents assessed 15 clinical cases while using a computer-based system. Each subject assessed the same cases, in the same order, over three periods – a pre-test, feedback and post-test period. Within each period they assessed five cases. The graphic below depicts this process.

Pre-Test Period	Feedback Period	Post-Test Period
Assess 5 cases **Thinking-Aloud During All Cases**	Assess 5 Cases Receive Information After Each Case	Assess 5 cases **Thinking-Aloud During All Cases**

An objective of my research is to determine how often the subjects use the cognitive heuristics **Anchoring and Adjustment** and **Confirmation Bias**. For each case solved, I will calculate (a) the average number of times the subject population *Anchored* on an initial diagnosis; (b) the manner in which the subjects *Adjusted* their diagnosis; and (c) the number of times they committed *Confirmation Bias*.

Anchoring and Adjustment

Anchoring occurs when a subject specifies an initial diagnosis for the case prior to identifying the final diagnosis.

Adjustment is how the subject transitions from the initial diagnosis to the final diagnosis. There are several types of Adjustment including:

- **Sufficient Adjustment** - if a subject specifies an incorrect initial diagnosis, then a correct final diagnosis
- **Insufficient Adjustment** - if a subject does not arrive at the correct final diagnosis (regardless of the initial diagnosis)
- **No Adjustment Necessary** - if a subject Anchors on the correct diagnosis and specifies the correct final diagnosis (they did not have to adjust their diagnosis)
- **Adjust Away From Correct** – when a subject Anchors on the correct initial diagnosis and specifies an incorrect final diagnosis.

Confirmation Bias

For each case there are critical pieces of information within the case that, if considered, should lead the subject to the correct diagnosis. Commonly when a subject Anchors, they seek information within the case to confirm their diagnosis and ignore information that disconfirms the diagnosis. Often times the critical information is ignored if it disconfirms an initial diagnosis. When the critical information is not considered, this is called **Confirmation Bias**. The formal definition for Confirmation Bias is when one ignores critical information that should lead them to the correct diagnosis.

During my study, when a subject specifies an initial and/or final diagnosis, they are required to specify the data they used to arrive at the final diagnosis by checking a checkbox adjacent to the data. If they do not select the

critical data within the case, it is assumed they did not use the critical data - this is considered commission of Confirmation Bias.

Think-Aloud Protocol Coding

As the subjects assessed the cases in the pre- and post-test periods, they were asked to think out loud and verbalize their thoughts as they reasoned through the cases. The subject's words were audio taped and transcribed. For each case, the transcripts need to be assessed to determine if the subjects Anchored, how they Adjusted, and if they committed Confirmation Bias. This assessment will involve "coding" the transcripts which will involve:

1. **Data Segmentation** – this involves breaking the subjects' words into segments – each segment should represent a single thought.
2. **Assign an Operator and Knowledge State** – this will involve looking for the following events and placing the correct codes on the line where the event occurs.
    a. Determine if the subject Anchored
    b. Determine how the subject Adjusted (based on the above categories)
    c. Determine if the subject committed Confirmation Bias (used the critical data)

When using the computer-based system, the subjects were required to enter an initial diagnosis, so there should always be an *Anchor* point for each case. However, the subject may have just typed the diagnosis and not verbalized it. Therefore, there may be cases where you will not find their *Anchor* point. If this is the case, you will not be able to determine how the subject *Adjusted* their diagnosis. Subjects also may not have verbalized the data items they used to arrive at their diagnosis (they may have just checked the boxes and didn't say what they were checking). Therefore, you may not be able to tell if they committed *Confirmation Bias*.

There are a total of 200 transcripts (cases) that need to be assessed (coded). My committee requires multiple coders to code a portion (30%) of the cases and the coding differences to be resolved. So, each coder will independently code the same 60 cases. I will compare the coding to determine the coding differences and compute an inter-rater reliability (level of agreement). I will then facilitate a meeting between the coders to resolve the coding differences. Once all the differences are resolved, the remaining 140 cases will be evenly divided amongst the coders for independent coding. Ten of these cases will be overlapped (coded by all coders) and another inter-rater reliability be calculated.

For this project there are two coders. You both will independently code 60 of the cases; I will review the coded transcripts, determine the level of agreement, note the differences and facilitate a meeting between you to discuss the differences. Once we all agree to the manner in which these 60 cases should be coded, the remaining 140 cases will be evenly divided between the two of you to independently code the remaining cases (70 each). Ten of these cases will also be coded by another coder so as to ensure the inter-rater reliability (level of agreement) remains at an acceptable level.

The list of cases to be coded, along with coding assignments can be found at the following url **http://velmalpayne.com/transcript_list/transcript_list.html** Each file on the website is clickable and will open a Word document which contains five cases. Also on the website (at the top of the page) you will find a file containing the correct diagnosis and critical data for each case (under the link "Case Diagnosis and Critical Data") as well as an Excel spreadsheet to enter a summary of your findings for each case – this file is further described below.

Data Segmentation

Before coding each case, you will need to segment the data into single thoughts. I have segmented the first file located on the website listed above as an example. For the 60 cases that both of you will code, I have assigned files 3 – 7 to Kayse to segment and files 8 – 12 to Tom to segment. I will segment file 2.

The transcripts that need to be segmented will contain text in paragraph form. When segmenting the data, simply press the *Enter* key after each segment to insert a new line character into the file. Each segment should be on a separate line. Then use the mouse and select all the lines you just created. Within Word (2007), select the "Table" menu, then select "Convert Text to Table", enter "4" for the number of columns, select "Auto" – this will convert your text into a table such as the table in File 1 on the above mentioned website. You can enter the data into a table this way, or by whatever means you feel is easiest.

Prior to coding complete the segmentation of these files, send me the files. I will put them on the website for coding

Assigning Operator and Knowledge State

For each segmented passage, you will assign an *Operator* and *Knowledge State*. The *Operators* that you use are listed in the table below. The Knowledge State is the specific item associated with the Operator.

