HYPOTHESIS-DRIVEN STORY BUILDING: COUNTERACTING HUMAN COGNITIVE BIASES TO IMPROVE MEDICAL DIAGNOSIS SUPPORT

A Dissertation in
Information Sciences and Technology

by

Shizhuo Zhu

© 2010 Shizhuo Zhu

Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

December 2010
The dissertation of Shizhuo Zhu was reviewed and approved* by the following:

John Yen  
University Professor of Information Sciences and Technology  
Dissertation Co-Advisor  
Co-Chair of Committee

Madhu C. Reddy  
Associate Professor of Information Sciences and Technology  
Dissertation Co-Advisor  
Co-Chair of Committee

Michael McNeese  
Professor of Information Sciences and Technology

Christopher N. Sciamanna  
Professor of Medicine and Public Health Sciences

Fred Fonseca  
Associate Professor of Information Sciences and Technology  
Interim Associate Dean for Education

*Signatures are on file in the Graduate School
ABSTRACT

Clinical decision-making is challenging mainly because of two factors: (1) patient conditions are often complicated with partial and changing information; (2) people have cognitive biases in their decision-making and information-seeking. Consequentially, misdiagnoses and ineffective use of resources may happen.

To better support clinical decision-making, a framework named Hypothesis-Driven Story Building (HDSB) was proposed to address the information challenges during clinical diagnosis. When given partial information, HDSB generates a list of hypotheses that can explain the current information and rank the hypotheses based on their likelihoods. If more information is needed, HDSB recommends a list of possible actions for information seeking, ranked in their effectiveness in differentiating the potential hypotheses. Whenever new information arrives, the HDSB framework updates the potential hypotheses accordingly.

The HDSB framework was built based on Multi-Layer Bayesian Network (MLBN), which is an extension of standard Bayesian network with relational representation on each node, enabling the probabilistic causal inferences for relations based on variable bindings. In MLBN, different conditional probability tables can be defined for different variable bindings so that the Bayesian inferences can be specialized or personalized and abductive reasoning can be conducted.

A web-based clinical diagnostic decision support prototype SRCAST-Diagnosis was developed based on the HDSB framework. In a given scenario, SRCAST-Diagnosis will display patient conditions, recommend differential diagnoses, and rank lab tests to
the user. It was evaluated through a controlled experiment conducted at Hershey Medical Center. Participants including nurses, residents, and physicians were divided into a control group and an experimental group. Their actions (ordering lab tests and making diagnoses) were recorded.

The result showed that SRCAST-Diagnosis can significantly improve the diagnosis accuracy and reduce the cost of resources overall, although the performances for different role players may vary. The data also showed that the tool helped more decision-makers who made the wrong initial diagnosis eventually find the correct diagnosis (counteracting anchoring heuristics) and helped them quickly figure out the correct diagnosis with significantly less resource cost (counteracting confirmation biases). It can be concluded that misdiagnosis and ineffective use of resources are associated with human cognitive biases, and a well-designed decision support system is able to improve diagnosis accuracy and resource efficiency by counteracting the cognitive biases.
TABLE OF CONTENTS

LIST OF FIGURES .................................................................................................................. viii

LIST OF TABLES ................................................................................................................... x

ACKNOWLEDGEMENTS .................................................................................................... xi

Chapter 1 INTRODUCTION ................................................................................................. 1
  1.1 The Motivation .............................................................................................................. 3
    1.1.1 Complex Patient Conditions .............................................................................. 4
    1.1.2 Human Cognitive Biases .................................................................................. 5
    1.1.3 Research Gaps ................................................................................................ 6
  1.2 Research Questions .................................................................................................... 8
  1.3 Research Approach ................................................................................................... 9
  1.4 Dissertation Roadmap ............................................................................................. 11

Chapter 2 BACKGROUND .................................................................................................. 13
  2.1 Human Errors and Cognitive Biases ....................................................................... 13
  2.2 Medical Decision-Making and Clinical Decision Support Systems .................. 19
    2.2.1 Medical Errors and Misdiagnosis ................................................................. 19
    2.2.2 Clinical Decision Support Systems ............................................................. 22
      2.2.2.1 Roles and Purposes .............................................................................. 22
      2.2.2.2 Clinical Decision Support Systems for Diagnosis: Examples ............ 24
  2.3 Recognition-Primed Decision-making and Story Building .................................. 29
  2.4 Related Techniques ................................................................................................. 32
    2.4.1 Abductive Reasoning ..................................................................................... 33
    2.4.2 Bayesian Networks ....................................................................................... 35
    2.4.3 Combining Logic with Probability ................................................................. 37
    2.4.4 Value of Information from Decision Theory .............................................. 40
    2.4.5 Information Theory ...................................................................................... 41
      2.4.5.1 Information Entropy ............................................................................ 42
      2.4.5.2 Information Gain ................................................................................ 43
  2.5 Summary .................................................................................................................... 44

Chapter 3 MULTI-LAYER BAYESIAN NETWORK: MODELING PROBABILISTIC CAUSALITY WITH VARIABLE BINDING ................. 46
  3.1 Introduction ................................................................................................................. 46
  3.2 Multi-Layer Bayesian Network: Definitions ............................................................ 50
  3.3 Inferences with MLBN ............................................................................................. 60
  3.4 Summary .................................................................................................................... 67
Chapter 4  HYPOTHESIS-DRIVEN STORY BUILDING.........................................68

4.1 The Framework ..................................................................................................68
4.2 The Formal Definitions of HDSB .........................................................................71
4.3 The Methods .........................................................................................................81
  4.3.1 Hypothesis Generation ........................................................................81
  4.3.2 Hypothesis Evaluation ........................................................................84
  4.3.3 Hypothesis-Driven Information Seeking ........................................88
  4.3.4 Hypothesis Revision ........................................................................94

Chapter 5  SYSTEM IMPELMENTATION ...............................................................99

5.1 System Architecture ..........................................................................................99
5.2 Knowledge Repositories ..................................................................................104
5.3 SR-CAST HDSB Interface ...........................................................................107
5.4 SRCAST-Diagnosis: A Web-based Clinical Diagnostic Decision Support Tool..................................................................................................................111

Chapter 6  EXPERIMENT DESIGN...........................................................................114

6.1 Experiment Objectives ......................................................................................114
6.2 Experiment Hypotheses ..................................................................................115
6.3 Experiment Design ..........................................................................................116
6.4 Experiment Process .........................................................................................116
  6.4.1 Knowledge Acquisition ........................................................................120
  6.4.2 Scenario Design ....................................................................................127
  6.4.3 Participant Recruitment ........................................................................130
  6.4.4 Data Collection .....................................................................................131
6.5 Summary...........................................................................................................134

Chapter 7  DATA ANALYSIS AND RESULTS ........................................................136

7.1 Diagnosis Accuracy .........................................................................................137
  7.1.1 Participant’s Perspective of Diagnosis Accuracy ....................................139
  7.1.2 Scenarios’ Perspective on Diagnosis Accuracy ....................................143
  7.1.3 Summary ...............................................................................................148
7.2 Cost Efficiency ................................................................................................149
  7.2.1 Analyses on Resource Cost .................................................................149
  7.2.2 Summary ...............................................................................................156
7.3 Counteracting Cognitive Biases ........................................................................157
  7.3.1 Effect of Anchoring Heuristics on Diagnosis Accuracy .......................158
  7.3.2 Effect of Confirmation Biases on Cost Efficiency ...............................161
7.4 Summary...........................................................................................................168

Chapter 8  DISCUSSION ............................................................................................170
8.1 Implications of Experiment Results.................................................................170
  8.1.1 SRCAST-Diagnosis and Cognitive Biases................................................170
  8.1.2 Other Types of Cognitive Biases in Experiment........................................173
    8.1.2.1 Premature Closure ..............................................................................174
    8.1.2.2 Availability Heuristics ......................................................................175
    8.1.2.3 Overchoice .........................................................................................176
8.2 MLBN: Representing Probabilistic Causal Knowledge with Influencing Factors ........................................................................................................177
8.3 HDSB: Modeling Human Decision-Making with Partial Information Arriving Over Time..................................................................................................178
8.4 Potential Use of SRCAST-Diagnosis............................................................182
8.5 Study Limitations..........................................................................................183
8.6 Summary......................................................................................................186

Chapter 9 CONCLUSION .....................................................................................187
  9.1 Answering Research Questions ......................................................................189
  9.2 Contributions ...............................................................................................192
    9.2.1 Contributions to the AI Community.....................................................192
    9.2.2 Contributions to the Medical Informatics Community ........................195
  9.3 Future Work................................................................................................197
    9.3.1 Large-scale Experiment.........................................................................197
    9.3.2 Extension to Collaborative Settings ....................................................197
    9.3.3 Implementation on Mobile Devices .....................................................198
    9.3.4 HDSB with Data Mining ......................................................................198
    9.3.5 HDSB with Natural Language Processing ..........................................199
  9.4 Closing Remarks..........................................................................................199

Bibliography .........................................................................................................201

Appendix A Diseases and Related Content ..........................................................207
  A.1 Viral Pneumonia (ICD-9: 480) .....................................................................207
  A.2 Pulmonary Embolism (ICD-9: 415.1) .............................................................208
  A.3 Cardiogenic Pulmonary Edema (ICD-9: 514) .............................................211
  A.4 Myocardial Infarction (ICD-9: acute 410) ....................................................214
  A.5 Status asthmaticus (ICD-9: 493.01, 493.02) ..............................................215
  A.6 Acute Bronchitis (ICD-9: 466) .....................................................................217
  A.7 Tension Pneumothorax (ICD-9: 512.0) ..........................................................219
  A.8 Dissecting Thoracic Aneurysm (ICD-9: 441.0) .........................................220

Appendix B Conditional Probability Tables ........................................................223
  B.1 Conditional Probabilities for Symptoms....................................................223
  B.2 Conditional probabilities on Physical Examinations ................................224
  B.3 Conditional probabilities of Lab Tests ......................................................225
LIST OF FIGURES

Figure 2-1: Recognition-Primed Decision (RPD) Model ............................................... 31
Figure 3-1: A Classical Bayesian Network.................................................................... 51
Figure 3-2: A revised Bayesian network that differentiates causes and constraints.... 52
Figure 3-3: A Sample Multi-Layer Bayesian Network................................................. 59
Figure 4-1: Hypothesis-Driven Story Building Framework ........................................ 70
Figure 4-2: Observations in a MLBN ......................................................................... 72
Figure 4-3: Observation Space .................................................................................. 73
Figure 4-4: Explanandum Space and Evidence Space................................................ 76
Figure 4-5: Explanans Space .................................................................................... 77
Figure 4-6: Explanation .............................................................................................. 78
Figure 4-7: Relations of the Components of a Story .................................................... 79
Figure 4-8: Story ........................................................................................................ 79
Figure 4-9: A MLBN with Medical Conditions ........................................................... 80
Figure 4-10: Binary Entropy Plot ............................................................................... 86
Figure 4-11: RI(H) Plot .............................................................................................. 87
Figure 4-12: Plot of Revised RI(H) .............................................................................. 88
Figure 4-13: Hypothesis Removed by Node Changed from Unknown to Normal ...... 96
Figure 4-14: Hypothesis Added by Node Changed from Unknown to Abnormal ...... 96
Figure 4-15: Hypothesis Added by Node Changed from Normal to Abnormal ......... 97
Figure 5-1: SR-CAST Architecture ............................................................................ 100
Figure 5-2: Syntax of Causal Relationship Definition............................................... 105
Figure 5-3: A Multi-Layer Definition of Causal Relationship .................................... 106
Figure 5-4: SR-CAST HDSB Interface ................................................................. 107
Figure 5-5: Editing Window for a Predicate Node ............................................. 108
Figure 5-6: Editing Window for Conditional Probabilities ............................... 108
Figure 5-7: SRCAST-Diagnosis Architecture .................................................... 111
Figure 5-8: SRCAST-Diagnosis Interface ............................................................ 112
Figure 6-1: Interface for Control Group Users ................................................... 118
Figure 7-1: (a) Assigned Group; (b) Gender; (c) Years of medical experience; (d) Position in Emergency Department ................................................................. 137
Figure 7-2: Histogram of Exact Accuracy in Control Group and Experimental Group ............................................................................................................. 139
Figure 7-3: Check of Normal Distribution .......................................................... 140
Figure 7-4: Histogram of Resource Cost in Control Group and Experimental Group ............................................................................................................. 150
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Examples of CDSS Interventions by Target Area of Care (Berner, 2009)</td>
<td>23</td>
</tr>
<tr>
<td>2-2</td>
<td>Clinical Decision Support Systems</td>
<td>28</td>
</tr>
<tr>
<td>4-1</td>
<td>Revision Strategy for Condition Changes</td>
<td>98</td>
</tr>
<tr>
<td>6-1</td>
<td>Price and Time of Lab Tests</td>
<td>127</td>
</tr>
<tr>
<td>7-1</td>
<td>Two-sample Student’s T-test for Exact Accuracy (All Participants)</td>
<td>141</td>
</tr>
<tr>
<td>7-2</td>
<td>Two-sample Student’s T-test for Exact Accuracy (Nurses)</td>
<td>141</td>
</tr>
<tr>
<td>7-3</td>
<td>Two-sample Student’s T-test for Exact Accuracy (Residents)</td>
<td>141</td>
</tr>
<tr>
<td>7-4</td>
<td>Two-sample Student’s T-test for Exact Accuracy (Physicians)</td>
<td>142</td>
</tr>
<tr>
<td>7-5</td>
<td>The Diagnosis Accuracy on Scenarios (All Participants)</td>
<td>144</td>
</tr>
<tr>
<td>7-6</td>
<td>P-Values Based on Scenarios (All Participants)</td>
<td>146</td>
</tr>
<tr>
<td>7-7</td>
<td>P-Values Based on Scenarios (by Roles)</td>
<td>147</td>
</tr>
<tr>
<td>7-8</td>
<td>P-Value of Resource Cost</td>
<td>151</td>
</tr>
<tr>
<td>7-9</td>
<td>Cost Analysis on Scenarios (All Participants)</td>
<td>152</td>
</tr>
<tr>
<td>7-10</td>
<td>Means of Resource Cost</td>
<td>153</td>
</tr>
<tr>
<td>7-11</td>
<td>P-Values for Resource Cost</td>
<td>154</td>
</tr>
<tr>
<td>7-12</td>
<td>Number of Labs</td>
<td>155</td>
</tr>
<tr>
<td>7-13</td>
<td>Effect of Anchoring Heuristics on Diagnosis Accuracy</td>
<td>159</td>
</tr>
<tr>
<td>7-14</td>
<td>Resources Used When Initial Diagnosis Revised (All Participants)</td>
<td>162</td>
</tr>
<tr>
<td>7-15</td>
<td>Resources Used When Initial Diagnosis Revised (Nurses)</td>
<td>163</td>
</tr>
<tr>
<td>7-16</td>
<td>Resources Used When Initial Diagnosis Revised (Residents)</td>
<td>164</td>
</tr>
<tr>
<td>7-17</td>
<td>Resources Used When Initial Diagnosis Revised (Physicians)</td>
<td>165</td>
</tr>
<tr>
<td>7-18</td>
<td>P-values for Hypothesis Switch</td>
<td>166</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

I would like to first thank my parent, Pengnian Zhu and Qiuzhen Ye, for their unconditional love and support during this long journey. They were my first teachers who educated me with their love, bravery, diligence, and dedication. Their encouragement have accompanied me during this lonely and sometimes struggling journey, and will continue be with me along the journey of my entire life.