Item	Operator	Explanation
1	Specified Initial Diagnosis (Anchor Point)	The point at which the subject specified an initial diagnosis
2	Extended Initial Diagnosis	If the subject added a disease to their initial diagnosis
3	Changed Initial Diagnosis	If the subject removed a disease from their initial diagnosis
4	Used Critical Data	If the subject mentioned the case critical data
5	Specified Final Diagnosis	When the subject states the final diagnosis

When assigning Operators and Knowledge States, it may be challenging to determine what the subject is reading from the screens and what verbiage is associated with their diagnostic reasoning. In order to assist you in determining the data they are reading vs. reasoning, I have provided a copy of the screens for each case on at this url **http://velmalpayne.com/screenshots/screen_shots.html** For each case, there are 3 screens. The links on this webpage are clickable and will display the screen. You will need to press the *Back* button on the browser to return to the list of screens. Some screens contain a large amount of data; so when you load them the text will be small. For each screen you can zoon in/out to adjust the size of the text. To do this, simply click on the screen to zoom in – this will most likely create scroll bars on the bottom and right side to allow you to scroll to see the additional data on the screen.

File 1 has been partially coded as an example of how to code the protocols. Each of you should finish coding the remaining cases in file 1, then move to files 2, 3, 4, etc. The "**Case Diagnosis and Critical Data**" file contains the correct diagnosis and critical data items for each case to assist you in determining the proper Adjustment value and to determine if the subject reviewed the critical data (committed Confirmation Bias). Some of the cases have multiple critical data items.

Once you have completed the coding for each case, please fill out the columns in the "**Coding Summary Excel Spreadsheet**" included on the website. In this file, enter your name, the file number being processed (found on the website left-most column), the case number within that file, the line number where the Anchor point was designated, the line number where the final diagnosis was specified, the line number (or numbers if there are multiple items) where the subject used the critical data (enter NA if they did not use the critical data), and the appropriate Adjustment category. Once you have completed coding a file, send me the coded file and the Excel spreadsheet via email so I can back them up.

When you open the files from the website, please save the files locally on your pc to complete the coding and update the Excel spreadsheet.

Don't forget to save the files from the website to your PC

BIBLIOGRAPHY

[1] Institute of Medicine. *To Err is Human: Building a Safer Health System*. Washington, D.C.: National Academy Press 2000.

[2] Graber M, Gordon R, Franklin N. Reducing Diagnostic Errors in Medicine: What is the Goal? . Academic Medicine. 2002;77(10):981-992.

[3] Leape LL, Brennan TA, Laird N, Lawthers AG, Localio R, Barnes BA, et al. The Nature of Adverse Events in Hospitalized Patients Results of the Harvard Medical Practice Study II. The New England Journal of Medicine. 1991;324(6):377-384.

[4] Baretlett E. Physicians' Cognitive Errors and Their Liability Consequences. Journal of Healthcare Risk management. 1998;Fall 1998:62-69.

[5] Tai D, El-Bilbeisi H, Tewari S, Mascha E, Wiedermann H, Arroliga A. A Study of Consecutive Autopsies in a Medical ICU: A Comparison of Clinical Cause of Death and Autopsy Diagnosis. Chest. 2001;119:530-536.

[6] Bordage G. Why Did I Miss the Diagnosis? Some Cognitive Explanations and Educational Implications. Academic Medicine. 1999;74(10):S138-S143.

[7] Elstein AS, Schwartz A. Clinical Reasoning in Medicine. In: Higgs J, Jones M, eds. *Clinical Reasoning in the Health Professions*. Woburn, Mass: Butterworth-Heinemann 1995:49-59.

[8] Kassirer JP, Kopelman RI. Cognitive Errors in Diagnosis: Instantiation, Classification and Consequences. The American Journal of Medicine. 1989;86(4):433-441.

[9] Parmley MC. The Effects of the Confirmation Bias on Diagnostic Decision-Making: Drexel University; 2006 (PhD Dissertation).

[10] Regehr G, Norman GR. Issues in Cognitive Psychology: Implications for Professional Education. Academic Medicine. 1996;71:988-1001.

[11] Schmidt HG, Norman GR, Boshuizen HP. A Cognitive Perspective on Medical Expertise: Theory and Implication. Academic Medicine. 1990;65;611.

[12] Kahnamen D, Tversky A. On the Psychology of Prediction. Psychological Review. 1973;80:237-51.

[13] Tversky A, Kahneman D. Availability: A Heuristic for Judging Frequency and Probability. Cognitive Psychology. 1973;5:207-232.

[14] Tversky A, Kahneman D. Judgment under Uncertainty: Heuristics and Biases. Science. 1974;185:1124-1131.

[15] Gilovich T, Griffin D. Heuristics and Biases: Then and Now. In: Gilovich T, Griffin D, Kahneman D, eds. *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York, NY: Cambridge University Press 2005.

[16] Friedman CP, Gatti GG, Franz TM, Murphy GC, Wolf FM, Heckerling PS, et al. Do Physicians Know When Their Diagnoses Are Correct? Journal of General Internal Medicine. 2005;20:334-339.

[17] Hershberger PJ, Markert RJ, Part HM, Cohen SM, Finger WW. Understanding and Addressing Cognitive Bias in Medical Education. Advances in Health Sciences Education. 1997;1:221-226.

[18] Hershberger, PJ; Part, HM; Markert, RJ; Cohen, SM; Finger, WW. Development of a Test of Cognitive Bias in Medical Decision-Making. Academic Medicine. 1994; 69(10):839-842

[19] Payne VL, Crowley RS. Assessing Use of Cognitive Heuristic Representativeness in Clinical Reasoning. In: Association AMI, editor. American Medical Informatics Association; 2008; Washington, D.C.: American Medical Informatics Association; 2008. p. 571-575.