I am grateful to my advisors, Dr. John Yen and Dr. Madhu Reddy, for their excellent guidance over the past years. Their joint advising has led me to an area that I had never imagined and a stage that I had never reached. Dr. Yen encouraged my independent thinking and attitude of self-challenging. Dr. Reddy demonstrated me how to articulate my work. The accomplishment of this dissertation was not possible without their supports and advises. I would also like to thank Dr. Michael McNeese and Dr. Chris Sciamanna for serving on my thesis committee and for their insightful comments.

I am also grateful to Dr. Christopher DeFlitch for his significant help during the entire process of system implementation and experiment conducting. He thoroughly reviewed the knowledge base and verified the scenarios of SRCAST-Diagnosis and coordinated the scheduling of experiment. I should also thank Dr. Uchenna Ofoma and Dr. Steward McCarver for their input of the related medical knowledge. I also thank Darrell Walter, David Snell, Nancy Flint, Laura Martin and Lisa Haas for their helps in conducting knowledge engineering and experiment.

I am also thankful to my colleagues and friends, Xiaocong Fan, Haizheng Zhang, Shuang Sun, Cong Chen, Rui Wang, Kaivan Kamali, Bingjun Sun, Po-Chun Chen,
Soyoung Oh, Baojun Qiu, Kang Zhao, Hyun-Woo Kim, Qi Fang, Eun-Yeong Ahn, Mu Qiao, Hao Jiang, Sharoda Paul, Joanna Abraham, and Honglu Du, who offered useful inputs and suggestions to my dissertation.

My special thanks go to my girlfriend, Lu Fan, who gave me the bright sunshine during my last time of completing the dissertation. Thank you, Mianmian.
Chapter 1

INTRODUCTION

“…we’ve estimated that most of this plan can be paid for by finding savings within the existing health care system, a system that is currently full of waste and abuse. Right now, too much of the hard-earned savings and tax dollars we spend on health care don’t make us any healthier.”

- President Barack Obama

remarks to a Joint Session of Congress on health care

September 9, 2009

The health care system in the United States has long been a controversial issue, and its high cost is one of the major reasons. In 2009, it was estimated that the spending on healthcare in the U. S. was approximately 16.2% of its GDP, or $7,681 per capita (CMS, 2008). A recent study (Dalen, 2009) found that medical expenditures were a significant factor that contributed to 62% of the country’s personal bankruptcies in 2007. The proposed health care reform, therefore, aims to offer quality and affordable health care to all Americans. Significant milestones have been made with two bills—the Patient Protection and Affordable Care Act and the Health Care and Education Reconciliation Act—passed in 2010 by Congress and signed by President Barack Obama.

In the overall cost of health care, a great portion is related to medical liability and defensive medicine. According to 2006 analysis conducted by Price Waterhouse Coopers, this cost is about 10% of the overall health care cost. Medical liability and defensive medicine are rooted in medical malpractices, which are the failures of health care providers to meet the professional standards by which patient injuries or deaths are
caused (Danzon, 1985). Many factors can cause malpractices. Among them, misdiagnosis is one of the major causes. Studies have disclosed that one third of all malpractice claims concerned misdiagnosis (Phillips et al., 2004), making it one of the fundamental reasons for defensive medicine and its related cost.

Defensive medicine involves the practice of doctors’ diagnostic or therapeutic measures primarily (but not necessarily or solely) to reduce their exposure to malpractice liability, instead of ensuring the health of patients (Congress, 1994). One typical action that a physician may engage in to reduce malpractice liability is to order extra tests or procedures, which is considered positive defensive medicine. The fear of malpractice has lead to significant changes in emergency department decision-making and has been associated with the increased hospitalization of low-risk patients as well as the increased use of diagnostic tests (Katz et al., 2005). In another study, it was reported that “nearly all” (93%) of the physicians who participated in the study had practiced defensive medicine (Studdert et al., 2005). For instance, 43% participants reported using image technology in clinically unnecessary circumstances, as an “assurance behavior” (Studdert et al., 2005; Manner, 2007). The practice of ordering unnecessary tests in order to reduce the risk of malpractice litigation is known as “defensive testing” (DeKay & Asch, 1998).

There is a strong link between misdiagnosis and the high cost of health care. The misdiagnosis-cost relationship has been the dilemma: if health care providers are constrained to reduce the number of tests to control the cost, they may potentially be exposed to the risk of increasing the misdiagnosis rate, which may eventually cost them the medical liability awards; if health care providers are encouraged to practice defensive medicine and conduct as many lab tests as they think are clinically assured (but more
than are necessary), the entire society and ultimately each individual must share the cost in the long run, even though a specific patient or physician may receive a short-term benefit.

In his 2009 speech to the Joint Session of Congress, President Barack Obama proposed to “put patient safety first and let doctors focus on practicing medicine” while at the same time bringing down the cost of health care. His passion and ambition motivated me to explore the possibility of breaking the misdiagnosis-cost dilemma.

1.1 The Motivation

Misdiagnosis and its consequences, including medical malpractice litigation and defensive medicine, have contributed a substantial portion to the high cost of health care in the United States. Therefore, it is critical to understand the causes of misdiagnosis before potential methods can be explored for solving the misdiagnosis-cost dilemma.

A misdiagnosis is an error in diagnosis. It may be in various formats, such as a wrong diagnosis, a missed diagnosis, a delayed diagnosis, or missed complications. A misdiagnosis is the erroneous result of a physician’s cognitive assessment of a patient’s condition. Therefore, it is the common effect of complicated patient conditions (externally) and physicians’ cognitive constraints (internally).
1.1.1 Complex Patient Conditions

A misdiagnosis is most likely to occur with patients having complex situations, especially in those whose symptoms can be explained by more than one disease, and whose conditions can change over time. A typical patient diagnosis process usually faces three challenges:

1. Patient conditions are only partially known.

Patient conditions are only partially known most of the time during the diagnosis process. A diagnosis is an effort of assessing the true patient condition and the actual cause of the patient’s current situation based on limited information such as symptoms, vital signs, physical examination results, and lab results. With the accumulation of information, the assessment will become more and more accurate, but it will still be a best guess rather than the truth. Thus, incomplete information is the first challenge to making a correct diagnosis.

2. Patient conditions can be obtained only if they are requested.

Some patient conditions or symptoms can be observed by physicians or nurses; however, other conditions can be obtained only if they are requested, such as lab test results, and will remain unknown until the lab test is explicitly ordered. Resources for lab tests are oftentimes costly or tightly scheduled. Ordering redundant lab tests is not only economically inefficient for the patient but also may deprive other patients with urgent needs of the opportunities to use the lab test resources. Therefore, the second challenge for a physician is to order the right amount of information for making a correct diagnosis.