[20] Pohl R, Eisenhauer M, Hardt O. SARA: A Cognitive Process Model to Simulate the Anchoring Effect and Hindsight Bias. Memory. 2003;11(4/5):337-356.

[21] Poses RM, Anthony M. Availability, Wishful Thinking, and Physicians' Diagnostic Judgments for Patients with Suspected Bacteremia. Medical Decision-Making. 1991;11:159-168.

[22] Redelmaier D, Shafir E. Medical Decision-Making in Situations That Offer Multiple Alternatives. Journal of the American Medical Association. 1995;273(4):302-305.

[23] Redelmeier D. The Cognitive Psychology of Missed Diagnoses. Annals of Internal Medicine. 2005;142:115-120.

[24] Redelmeier D, Ferris L, Tu J, Hux J, Schull M. Problems for Clinical Judgment: Introducing Cognitive Psychology as One More Basic Science. Canadian Medical Association. 2001;164(3):358-360.

[25] Roswarski TE, Murray MD. Supervision of Students May Protect Academic Physicians from Cognitive Bias: A Study of Decision-making and Multiple Treatment Alternatives in Medicine. Medical Decision-Making. 2006;26:154-161.

[26] Hofer TP, Kerr EA, Hayward RA. What is an Error? Effective Clinical Practice. 2000;3:261-269.

[27] Brennan TA, Leape LL, Laird NM, Herbert L, Localio R, Lawthers A, et al. Incidence of Adverse Events and Negligence in Hospitalized Patients. The New England Journal of Medicine. 1991;324(6):370-6.

[28] Thomas EJ, Studdert DM, Burstin HR, Orav EJ, Zeena T, Williams EJ, et al. Incidence and Types of Adverse Events and Negligent Care in Utah and Colorado. Medical Care. 2000;38(3):261-271.

[29] Wilson RM, Runciman WB, Gibberd RW, Harrison BT, Newby L, Hamilton JD. The Quality in Australian Health Care Study. The Medical Journal of Australia. 1995;163:458-471

[30] Wilson RM, Harrison BT, Gibberd RW, Hamilton JD. An analysis of the causes of adverse events from the Quality in Australian Health Care Study. The Medical Journal of Australia. 1999;170:411-415

[31] Bates DW, Cullen DJ, Laird N, Petersen LA, Small SD, Servi D, et al. Incidence of Adverse Drug Events and Potential Adverse Drug Events: Implications for Prevention. JAMA 1995:274:29-34

[32] Sox HC, Woloshin S. How many deaths are due to medical error? Getting the number right. Effective Clinical Practice. 2000;6:277-283.

[33] Hayward RA, Hofer TP. Estimating hospital deaths due to medical errors: Preventability is in the eye of the reviewer. JAMA. 2001;286(4):415-420.

[34] McDonald CJ, Weiner M, Hui SL. Deaths due to medical errors are exaggerated in Institute of Medicine Report. JAMA. 2000;284(1):93-95.

[35] Graber ML, Franklin N, Gordon R. Diagnostic Error in Internal Medicine. Archives of Internal Medicine. 2005;165:1493-1499.

[36] Graber ML. Next steps: envisioning a research agenda. Advances in Health Sciences Education. 2009;Online DOI10.1007/s10459-009-9183-1(Online print only).

[37] Graber M. Diagnostic error in medicine: a case of neglect. Joint Commission Journal on Quality and Patient Safety. 2005;21(2):106-113.

[38] Shojania K, Burton E, McDonald K, Goldman L. Changes in rates of autopsy-detected diagnostic errors over time: a systematic review. JAMA. 2003;289(21):2849-2856.

[39] Kachalia A, Gandhi TK, Puopolo AL, Yoon C, Thomas EJ, Giffey R, et al. Missed and delayed diagnoses in the emergency department: a study of closed malpractice claims from four liability insurers. Annals of Emergency Medicine. 2007;49(2):196-205.

[40] Leape L, Berwick D, Bates D. Counting deaths due to medical errors. JAMA. 2002;288(19):2405.

[41] Chimowitz MI, Logigian EL, Caplan LR. The accuracy of bedside neurological diagnoses. Annals of Neurology. 1990;28(1):78-85.

[42] Weinberg N, Statson W. Managing quality in hospital practice. International Journal of Quality Healthcare. 1998;10:295-302.

[43] Gruver R, Freis E. A study of diagnostic errors. Annuals of Internal Medicine. 1957;47:108-120.

[44] Plous S. Heuristics and Biases. The Psychology of Judgment and Decision-making. New York, NY: McGraw-Hill, Inc. 1993.

[45] Tversky A, Kahnamen D. Availability: A Heuristic for Judging Frequency and Probability. In D Kahneman, P Slovic and A Tversky (Eds), *Judgment Under Uncertainty: Heuristics and Biases* 1982:163-178.

[46] Tversky A, Kahneman D. Judgments of and by representativeness. In: Tversky A, Kahneman D, eds. *Judgment under uncertainty: Heuristics and Biases.* Cambridge, England: Cambridge University Press 1982:84-98.

[47] Tversky A, Kahneman D. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review. 1983;90:293-315.

[48] Croskerry P. Achieving quality in clinical decision-making: cognitive strategies and detection of bias. Academic Emergency Medicine. 2002;9:1184-1204.

[49] Kassirer JP, Kopelman RI. Learning Clinical Reasoning. Baltimore, Maryland: Lippincott Williams and Wilkins 1991.

[50] Kahneman D, Slovic P, Tversky A. *Judgment under uncertainty: Heuristics and biases*. New York, NY: Cambridge University Press 1982.

[51] Elstein AS. Heuristics and Biases: Selected Errors in Clinical Reasoning. Academic Medicine. 1999;74:791-794.

[52] Dawson N, Arkes HR. Systematic errors in medical decision-making: Judgment limitations. Journal of General Internal Medicine. 1987;2:183-187.

[53] Kempainen RR, Migeon MB, Wolf FM. Understanding our mistakes: a primer on errors in clinical reasoning. Medical Teacher. 2003;25(2):177-181.

[54] Croskerry P. The Cognitive Imperative: Thinking About How We Think. Academic Emergency Medicine. 2000 November;7(11):1223-1231.

[55] Elstein AS, Holzman GB, Ravitch MM, Metheny WA, Holmes MM, Hoppe RB, et al. Comparison of physicians' decisions regarding estrogen replacement therapy for menopausal women and decisions derived from a decision analytic model. American Journal of Medicine. 1986;80:246.

[56] Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Academic Medicine. 2003;78:775-780.

[57] Aberegg SK, Haponik EF, PB Terry. Omission bias and decision-making in pulmonary and critical care medicine. Chest. 2005; 128:1497-1505

[58] Gruppen, LD; Margolin, J; Wisdom, K; Grum, CM. Outcome bias and cognitive dissonance in evaluating treatment decisions. Academic Medicine. 1994; 69, S57-S59

[59] Arkes, HR; Saville, PD; Wortmann, RL; Harkness, AR. Hindsight bias among physicians weighing the likelihood of diagnoses in J Dowie, A Elstein, *Professional Judgment: A reader in clinical decision-making*. New York: Cambridge, University Press, 1998; 374-378

[60] Christensen, C; Heckerling, PS; Mackesy, ME; Bernstein, LM; Elstein, AS. Framing bias among expert and novice physicians. Academic Medicine. 1991; 66, S76-S78

[61] Detmer, DE; Fryback, DG; Gassner, K. Heuristics and biases in medical decision-making. Journal of Medical Education. 1978. 53; 682-683

[62] Moskowitz, AJ; Kuipers, BJ; Kassirer, JP. Dealing with uncertainty, risks and tradeoffs in clinical decisions. Annals of Internal Medicine. 1988. 108; 435-449

[63] Klein, JG. Five pitfalls in decisions about diagnosis and prescribing. British Medical Journal. 2005. 330:2; 781-784

[64] Pines, JM; Profiles in Patient Safety: Confirmation Bias in Emergency Medicine. Academic Emergency Medicine. 2006. 13:1; 90-94

[65] Voytovich A, Rippey R, Suffredini A. Premature conclusions in diagnostic reasoning. Journal of Medical Education. 1985;60:302-307.

[66] Richards, MS; Wierzbicki, M. Anchoring effects in clinical like judgments. Journal of Clinical Psychology. 1990. 46:3, 358-365

[67] Hertwig, R; Pachur, T; Kurzenhauser, S. Judgments of risk frequencies: Tests of possible cognitive mechanisms. Journal of Experimental Psychology. 2005. 31:4, 621-642

[68] Ellis, MV; Robbins, ES; Schult, D; Ladany, N; Banker, J. Anchoring errors in clinical judgments: Type I error, adjustment or mitigation? Journal of Counseling Psychology. 1990. 37:3, 343-351

[69] Friedlander, ML; Stockman, SJ. Anchoring and publicity effects in clinical judgment. Journal of Clinical Psychology. 1983. 39:4, 637-644

[70] Friedlander, ML; Phillips, SD. Preventing anchoring errors in clinical judgment. Journal of Consulting and Clinical Psychology. 1984. 52:3, 366-371

[71] Arkes, HR. Impediments to accurate clinical judgment and possible ways to minimize their impact. Journal of Counseling and Clinical Psychology. 1981. 49:3, 323-330

[72] Arkes, HR; Faust, D; Guilmette, TJ; Hart, K. Eliminating hindsight bias. Journal of Applied Psychology. 1988. 73:2, 305-307

[73] Eva, KW; Norman GR. Heuristics and biases – a biased perspective on clinical reasoning. Medical Education. 2005. 39, 870-872

[74] Wolf FM, Gruppen LD, Billi JE. Differential diagnosis and the competing-hypotheses heuristic: A practical approach to judgment under uncertainty and Bayesian probability. Journal of American Medical Association. 1985;253(19):2858-2862.

[75] Hastie R, Dawes RM. *Rational Choice in an Uncertain World*. Thousand Oaks, CA: Sage Publications, Inc 2001.

[76] von Newmann J, Morgenstern O. *Theory of games and economic behavior.* Princeton, NU: Princeton University Press 1947.

[77] Dowie J, Elstein A. *Professional Judgment: A Reader in Clinical Decision-making*. Cambridge, MA: Cambridge University Press 1988.

[78] Kuipers B, Moskowitz AJ, Kassirer JP. Critical Decisions Under Uncertainty: Representation and Structure. Cognitive Science. 1988;12:177-210.

[79] Eddy D. Probabilistic Reasoning in Clinical Medicine: Problems and Opportunities. In: Kahneman D, Slovic P, Tversky A, eds. *Judgment under Uncertainty: Heuristic and Biases.* Cambridge, MA: Cambridge University Press 1982.

[80] Simon H. A Behavioral Model of Rational Choice. *Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting*. New York, NY: Wiley 1957.

[81] Eddy DM, Clanton CH. The art of diagnosis: solving the clinicopathological exercise. New England Journal of Medicine. 1982;306:1263-1268.

[82] Harris JM. The hazards of bedside Bayes. Journal of American Medical Association. 1981;246(22):2602-2605.

[83] Rathbun SW, Raskob GE, Whisett T. Sensitivity and specificity of helical computed tomography in the diagnosis of pulmonary embolism: a systematic review. Annuals of Internal Medicine. 2000;132(3):227-232.

[84] Graber ML. Educational strategies to reduce diagnostic error: Can you teach this stuff? Advances in Health Sciences Education. 2009; Only available online DOI 10.1007/s10459-009-9178-y: Available online.

[85] Poses RM, Cebul RD, Collins M, Fager Ss. The ability of experienced physicians' probability estimates for patients with sore throats. JAMA. 1985; 254:925-929

[86] Morris AH. Developing and implementing computerized protocols for standardization of clinical decisions. Annals of Internal Medicine. 2000;132(5):373-383.