3. Patient conditions can change.
The conditions of a patient may be unstable during the diagnostic process. A symptom may appear or disappear. A vital sign may have a significant jump or drop. Some of these changes may have a significant impact on a suspected diagnosis while others may not. Whenever a condition changes, it must be integrated into the current differential diagnosis. The significance level of the change and its consequence must be determined: should the list of differential diagnoses be changed, should only their ranking be changed, or should no change be made? Thus the third challenge for a physician is to adjust his differential diagnosis based on new information in a timely manner.

These challenges highlight some of the difficulties in making a correct medical diagnosis. All three are related to information use and seeking. However, human cognitive biases make a medical diagnosis process more error-prone.

1.1.2 Human Cognitive Biases

Human decision-makers’ activities when dealing with these three challenges are affected by their cognitive biases. A cognitive bias is a person’s tendency to make errors in judgment based on cognitive factors, a phenomenon studied in cognitive science and social psychology (Gilovich, Griffin, & Kahneman, 2002). Researchers have found that human cognitive biases will prevent individuals from making better decision choices (Hogarth, 1987; Reason, 1990). Within the healthcare domain, studies have shown the effects of human cognitive biases on clinical quality (Bornstein & Emler, 2001; Milstein & Adler, 2003; Redelmeier, 2005).
Among the various types of cognitive biases that affect decision-makers’ judgment, some are related to information using and seeking, which make the (above) three challenges more difficult. For example, during a diagnosis, a physician may over-trust or over emphasize one particular piece of information based on his past experiences. Consequently, he may stick to an initial suspected diagnosis made based on this previous information and overlook other possibilities (Tversky & Kahneman, 1974). Furthermore, when the decision-maker needs additional information to verify the diagnosis, he may again selectively pay more attention to information that will further support the favorable diagnosis than to other information (Plous, 1993).

Human cognitive biases are the systematic cognitive tendencies of human beings caused by the constraints of human brains, not errors made due to individual mistakes (Gilovich & Griffin, 2002). Therefore, the reward-penalty mechanism (which led to the practice of defensive medicine) may not work to reduce rate of misdiagnosis attributed to human cognitive biases, but may significantly increase the cost of health care.

1.1.3 Research Gaps

Many efforts have been made to reduce misdiagnosis rates and to control costs. However, most of these have focused only on one element instead of taking them into account jointly. For example, diagnostic decision support systems were designed primarily to reduce the misdiagnosis rate by offering diagnostic recommendations, but they did little to address the problems of defensive medicine and cost control. On the other hand, Fries et al. (1993) suggested that measurements on preventing disease,
reducing risky behaviors, improving self-management, and applying healthcare promotion programs at work could be made for analyzing potential cost reductions. These suggestions would be useful to reduce the need for medical services before a patient becomes ill, but none of them address the issue of cost control during the patient diagnosis process.

In summary, there are two important gaps in the current research on reducing misdiagnosis and controlling cost:

1. **The factor of cognitive biases is missing from the design of diagnostic decision support systems.**

   Although studies have shown the relationship between human cognitive biases and misdiagnosis (Bornstein & Emler, 2001; Milstein & Adler, 2003; Redelmeier, 2005), few efforts have addressed human cognitive biases in clinical diagnostic decision support systems as a fundamental measurement of reducing misdiagnosis. So far, efforts related to diagnostic decision support systems have focused on the design of algorithms, the representation of knowledge, incorporating more types of diseases, improving user interfaces, and offering links to searchable databases. Each of these activities will, no doubt, improve the accuracy of recommended diagnosis; however, none of the decision support systems have shown evidence that their improvements in diagnostic accuracy are associated with any counteraction of cognitive biases.

2. **Cost control is rarely considered when a diagnostic decision support system recommends possible diagnosis.**

   Many types of clinical decision support systems have been developed for different purposes and applied in different areas. Some focus on suggesting possible diagnoses that
match a patient’s signs and symptoms. Others focus on cost reduction. However, cost control exists in relatively simple format: it is achieved by monitoring medication orders, and by avoiding duplicate and unnecessary tests (Perreault & Metzger, 1999). Currently, no list of possible lab tests is given, ranked in importance level to differentiate the differential diagnoses. Requesting labs in the order of their importance level could be expected to improve the overall efficiency of lab resources.

The research gaps set the goal of my research: to break the misdiagnosis-cost dilemma (i.e., to reduce health care cost by increasing the effective use of lab resources) and, at the same time, to improve or retain diagnosis accuracy. This goal shaped my research questions.

### 1.2 Research Questions

To address the research gaps, I raised the following four questions which will be investigated in the remainder of this dissertation.

1. **What model can be developed to describe physicians’ hypothesis reasoning and information seeking process in clinical diagnosis?**

2. **Can a decision support tool, built based on the proposed model, reduce misdiagnosis?**

3. **Can the same tool reduce the cost of lab tests while reducing misdiagnosis?**

4. **Are these improvements caused by the counteraction of cognitive biases?**
The first question was aimed at setting up the requirements for a general framework to model the process by which a physician (or possibly a computer-based decision support system) makes diagnostic decisions, especially when facing incomplete and changing patient conditions. The framework must address: how the decision-maker generates differential diagnoses when information is only partially known; how the decision-maker identifies and collects missing information; and how the decision-maker revises previously made tentative decisions when given new information.

Questions 2 and 3 were aimed at evaluating the usefulness of the decision support tool in reducing both misdiagnosis and cost. The tool needs to show its capability for reducing both misdiagnosis and cost at the same time, demonstrating its potential to break the misdiagnosis-cost dilemma.

Question 4 was aimed at understanding how these improvements were possible. It must show whether the effects of certain human cognitive biases would be counteracted by the tool so that a decision-maker’s behaviors and choices would potentially be changed.

The next section will outline the approaches taken to answer the research questions.

1.3 Research Approach

To answer the first question, I developed the Hypothesis-Driven Story Building (HDSB) framework which describes, in a general perspective, a decision-maker’s activities while moving toward his final decision. Diagnosis is modeled as a story
building activity, which aims to find an explanation for the observed information. A story building process begins by generating hypotheses, then proceeds to seeking information and revising hypotheses, and ends by making a final decision. The HDSB framework was built on the Multi-Layer Bayesian Network (MLBN), an extension of the classical Bayesian network, to enable probabilistic inferences with predicate sentences.

To answer the second and third questions, a controlled experiment was designed and conducted. Participants with certain levels of medical decision-making experience were recruited for the experiment. Participants were divided into an experimental group and a control group. Those in the experimental group were backed up by the decision support tool while making their decisions, while those in the control group made decisions based on their own judgment. Each participant’s decisions about final diagnosis and lab tests were recorded. An analysis on the diagnosis accuracy and cost efficiency of each group was conducted.

To answer the fourth question, I tracked the intermediate behaviors of each participant’s decision-making process during the experiment. Whenever a decision-maker wanted to order lab tests, he was asked to select the diagnosis that he most suspected at that time. This allowed a comparison of the accuracy of the initial diagnosis, both with and without the decision support tool, and also tracked the time at which a participant switched to a correct diagnosis from an initially incorrect one. These data reflected the cognitive tendency in decision-making, especially the anchoring heuristics and confirmation biases. Thus, by tracking the intermediate behaviors and choices, an analysis of these two cognitive biases can be conducted.
This approach is composed of stages of design, implementation, experimentation, and analysis; therefore, it is a holistic approach. Through it, not only are the research questions answered, but also insights for future development are generated.

1.4 Dissertation Roadmap

The remainder of this dissertation is structured as follows:

Chapter 2 introduces related studies on cognitive biases, clinical decision-making, diagnostic decision support systems, and technologies for developing the decision support tool (abductive reasoning, Bayesian networks, value of information in decision theory, and entropy of information.)

Chapter 3 introduces the Multi-Layer Bayesian Network, which is fundamental to the HDSB framework.

Chapter 4 presents the Hypothesis-Driven Story Building framework, including the methodology for each specific step: hypothesis generation, hypothesis evaluation, hypothesis-driven information seeking, and hypothesis revision.

Chapter 5 describes the implementation of the HDSB-enabled decision support tool, built on the R-CAST platform but using a new MLBN component, and incorporating an HDSB module. It also presents an agent interface and a web-based application interface.

Chapter 6 discusses the experiment design used to evaluate the tool.
Chapter 7 provides the analysis of the experiment data. Diagnosis accuracy and cost efficiency are analyzed, and the relationship between the performance improvement and the counteraction of cognitive biases are discussed.

Chapter 8 offers a discussion of remaining issues, including cognitive biases, other potential use of the tool, and the limitations of the study.

Chapter 9 summarizes the contributions of my work and its possible future extension.
Chapter 2
BACKGROUND

This chapter provides an overview of studies related to this work. The first section is a review of the literature on human errors and cognitive biases, focusing on specific types of cognitive biases and their influence on decision-making quality in clinical diagnosis. The second section is a review of medical diagnostic decision-making and clinical decision support systems, with a focus on how a computerized decision support system can assist with making diagnostic decisions. The third section is the review of the Recognition-Primed Decision (RPD) model and the key concept: Story Building. The fourth section is a review of techniques related to or applied in this work, including abductive reasoning, Bayesian networks, value of information in decision theory, and information entropy.

2.1 Human Errors and Cognitive Biases

People make mistakes in decision-making and problem solving. Some of these mistakes and failures are caused by human cognitive biases (Tversky & Kahneman, 1974), especially when people are faced with complicated problems in complex situations. There are various types of cognitive biases. Some are decision-making and individual behavior-related biases, such as anchoring heuristics (Tversky & Kahneman, 1974) and confirmation biases (Plous, 1993); some are related to the estimation of probabilities and
beliefs, such as availability heuristics (Tversky & Kahneman, 1974) and belief bias (Evans, Barston, & Pollard, 1983); and some are related to one’s social activities, such as the bystander effect (Darley & Latane, 1968) and the false consensus effect (Ross, Greene, & House, 1977).

Although they may be used interchangeably from time to time, *heuristics* and *biases* refer to two phenomena closely associated but with slight differences. A heuristic is a mental shortcut that provides subjectively compelling and oftentimes serviceable solutions to decision problems (Gilovich & Griffin, 2002), while a bias is oftentimes used to refer to a deviation from the normative rational theory that serves as a marker and signature of the underlying heuristics (Gilovich & Griffin, 2002). Therefore, a heuristic is the cause of a bias, and a bias is the consequence of a heuristic.

A heuristic comes into play if a decision problem becomes complicated and sophisticated. Heuristics and biases were first brought to people’s attention by Amos Tversky and Daniel Kahneman (1974) who realized the relationships between uncertainty and heuristics. A decision oftentimes needs to be made based on uncertain information, where the uncertainty is usually represented in probabilities. However, not all probabilities can be obtained objectively and reliably, which is why subjective probabilities are used and where a heuristics can play a role. A heuristic reduces the complexity of a decision task to simpler judgment operations (Tversky & Kahneman, 1974) by bypassing the complicated and sophisticated computations. For example, a heuristic may occur when a starting point (anchor) is given to the subject, or the subject bases his first estimate on some incomplete and initial computation (anchoring heuristics) (Tversky & Kahneman, 1974). Also, a heuristic may occur in situations where the
estimate of possibilities is affected by the available information (availability heuristics) (Tversky & Kahneman, 1974). For instance, a physician may come up with an initial diagnosis for a patient based on his memory of similar cases from past experiences.

Heuristics are often used as a replacement for complicated and sophisticated computations based on the assumption of rational choice theory: decision-makers base their decisions on calculations of utilities using costs, benefits, and probabilities in order to make rational decisions that maximize the expected utility (Scott, 2000). The particular type of decision-making model built on the rational choice theory is called a descriptive decision-making model. It assumes that a decision-maker has clearly identified goals, complete information about the situation, exhaustive alternatives, and accurate predictions of their consequences (Chaffee, 1981; Browne, 1993; Goodwin & Wright, 2004). These assumptions were considered too strong and inconsistent with the actual way in which humans make decisions. For example, the capabilities of human brains may be limited in identifying problems, processing information, generating alternatives, evaluating alternatives, and implementing actions (March, 1994; Gigerenzer & Selten, 2002). Therefore, the heuristics that play a role in making up the gap between ideal rational choice and human cognitive limitations are no longer merely simple replacements of rational choice, but categorize the decision-making into a totally different type (Gilovich & Griffin, 2002). This type of decision-making is known as descriptive decision-making, which includes bounded rationality decision-making (Gigerenzer & Selten, 2002), behavioral decision-making (Browne, 1993), and naturalistic decision-making (Zsambok & Klein, 1997), each with different focus.
Descriptive decision-making aims to describe what actually happens during the human decision-making processes rather than prescribe what someone thinks should happen (Bell, Tversky, & Raiffa, 1988; Browne, 1993). Since prescriptive decision-making usually requires the environment and the problem to be well-defined, it may be constrained in solving real problems which are often ill-defined. Descriptive decision-making is therefore the effort to investigate how human decision-makers actually make decisions under non-ideal environments for ill-defined problems, where conditions may change over time, information may be ambiguous, goals may shift, and time may be stressed for the decision-making (Zsambok, Beach, & Klein, 1992; Zsambok & Klein, 1997). For example, as described in RPD (Klein, 1989), a decision-maker may generate a solution for a problem based on his past experiences by comparing the current situation with previous situations. This solution may not be optimal, but it may work to solve the problem. To have a solution in a timely manner is critically important to those problems with time pressure, such as fire fighting. This is where the value of human heuristics lies and why researchers are interested in investigating the actual decision-making process of human decision-makers.

As the consequence of heuristics, a bias may not necessarily be harmful. On one hand, as defined by Gilovich and Griffin (2002), a bias is a deviation of normative rational theory, and it may possibly lead to a better result than the rational result calculated from rational choice theory, since the rational result is oftentimes based on expected utility rather than the possible maximum utility. On the other hand, a less-optimal solution may not substantially impair the quality of the problem solving, but may significantly improve other performances, such as resource efficiency and time efficiency.
In many situations, the resource efficiency and time efficiency may also be the important factors that significantly affect the overall performance.

Despite their positive effects, the negative effects of cognitive biases are not ignorable. The remainder of this section reviews some of the biases and their negative effects related to decision-making and information seeking.

1. Out of sight, out of mind: This bias suggests that people may ignore the facts if related elements are hidden from them, but which would have been apparent and easy to understand if presented to them. Fischhoff et al. (1978) demonstrated the existence of this bias. In their experiment, the subjects were insensitive to the missing parts of things, even to major and commonly known but omitted components. Healthcare researchers have studied the influence of this bias on clinical quality failure (Milstein & Adler, 2003).

2. Anchoring heuristics: This bias describes the common human’s tendency to rely too heavily on one trait or one piece of information during decision-making (Tversky & Kahneman, 1974). A decision-maker tends to stick to the first hypothesis and decision made at the beginning of the process. Healthcare studies have shown that medical decision-makers also tend to rely on the first diagnosis made for a patient and therefore, misdiagnosis is likely to occur (Redelmeier, 2005).