[87] Edwards F, Davies R. Use of a Bayesian algorithm in the computer-assisted diagnosis of appendicitis. Journal of Pediatric Ophthalmol and Strabismus. 1984;158:219-222

[88] Velanovich V. Bayesian analysis of the reliability of peritoneal lavage. Surgical Gynecology and Obstetrics. 1990;170:7-1711

[89] Popescu I, Jovin G, Vasilescu C, Esanu C. The value of ecography for the diagnosis of acute cholecystitis (a Bayesian approach). Theoretical Surgery. 1992;7(1):10-13

[90] Place R, Velanovich V, Carter P. Fine needle aspiration in the clinical management of mammary masses. Surgical Gynecology and Obstetrics. 1993;177(1):7-11.

[91] Rifkin R, Hood W. Bayesian analysis of electrocardiographic exercise stress testing. New England of Medicine. 1977;297:681-696.

[92] Mai M, Henrich G, Larmon DV. Application of a Bayes program for classification of comma. Methods of Information in Medicine. 178;17:41-46

[93] Christensen-Szalanski J JJ Bushyhead JB. Physicians' use of probabilistic information in a real clinical setting. Journal of Experimental Psychology. 1981;7(4):928-935

[94] Johnson-Laird PN. Mental Models and Deduction. Trends in Cognitive Science. 2001;5(10):434-442.

[95] Stanovich KE. *Who is Rational? Studies of Individual Differences in Reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates 1999.

[96] Rips LJ. *The Psychology of Proof: Deductive Reasoning in Human Thinking*. Cambridge, MA: MIT Press 1994.

[97] Braine MDS, O'Brien DP. *Mental Logic*. Mahwah, NJ: Lawrence Erlbaum Associates 1998.

[98] Jeffrey R. *Formal Logic: Its Scope and Limits*. McGraw-Hill 1981.

[99] Johnson-Laird, PN, Byrne, RMJ. *Deduction.* Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc. 1991.

[100] Bara BG. In favor of a unified model of deductive reasoning. In: Madruga JAG, ed. Mental *Models in Reasoning*. Madrid Universidad Nacional de Educacion a Distancia 2000:69-81.

[101] Craik KJW. *The Nature of Explanation*. Cambridge University Press 1943.

[102] Johnson-Laird PN. Imagery, visualization, and thinking. In: Hochberg H, ed. *Perception and Cognition at Centruy's End*. Academic Press 1998:441-67.

[103] Johnson-Laird PN. Mental Models and Thought. In: Holyoak KJ, Morrison RG, eds. *The Cambridge Handbook of Thinking and Reasoning*. New York, NY: Cambridge University Press 2005:185-208.

[104] Baddeley A. *Working Memory*. Oxford University Press, USA 1987.

[105] Halford GS, Wilson WH, Phillips S. Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. Behavioral and Brain Sciences. 1998;21:803-864.

[106] Halford GS, Wilson WH, Phillips S. Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. Behavioral and Brain Sciences. 1998;21:803-864.

[107] Carreiras M, Santamaria C. Reasoning about relations: spatial and non-spatial problems. Think Reason. 1997;3:191-208.

[108] Knauff M, Rauh R, Schlieder C, Strube G. Mental Models in Spatial Reasoning. Spatial Cognition. Heidelberg: Springer Berlin 1998:267-291.

[109] Vandierendonck A, Vooght GD. Working memory constraints on linear reasoning with spatial and temporal contents. Quarterly Journal of Experimental Psychology. 1997;50A:803-820.

[110] Sloutsky VM, Johnaon-Laird PN. Problem representations and illusions in reasoning. In: Hahn M, Stones S, editors. 21st Annual Conference of the Cognitive Science Society; 1999: Erlbaum; 1999. p701-705

[111] Espino O, Santamaria C, Meseguer E, Carreiras M. Eye movements during syllogistic reasoning. In: Madruga JAG, Carriedo N, Gonzalez-Labra MJ, eds. *Mental Models in Reasonin*g. Madrid, Spain: Universidad Nacional de Educacion a Distancia 2000:179-188.

[112] Newstead SE, Griggs RA. Premise misinterpretation and syllogistic reasoning. Quarterly Journal of Experimental Psychology. 1999;52A:1057-1075.

[113] Ormerod TC. Mechanisms and strategies for rephrasing. In: Schaeken W, Vooght GD, Vandierendonck A, d'Ydewalle G, eds. *Deductive Reasoning and Strategies*. Erlbaum 2000.

[114] Sloutsky VM, Goldvarg Y. Effects of externalization on representation of indeterminate problems. In: Hahn M, Stones S, editors. Annual Conference of the Cognitive Science Society; 1999: Erlbaum; 1999.

[115] Stanovich KE, West RF. Cognitive ability and variation in selection task performance. Think Reason. 1998;4:193-230.

[116] Liberman N, Klar Y. Hypothesis testing in Wason's selection task: social exchange cheating detection or task understanding. Cognition. 1996;58:127-156.

[117] Roberts MJ, Gilmore DJ, Wood DJ. Individual differences and strategy selection in reasoning. British Journal of Psychology. 1997;88:473-492.

[118] Bond W, Dettrick L, Eberhardt M, Barr G, Kane B, Worrilow C, Arnold D, Croskerry P. Cognitive versus Technical Debriefing After Simulation Training. Academic Emergency Medicine. 2006;13:276-283.

[119] Croskerry P. Cognitive Forcing Strategies in Clinical Decision-making. Annals of Emergency Medicine. 2003;41(1):110-120.

[120] Croskerry P. The Theory and Practice of Clinical Decision-Making. Canadian Journal of Anesthesia. 2005;52(6):R1-R8.

[121] Croskerry P, Norman G. Overconfidence in Clinical Decision-making. Naples Collaborative. Naples 2007.

[122] Fischhoff B. Debiasing. In: Kahneman D, Slovic P, Tversky A, eds. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge, MA: Cambridge University Press 1982:422-444.