3. Availability heuristics: This bias indicates the tendency of a decision-maker to predict an event if an instance of it can be easily brought to mind (Tversky & Kahneman, 1974). For example, a clinical diagnostician may attribute a patient’s upper back pain and myalgias to viremia without considering other
options if such an association can be easily retrieved from the memory of previous experience (Redelmeier, 2005).

4. Premature closure: This bias is regarded as the tendency to offer only a single solution to a problem that logically allows multiple solutions due to insufficient or ambiguous information (Acredolo & Horobin, 1987). It has been considered one of the major factors causing misdiagnoses in medical decision-making. For example, Redelmeier (2005) discovered that within the healthcare domain, physicians may tire of searching for alternative solutions and thus close the search process too quickly.

5. Confirmation bias: This bias concludes that a person rapidly favors one available interpretation and subsequently comes to favor information that may support his hypothesis when faced with ambiguous information (Reason, 1990; Plous, 1993). Studies have shown that preliminarily formed hypotheses based on initial and incomplete data may interfere with later interpretations of better and abundant data (Greenwald, Pratkanis, Leipoe, & Baumgardner, 1986). Bornstein and Emler (2001) presented examples of this bias in medical decision-making, including the use of mammography not to dissuade a doctor from suggesting a biopsy but rather to confirm a hypothesis of malignancy, regardless of the patient’s symptoms or even the results of a mammogram.

Human biases are the cognitive tendencies of humans due to the cognitive constraints of their brains, not errors that are due to personal intention or carelessness. Therefore, the decision-making mistakes caused by biases should not be considered faults of the individuals, but rather the lack of a better system that takes into account the human
biases and helps reduce their negative effects. This situation can be changed through the use of decision support systems which address cognitive biases. The purposes of developing such a decision support system can include but are not limited to: (1) generating and identifying the correct hypothesis as early as possible, and (2) seeking the right information for the right purpose.

The next section will discuss types of medical errors, particularly misdiagnosis in medical decision-making, and the role of clinical decision support systems in reducing misdiagnosis.

2.2 Medical Decision-Making and Clinical Decision Support Systems

The first subsection contains a review of medical errors and types of misdiagnosis, since misdiagnosis is significant cause of medical errors. The second sub section reviews efforts toward making clinical decision support systems, especially those focused on making diagnostic recommendations. By reviewing the current situation of misdiagnosis in health care and the functionality of existing decision support systems, a gap between the capabilities of the current systems and the vision of the health care reform can be seen. This is where the motivation of my work is located.

2.2.1 Medical Errors and Misdiagnosis

According to a report by the Institute of Medicine (2000), between 44,000 to 98,000 unnecessary deaths and more than one million injuries each year are caused in the
United States by medical errors. In another study, researchers reviewed the medical charts of 30,121 patients admitted to 51 acute care hospitals in New York State in 1984, and reported that injuries caused by medical management occurred to 3.7% of admitted patients, and 69% of those injuries were due to medical errors (Brennan et al., 1991; Leape, Lawthers, Brennan, & Johnson, 1993). Another group of researchers reviewed the medical records of 14,179 admissions to 28 hospitals in New South Wales and South Australia during 1995, and found that adverse events occurred in 16.6% of admissions, resulting in permanent disability to 13.7% of the patients and death to 4.9%; among these adverse events, 51% were regarded as preventable (Wilson et al., 1995).

Medical errors can prolong a patient’s hospital stay and result in increased hospital costs. One study showed that adverse drug events increased hospital costs by $2,595, and extended the length of stay by 2.2 days (Bates et al., 1997).

Within all medical errors, diagnosis errors are more likely to result in permanent disability than are technical errors, and are more likely to have been preventable (Wilson, Harrison, Gibberd, & Hamilton, 1999; Weignart, Wilson, Gibberd, & Harrison, 2000). The types of diagnosis errors include (WD, 2010):

1. Wrong diagnosis: a patient is diagnosed to have a different disease from the actual disease.

2. Missed diagnosis: a patient is diagnosed as healthy, while in fact he has a disease.

3. Delayed diagnosis: a patient is not diagnosed in a timely manner. The patient may develop severe consequences because of the delayed diagnosis.
(4) Wrong diagnosis of subtype of disease: the overall diagnosis is correct, but the diagnosis of the subtype is wrong.

(5) Missed complications: although the patient’s primary condition may be diagnosed, additional complications caused by it may have been missed. It is critical to correctly diagnose and treat the complications.

(6) Missed diagnosis of underlying disease: the original condition is diagnosed; however, the actual disease that caused the original condition is missed.

(7) Missed diagnosis of medication side effect: medications may cause side effects. The diagnosis of a condition as a medication side effect may be overlooked.

(8) Missed diagnosis of related diseases: certain diseases tend to group together. Even if a physician correctly diagnoses a condition, he may also miss the diagnosis of a related disease.

Despite these varieties of diagnosis errors, there is little evidence showing that they are due to the presence of bad people in healthcare. Rather, good people are working in imperfect systems that need to be made safer (Medicine, 2000). It has also been found that medical errors are usually associated with inexperienced physicians, with the introduction of new procedures, and during complex care and urgent care (Weignart, Wilson, Gibberd, & Harrison, 2000). A 1997 report showed a misdiagnosis rate as high as 20%-40% in ICUs and emergency rooms (Associates, 1997). This condition is due to the time-critical and serious nature of diagnosis under crisis. Additionally, as noted earlier, human cognitive biases play an important role in making medical diagnosis error-
prone (Redelmeier, 2005). Hence, clinical decision support systems have been proposed and used in medical settings as a supplemental measurement for the imperfect system.

2.2.2 Clinical Decision Support Systems

The first sub section reviews the different roles and purposes of clinical decision support systems. The second sub section introduces a list of clinical decision support systems that focus on making diagnostic recommendations, highlighting their different features.

2.2.2.1 Roles and Purposes

Decision support systems are a class of computer systems capable of assisting human decision-makers, based on certain data and models (Eom, 2003). Healthcare is one of the major areas in which these systems are currently applied. Clinical Decision Support Systems (CDSS) are a particular type of decision support system used by healthcare providers. According to Wyatt and Spiegelhalter (1991), the CDSS are active knowledge systems which generate case-specific advice using two or more items of patient data.

Perreault and Metzger (1999) identified four key CDSS functions: administrative, managing clinical complexity and details, cost control, and decision support. The target areas of CDSS are summarized in Table 2-1 based on Berner (2009), where supporting diagnosis is only one of many areas.
Table 2-1: Examples of CDSS Interventions by Target Area of Care (Berner, 2009)

<table>
<thead>
<tr>
<th>Target Area of Care</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preventive care</td>
<td>Immunization, screening, disease management guidelines for secondary prevention</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Suggestions for possible diagnoses that match a patient’s signs and symptoms</td>
</tr>
<tr>
<td>Planning or implementing treatment</td>
<td>Treatment guidelines for specific diagnoses, drug dosage recommendations, alerts for drug-drug interactions</td>
</tr>
<tr>
<td>Follow-up management</td>
<td>Corollary orders, reminders for drug adverse event monitoring</td>
</tr>
<tr>
<td>Hospital, provider efficiency</td>
<td>Care plans to minimize length of stay, order sets</td>
</tr>
<tr>
<td>Cost reductions and improved patient convenience</td>
<td>Duplicate testing alerts, drug formulary guidelines</td>
</tr>
</tbody>
</table>

The table shows that the general goals of CDSS are to improve patient safety, care quality, and efficiency in healthcare delivery (Coiera, 2003). Cost controls and making diagnostic recommendations are realized separately. Cost controls are realized by activities at the hospital administrative level, such as minimizing the length of stay for patients or by alerting physicians to avoid duplicate testing, while diagnosis support is more patient-centric (Berner, 2009).

In a narrow sense, a CDSS has three major functions related to diagnostic decision support: managing information, focusing attention, and providing patient-specific recommendations (Wyatt & Spiegelhalter, 1991). A CDSS allows medical decision-makers to browse, store, retrieve, and edit information related to decision-making, creates alerts and reminders based on its knowledge and decision-makers’ needs,
and makes diagnostic or therapeutic suggestions based on patient data (Musen, Shahar, & Shortliffe, 2000). More specifically, Coiera (2003) listed the functions of a CDSS as alerts and recommendations, diagnosis assistance, therapy critiquing and planning, prescription assistance, information retrieval, image recognition and interpretation. Some examples of CDSS that focus on making diagnostic suggestions are introduced in the next section.

2.2.2.2 Clinical Decision Support Systems for Diagnosis: Examples

This section introduces a short list of clinical decision support systems that focus on making diagnostic recommendations. Across these systems, different methods have been developed, different medical areas have been targeted, and different user interfaces have been designed. However, they serve the same major purpose: making diagnostic recommendations based on input information.

MYCIN

MYCIN was initially developed as an expert system in the mid-1970s at Stanford University. Its programming language was Lisp, and its inference derived from a knowledge base that contains more than 600 predefined judgment rules. It was initially designed to identify bacteria in blood infections. MYCIN required a user to answer a series of simple yes/no or textual questions. Then, it provided a list of possible diagnoses, ranked from high to low probability, and included a recommended course of treatment for each diagnosis.

INTERNIST-I / QMR
INTERNIST-I was a rule-based system implemented in the early 1970s at the University of Pittsburgh, and its database reportedly covered 70-80% of all possible diagnoses in internal medicine by 1982. It used partitioning algorithms to create problem areas and exclusion functions to eliminate possible diagnoses, based on the hierarchical decision-tree logic, which allows a disease to belong to only one disease class. QMR (Quick Medical Reference) was its successor and aimed at correcting the technical and philosophical deficiencies of INTERNIST-I, although the major algorithms remained unchanged.

**CADUCEUS**

CADUCEUS was an expert system implemented in mid-1980s at the University of Pittsburgh, a successor to INTERNIST-I. It extended the reference domain to all internal medicine and was eventually able to diagnose up to 1000 differential diseases.

**DXplain**

DXplain, one of the first commercialized systems, was able to generate stratified diagnoses based on patient symptoms, lab results, and other clinical findings. Along with the recommended differential diagnosis, the evidence for each and recommended follow-ups were presented. Begun in 1984 and still in use, DXplain contains more than 4,900 clinical manifestations associated with more than 2,200 different diseases, which can generate more than 230,000 unique finding-disease interconnections. It uses a pseudo-probabilistic algorithm to assess the likelihood of a diagnosis.

**Iliad**

Iliad was developed at the University of Utah as a clinical decision support system for internal medicine diagnosis. It uses Bayesian networks to calculate posterior
probabilities for differential diagnoses. Iliad contains 1500 diagnoses and is still widely used as a training tool.

**VisualDX**

VisualDX ([www.visualdx.com](http://www.visualdx.com)) is a stand-alone, web-based diagnostic resource that is able to visually offer differential diagnoses based on a patient’s signs, symptoms, medical history, and other conditions. It was initially developed by Logical Images for physicians to diagnose dermatologic diseases. Its knowledge base contains 17,000 images and 1,000 visually identifiable diseases, drug reactions, and infections represented across all age ranges and skin types. VisualDX displays images of differential diagnoses ranked by number of matched criteria.

**SimuConsult**

SimuConsult ([www.simuconsult.com](http://www.simuconsult.com)) is online medical decision support software that offers consultation about a patient’s diagnosis based on the signs, symptoms, medical history, and lab findings. Begun in 2002, it uses a Bayesian model to produce differential diagnoses ranked by probability and contains more than 2,200 diagnoses. It will also suggest additional findings that the system thinks might be useful. All subsequent input changes the probabilities and rankings of the diagnoses.

**Isabel**

Isabel ([www.isabelhealthcare.com](http://www.isabelhealthcare.com)) is primarily a diagnostic checklist or reminder system which utilizes unstructured information culled from standard medical content. It contains approximately 3,500 diagnoses in a pre-designed diagnostic tree. Commercial software with advanced pattern recognition techniques was used to extract knowledge
and concepts from unstructured documents. Bayesian inference and Shannon’s information theory in the software enables those sophisticated extractions.

**DiagnosisPro**

DiagnosisPro ([www.diagnosispro.com](http://www.diagnosispro.com)) generates a hierarchical list of diseases based on signs, symptoms, and lab results. Its database contains 10,000 diseases, 20,000 findings, and 250,000 relationships. The system is able to provide a detailed review of each disease, including clinical presentations, abnormal lab findings, rule-outs, complications, treatments, and much more. It is particularly useful when a problem falls outside a physician’s specialty, since it covers many sub-areas of the medical domain.

**Problem Knowledge Couplers (PKC)**

PKC ([www.pkc.com](http://www.pkc.com)) has two major types of couplers: diagnostic couplers and management couplers. These either provide users with relevant possible diagnoses or present follow-up management options based on patient information collected through a series of questions and statements. The system couples individual findings with specific possible diagnoses and uses probability to measure the likelihood of each.

Table 2-2 summarizes the above clinical decision support system, with main features highlighted. It can be seen that most of the systems recommend and rank differential diagnoses, however only a few of them recommend getting further information, and none of them recommend lab tests based on their usefulness for differentiating differential diagnoses.
<table>
<thead>
<tr>
<th>Name</th>
<th>Time Developed</th>
<th>Main Reference</th>
<th>Major Technology</th>
<th>Medical Area</th>
<th>Ranking of differential diagnosis</th>
<th>Recommendations on Lab Tests?</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MyCIN</td>
<td>1970’s</td>
<td>(Shortliffe, 1976)</td>
<td>Expert system, Lisp, Knowledge base, inference rules,</td>
<td>Initially to identify bacteria that causes infections</td>
<td>Yes with certainty factors</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>CADUCEUS</td>
<td>1970’s-1980’s</td>
<td>(Pople, 1985)</td>
<td>Expert system, knowledge engineering, abductive reasoning</td>
<td>Internal medicine, contains 1000 diagnoses</td>
<td>Similar to MyCIN</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Internist-I/QMR</td>
<td>1970’s</td>
<td>(Miller, Pople, &amp; Myers, 1982)</td>
<td>Partitioning algorithm, exclusion functions, decision-tree logic</td>
<td>70-80% diagnoses in internal medicine</td>
<td>Yes with domain heuristics</td>
<td>Yes, by asking questions, recommend for further tests</td>
<td></td>
</tr>
<tr>
<td>DXplain</td>
<td>1980’s</td>
<td>(Barnett, Cimino, Hupp, &amp; Hoffer, 1987)</td>
<td>Pseudo-probabilistic algorithm, Bayesian logic, 1-5 importance level</td>
<td>2,200 different diseases, 4,900 clinical manifestations, 230,000 finding-disease interconnections</td>
<td>Yes based on likelihood</td>
<td>Yes by recommending follow-ups</td>
<td></td>
</tr>
<tr>
<td>VisualDX</td>
<td>2000’s</td>
<td><a href="http://www.visualDX.com">www.visualDX.com</a></td>
<td>60,000 images, 1,000 visually identifiable diseases, drug reactions, and infections</td>
<td>Rank of number of matched criteria</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>SimulConsult</td>
<td>2000’s</td>
<td><a href="http://www.simuconsult.com">www.simuconsult.com</a></td>
<td>Bayesian model, temporal information</td>
<td>2,000 diagnoses in congenital neurological and metabolic diseases</td>
<td>Ranked in probability</td>
<td>Suggest additional findings for the user to select</td>
<td></td>
</tr>
<tr>
<td>Isabel</td>
<td>2000’s</td>
<td><a href="http://www.isabelhealthcare.com">www.isabelhealthcare.com</a></td>
<td>Diagnostic tree, pattern recognition from unstructured documents, Bayesian inference, information theory</td>
<td>3,500 diagnoses</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>DiagnosisPro</td>
<td>2000’s</td>
<td><a href="http://www.diagnosispro.com">www.diagnosispro.com</a></td>
<td>15,000 diseases, 20,000 findings, 250,000 relationships</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Problem Knowledge Couplers (PKC)</td>
<td>1980’s</td>
<td><a href="http://www.pkc.com">www.pkc.com</a></td>
<td>Question answering, Over couplers 100 relationships</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>