[123] Fong GT, Krantz DH, Nisbett RE. The effects of statistical training on thinking about everyday problems. Cognitive Psychology. 1986; 18:253-292.

[124] Hirt ER, Markman KD. Multiple Explanation: A Consider-an-Alternative Strategy for Debiasing Judgments. Journal of Personality and Social Psychology. 1995;69(6):1069-1086.

[125] Hodgkinson GP, Brown NJ, Maule AJ, Glaister KW, Pearman AD. Breaking the Frame: An Analysis of Strategic Cognition and Decision-making Under Uncertainty. Strategic Management Journal. 1999;20:977-985.

[126] Koriat A, Bjork RA. Mending Metacognitive Illusions: A Comparison of Mnemonic-Based and Theory-Based Procedures. Journal of Experimental Psychology: Learning, Memory and Cognition. 2006;32(5):1133-1145.

[127] Kosonen P, Winne PH. Effects of Teaching Statistical Laws of Reasoning About Everyday Problems. Journal of Educational Psychology. 1995;87(1):33-46.

[128] McKenzie CRM. Increased Sensitivity to Differentially Diagnostic Answers Using Familiar Materials: Implications for Confirmation Bias. Memory & Cognition. 2006;34(3):577-588.

[129] Mumma GH, Wilson SB. Procedural Debiasing of Primacy/Anchoring Effects. Journal of Clinical Psychology. 1995;51(6):841-853.

[130] Nisbett R, Krantz D, Jepson C, Fong G. Improving Inductive Inference. In: Kahneman D, Slovic P, Tversky A, eds. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge, MA: Cambridge University Press 1982.

[131] Simon AF, Fagley NS, Halleran JG. Decision Framing: Moderating Effects of Individual Differences and Cognitive Processing. Journal of Behavioral Decision-making. 2004;17:77-93.

[132] Wolf FM, Gruppen LD, Billi JE. Use of Competing-Hypotheses Heuristic to Reduce 'Pseudodiagnosticity'. Journal of Medical Education. 1988;63:548-554.

[133] Faul, Erdfelder, Lang, Buchner, G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences, 2007.

[134] Gonzalez C. Decision support for real-time, dynamic decision-making tasks. Organizational Behavior and Human Decision Processes. 2005;96:142-154.

[135] Sengupta K, Abdel-Hamid TK. Alternative Conceptions of Feedback in Dynamic Decision Environments: An Experimental Investigation. Management Science. 1993;39(4):411-428.

[136] Edwards W. Dynamic decision theory and probabilistic information processing. Human Factors. 1962;4:59-73

[137] Brehmer B. In one word: Not from experience. Acta Psychologica. 1980;45:223-241.

[138] Balzer WK, Doherty ME, Raymond O'Connor J. Effects of cognitive feedback on performance. Psychological Bulletin. 1989;106(3):410-433.

[139] Lerch FJ, Harter DE. Cognitive support for real-time dynamic decision-making. Information Systems Research. 2001;12(1):63-82.

[140] Kuhn G. Diagnostic Errors. Academic Emergency Medicine. 2002;9:740-750.

[141] Dubeau E, Voytovich A, Rippey R. Premature conclusions in the diagnosis of iron-deficiency anima: cause and effect. Medical Decision-making. 1986;6:169-173.

[142] Croskerry P, Cosby KS, Schenkel SM, Wears RL. *Patient Safety in Emergency Medicine*. Philadelphia, PA: Lippincott Williams & Wilkins 2009.

[143] Elstein AS. Thinking about diagnostic thinking: a 30-year perspective. Advances in Health Sciences Education. 2009;Online publication DOI 10.1007/s10459-009-9184-0(Online publication).

[144] Schiff GD, Kim S, Abrams R, Cosby K, Lambert B, Elstein AS, et al. Diagnosing diagnosis errors: Lessons from a multi-institutional collaborative project. Advances in Patient Safety. 2005;2:255-278.

[145] Croskerry P. The Feedback Sanction. Academic Emergency Medicine. 2000;7(11):1232-1238.

[146] JAMA evidence website. Using evidence to improve care located at http://www.jamaevidence.com/search/result/57521

[147] Corbett A, Anderson J. Locus of feedback control in computer-based tutoring: impact on learning rate, achievement and attitudes. In: Jacko J, Sears A, Beaudouin-Lafon M, Jacob R, editors. SIG CHI 2001 *Conference on Human Factors in Computing Systems*; 2001; Seattle, Washington, New York: ACM Press; 2001. p. 245-252.

[148 Crowley RS, Legowski E, Medvedeva O, Tseytlin E, Roh E, Jrkic D. Evaluation of an Intelligent Tutoring System in Pathology: Effects of External Representation on Performance Gains, Metacognition and Acceptance. Journal of the American Medical Informatics Association 2007;14(2):182-190.

[149] El Saadawi GM, Azevedo R, Castine M, Payne V, Medvedeva O, Tseytlin E, et al. Factors affecting feeling-of-knowing in a medical intelligent tutoring system: the role of immediate feedback as a metacognitive scaffold. Advances in Health Sciences Education. 2009;Online print only.

[150] Friedman CP, Elstein AS, Wolf FM, Murphy GC, Franz TM, Heckerling PS, et al. Enhancement of Clinicians' Diagnostic Reasoning by Computer-Based Consultation: A Multisite Study of 2 Systems. JAMA. 1999;282(19):1851-1856.

[151] Sherbino J, Dore K, Siu E, Upadhye S, Norman G. Diagnostic error in emergency medicine: The effectiveness of Cognitive Forcing Strategies. Working Paper. 2009:1-20

[152] Croskerry P. A Universal Model of Diagnostic Reasoning. Academic Medicine. 2009;84(8):1022-1028.

[153] Schneider W, Chein JM. Controlled and automatic processing: behavior, theory, and biological mechanisms. Cognitive Science. 2003;27:5525-5559.

[154] Atkinson RC, Shiffrin RM. Human memory: A proposed system and its control processes. In: Spence KW, Spence Jt, eds. *The psychology of learning and motivation: Advances in research and theory*. New York, NY: Academic Press 1968.

[155] Shiffrin RM, Gardner GT. Visual processing capacity and attentional control. Journal of Experimental Psychology. 1972;93(1):72-82.

[156] Shiffrin RM, McKay DP, Shaffer WO. Attending to forty-nine spatial positions at once. Journal of Experimental Psychology. 1976;93(1):14-22.

[157] Schneider W, Shiffrin RM. Controlled and automatic human information processing. I. Detection, search, and attention. psychological Review. 1977;84(1):1-66.

[158] Shiffrin RM, Schneider W. Controlled and automatic human information processing. II. Perceptual learning, automatic attending and a general theory. Psychological Review. 1977;84(2):127-190.

[159] Anderson JR. Automaticity and the ACT* Theory. American Journal of Psychology. 1992;105(2):165-180.

[160] Logan GD. Attentional and automaticity in Stroop and priming tasks: Theory and data. Cognitive Psychology. 1980;12(4):523-553.

[161] Pashler H, Johnston JC, Ruthruff E. Attention and performance. Annual Review of Psychology. 2001;52:629-651.

[162] Stanovich KE. The impact of automaticity theory. Journal of Learning Disabilities. 1987;20(3):167-168.

[163] Evans JSBT. Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. Annual Review of Psychology. 2008;59:255-278.

[164] Norman G. Dual processing and diagnostic errors. Advances in Health Sciences Education. 2009;Available online. DOI 10.1007/s10459-009-9179-x. At this time, available online only.

[165] Tulving E. How many memory systems are there? American Psychologist. 1985;40:385-398.

[166] Klein G. *Sources of power*. Cambridge, Mass: MIT Press 1999.

[167] Ambady N, Rosenthal R. Thin slices of behavior as predictors of interpersonal consequences: A meta-analysis. Psychological Bulletin. 1992;2:256-274.

[168] Ullman DG. *Making robust decisions: Decision management for technical, business and service teams*. Victoria, BC: Trafford Publishing, 2006

[169] Groves JE. Taking care of the hateful patient. New England Journal of Medicine. 1978;298:883-887

[170] Zajonc RB. Feeling and thinking: Preferences need no inferences. American Psychologist. 1980;35:151-157

[171] Stanovich KE. *The Robot's Rebellion: Finding Meaning in the Age of Darwin*. Chicago, Ill: The University of Chicago Press 2004.

[172] Hammond K. *Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice*. New York, NY: Oxford University Press 2000.

[173] Croskerry P. Clinical cognition and diagnostic error: Applications of a dual process model of reasoning. Advances in Health Sciences Education. 2009; DOI: 10.1007/x10459-009-9182-2. Available online.

[174] Pretz JE. Intuition versus analysis: Strategy and experience in complex everyday problem-solving. Memory and Cognition. 2008;36:554-566

[175] Coderre S, Mandin H, Harasym P, Fick G. Diagnostic reasoning strategies and diagnostic success. Medical Education. 2003;37:695-703

[176] Eva K, Regehr G. Self-assessment in the health professions: A reformulation and research agenda. Academic Medicine. 2005;80:S46-S54

[177] Norman G, Brooks L, Allen S, Rosenthal D. Sources of observer variation in dermatologic diagnosis. Academic Medicine. 1990;64:S19-S21

[178] Oskamp S. Overconfidence in case study judgments. Journal of Consulting Psychology. 1965;29:261-265

[179] Redelmeier D, Shafir E, Aujla P. The beguiling pursuit of more information. Journal of Medical Decision-making. 2001;21:376-381

[180] Wason P. On the failure to eliminate hypotheses in a conceptual task. Quarterly Journal of Experimental Psychology. 1960;12;129-140

[181] Groopman J. *How Doctors Think*. New York, NY: Houghton Mifflin Company 2007.

[182] Coderre S, Mandin H, Harasym P, Fick G. Diagnostic reasoning strategies and diagnostic success. Medical Education. 2003;37:695-703

[183] Schmidt H, Norman G, Boshuizen H. A cognitive perspective on medical expertise: Theory and implications. Academic Medicine. 1990;65:611-621

[184] Lakoff G. Cognitive models and prototype theory. In: Neisser U, ed. Concepts and conceptual development. Cambridge, England: Cambridge University Press 1987:63-102

[185] Schmidt H, Boshuizen H, Hobus P. Transitory stages in the development of medical expertise: the "Intermediate Effect" in clinical case representation studies. Proceedings of the 10th Conference on Cognitive Science Society; 1988; Hillsdale, New Jersey: Erlbaum; 1988. p 139-145

[186] Feltovich P, Barrows H. Issues of generality in medical problem solving. In: Schmidt H, Devolder M, eds. *Tutorials in problem-based learning: A new direction in teaching the health professions*. Assen, The Netherlands: Van Gorcum 1984:128-142

[187] Claessen H, Boshuizen H. Recall of medical information by students and doctors. Medical Education. 1985;19:61-67

[188] Groen G, Patel VL. The relationship between comprehension and reasoning in medical expertise. In : MTH Chi et al., ed. *The nature of expertise*. Hillsdale, New Jersey: Erlbaum 1988.

[189] Coughlin L, Patel VL. Text comprehension and expertise in the domain of medicine. Annual meeting of the American Educational Research Association; 1986; San Francisco, California; 1986.

[190] Norman G, Brooks L, Allen S. Recall by expert medical practitioners and novices as a record of processing attention. Journal of Experimental Psychology: Learning, Memory and Cognition. 1989;13:1166-1174

[191] Schmidt H, Hobus P, Patel V, Boshuizen H. Contextual factors in the activation of first hypotheses: Expert-novice differences. Annual meeting of the American Educational Research Association; 1987; Washington, D.C. 1987

[192] Hobus P, Hofstra m, Boshuizen H, Schmidt H. De context van de klatch als diagnosticum [The context of a complaint as a diagnostic tool]. Huisarts en Wetenschap. 1988;31:261-267

[193] Bordage G, Zacks R. The structure of medical knowledge in the memories of medical students and general practitioners: Categories and prototypes. Medical Education. 1984;18:406-416

[194] Hassebrock F, Pretula M. Autobiographical memory in medical problem solving. American Educational Research Association; Boston Massachusetts; 1990

[195] Van Rossum H, Bender W. What can be learned from a boy with acute appendicitis? Persistent effects of a case presentation on the diagnostic judgment of family doctors. Fourth Ottawa Conference; 1990; Ottawa, Ontario, Canada; 1990

[196] Allen S, Brooks L, Norman G. Effect of prior examples on rule-based diagnostic performance. Proceedings of the Twenty-fifty Annual Conference on Research in Medical Education; 1988; Washington, D.C.: Association of American Medical Colleges; 1988

[197] Hansson SO. Decision Theory. A Brief Introduction. Stockholm 2005.

[198] Wikipedia. Decision Theory. http://en.wikipedia.org/wiki/Decision_theory

[199] Baddeley, Alan. Working Memory: The Interface between Memory and Cognition. Journal of Cognitive Neuroscience. 1992:4(3):281-288

[200] Sherbino J, Dore K, Siu E, Upadhye S, Norman G. Diagnostic Error in Emergency Medicine: The Effectiveness of Cognitive Forcing Strategies. Working Paper. 2009

[201] Dragononiu A, Robinson D, Patel B, Patel V. Cognitive Biases in Decision Making by Clinicians. AMIA 2009 Poster. AMIA-0677-A2009.R1.

[201] Sherbino J, Norman G. Diagnostic Error in Emergency medicine: The de-biasing effect of Cognitive Forcing Strategies. Poster presented at the Canadian Association of Emergency Physicians (CAEP). 2009.

[202] El Saadawi GM; Tseytlin E, Legowski E, Jukic D, Castine M, Fine J, Gormley R, Crowley RS; A natural language intelligent tutoring system for training pathologists: implementation and evaluation. Advances in Health Sciences Education. 2008; 13:709-722

[203] Mello-Thoms C, Hardesty L, Sumkin J, Ganott M, Hakim C, Britton C, Stalder J, Maitz G; Effects of lesion conspicuity on visual search in mammogram reading. Academic Radiology. 2005; 12:830-840.

[204] Just M, Carpenter P, Miyake A. Neurindicies of cognitive workload: neuroimaging, pupillometric and event-related potential studies of brain work. Theoretical Isses Ergom Sci. 2003;4:56-88.

[205] Backs R, Walrath L. Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. Appl Ergon. 1992;23:243-254.

[206] Ericsson K. An expert-performance perspective of research on medical expertise: The study of clinical performance. Medical Education. 2007;41:1124-1130

[207] Payne VL, Medvedeva O, Legowski E, Castine M, Tseytlin E, Jukic D, Crowley RS. Effect of a limited-enforcement intelligent tutoring system in dermatopathology on student errors, goals and solution paths. Artificial Intelligence in Medicine. 2009;47:175-197

[208] Patton M. *Qualitative Research & Evaluation Methods*. 3rd Edition. California: Sage Publications. 2002; pp 385

[209] Ericsson K, Simon H. *Protocol Analysis; Overview of Methodology of Protocol Analysis* (Revised Edition) Massachusetts: MIT Press 1993

[210] Ericsson K, Simon H. *Protocol Analysis: Verbal Reports as Data*. Massachusetts: MIT Press 1982

[211] Elstein AS, Shulman LS, Sprafka SA. *Medical Problem Solving: An Analysis of Clinical Reasoning*. Cambridge, Mass: Harvard University Press 1978

[212] Anderson J. Methodologies for studying human knowledge. Behavior and Brain Science. 1978;10:467-505

[213] Kuipers B, Kassirer J. Knowledge acquisition by analysis of verbatim protocols. In: Kidd A, ed. *Knowledge Acquisition for Expert Systems: A Practical Handbook*. New York, NY: Plenum Press; 1987.

[214] Field A. Logistic Regression. Discovering Statistics Using SPSS. London, Thousand Oaks, New Delhi: SAGE Publictions Ltd; 2005. p. 218-268.

[215] Doherty, M.E., Mynatt, C.R., Tweney, R.D., Schiavo, M.D. Pseudo-diagnosticity. Acta Psychology., 1979;53:111-121

[216] Kern, L., Doherty, M. Pseudo-diagnosticity in an Idealized Medical Problem Solving Environment. J. Med Educ., 1982,57:100-104

[217] Wolf, F.M., Gruppen, L.D., Billi, J.E. Differential Diagnosis and the Competing-Hypotheses Heuristic: A Practical Approach to Judgment Under Uncertainty and Bayesian Probability. JAMA., 1985, 253:2858-2862

[218] Hogart, R.M. *Educating Intuition*. Chicago, Illinois: The University of Chicago Press, Ltd. 2001

[219] Schooler, J.W., Dougal, S. The Symbiosis of Subjective and Experimental Approaches to Intuition. Journal of Consciousness Studies, 1999,6:280-287.

[220] Schooler, J.W., Engster-Schooler, T.Y. Verbal Overshadowing of Visual Memories: Some Things Are Better Left Unsaid. Cognitive Psychology, 1990, 22:36-71

[221] Schooler, J.W., Ohlsson, S., Brooks, K. Thoughts Beyond Words: When Language Overshadows Insight. Journal of Experimental Psychology: General. 1993,122:166-183

[222] Campbell, S.G., Croskerry, P, Bond, W.F. Profiles in Patient Safety: A "Perfect Storm" in the Emergency Department. Academic Emergency Medicine 2007,14:743-749