2.3 Recognition-Primed Decision-making and Story Building

Recognition-Primed Decision (RPD), developed in the late 1980s (Klein, 1989), is a naturalistic decision-making model that focuses on the environment, where conditions may change over time, information may be ambiguous, and the plausibility of potential goals and courses of action may shift (Zsambok & Klein, 1997). Unlike prescriptive decision-making that attempts to prescribe how a decision should be made (Chaffee, 1981), RPD is a type of descriptive decision-making that attempts to describe how a decision is actually made in emergent situations.

RPD was developed to describe how experts make decisions in operational settings (Klein, 1989; Klein & Klinger, 1991). Within the RPD, the decision-maker has a set of experiences, each linking a situation to a course of actions. When the decision-maker faces a given situation, he refers to his past experiences in order to discern the most similar situation. If a situation is similar enough, the decision-maker chooses the same course of action that worked in the past. Figure 2-1 illustrates the complete RPD process.

When facing a new situation, a decision-maker will determine whether or not this situation is typical. If it is typical, the decision-maker will produce four byproducts:

- *relevant cues*: the factors that are used to compare the current situation with previous situations;
- *expectancies*: the consequences that are expected to happen after the decision is made and executed;
- *plausible goals*: the purpose of the decision;
- *course of actions*: a series of actions corresponding to the situation.

Next, the course of action will be evaluated by mental simulation. If it works in mental simulation, the course of action will be implemented immediately, with or without some modification. Otherwise, the decision-maker will go back to check other options or to reassess the situation.

If a situation is not matched to previous experiences or if an anomaly is found, the decision-maker will recognize that an unexpected situation exists which may be caused by an inaccurate assessment, possibly due to a lack of information critical to making the decision. The decision-maker must identify which information is missing. This and the existing information together will describe a situation, or a *story*. Therefore, the procedure of identifying and collecting missing information until a decision is made is called *story building*. 
As its name suggests, the RPD model relies on the recognition of situations. The critical step in RPD is checking whether or not a situation is typical in the decision-maker’s experiences. To respond quickly to a situation, the decision-maker must be alert to any change of the situation. This cognitive feature is termed *situation awareness* (Endsley, 1988).

Story building is part of situation assessment (Endsley, 2000). If only partial information is available, a decision-maker will need to assess it and determine what other
information can help connect the separate pieces of information to make the situation comprehensible. The Hypothesis-Driven Story Building framework which I have developed to model the behaviors of a decision-maker’s diagnosis-like activities derives from the basic understanding of story building proposed in RPD.

2.4 Related Techniques

Making a clinical diagnosis is essentially a process of finding explanations for unexplained symptoms by collecting evidence from physical examinations and lab tests. Many automatic methods, including logics, data structures, and algorithms have been developed to describe or solve similar problems. *Abductive reasoning* is a formal logic used to describe the explanatory relationship. *Bayesian networks* can be used for making probabilistic inferences, including calculating the probability of a cause from a set of observed facts. Bayesian networks have many types of extensions, some of which have attempted to integrate Bayesian probabilistic reasoning with logic representations. These *extended Bayesian networks with logic representations* attempted to enable probabilistic inferences of predicate-represented knowledge. These methods are related to generating and evaluating explanations.

*Value of information (VOI)* in decision theory acknowledges the value of information as guidance for seeking missing information. Another approach to defining VOI is based on *information theory*, where information entropy is used to measure the uncertainty of a random variable, and information gain is used to measure the capability
of an action to change the uncertainty of a random variable. Both theories are related to the information seeking.

### 2.4.1 Abductive Reasoning

Deduction, induction, and abduction are three major forms of logic reasoning. Deduction is the process of deriving propositions from other propositions based on certain logic rules, and is an inference procedure from generic to specific. Induction is the process of generalizing a series of specific instances with common features, and is an inference procedure from specific to general. The inductive reasoning is considered a reverse procedure of deductive reasoning.

Different from both deduction and induction, abduction is the process of discovering *explanations* for a phenomenon or a problem. Peirce (1903) described abduction using a modern logic perspective. He described abduction as follows:

“The surprising fact, \(C\), is observed;

*But if \(A\) were true, \(C\) would be a matter of course;*

*Hence, there is reason to suspect that \(A\) is true.*” (pp. 188-189)

Peirce described abductive reasoning as a process of finding out explanations (which are “\(A\)” in his example) for surprising observations (which are “\(C\)” in his example).

Abductive reasoning was formalized in a logic reasoning system called Theorist (Poole, Aleliumas, & Goebel, 1987), where \(\Sigma\) represents the knowledge set, and \(H\) represents the hypothesis set; a hypothesis \(h\) is generated from \(H\). For the observation set \(O\), if \(h\) satisfies:
(1) $\Sigma \not\models O$: The observation set $O$ can not be implied from current knowledge set $\Sigma$;

(2) $\Sigma \cup h \models O$: Together with $h$, $\Sigma$ can imply $O$;

(3) $\Sigma \cup h \models \square$: $\Sigma$ and $h$ together do not imply any contradiction;

then $h$ can be considered as a hypothesis (Poole, Aleliumas, & Goebel, 1987).

Other logic systems have been proposed to formalize abductive reasoning (Mayer & Pirri, 1993, 1995; Flack, 2000; Gouveia & Sernadas, 2001). For example, Flack (2000) claimed that

$$\models \alpha \leftrightarrow \beta, \alpha \prec \gamma \begin{array}{c} \hline \beta \prec \gamma \end{array} \tag{2-1}$$

and

$$\models \alpha \leftrightarrow \beta, \gamma \prec \alpha \begin{array}{c} \hline \gamma \prec \beta \end{array} \tag{2-2}$$

where the symbol “$\prec$” is read as “can be explained by”. Therefore, Eq. 2-1 means that if $\alpha$ and $\beta$ are logically equivalent, and $\alpha$ can be explained by $\gamma$, then $\beta$ can also be explained by $\gamma$. Similarly as indicated by Eq. 2-2, if $\alpha$ and $\beta$ are logically equivalent, and $\beta$ can be explained by $\gamma$, then $\alpha$ can also be explained by $\gamma$.

Since abduction is an inference method for finding explanations for observations, there may be more than one explanation for the same observation. Among those explanations, some may be more appropriate than others. From a logic perspective, certain criteria have been proposed to determine the appropriateness of explanations. The “basic” criterion stipulates that a hypothesis should be elementary and not require any
other explanation. The “minimal” criterion suggests that a better hypothesis is composed of fewer sentences (Kakas, Kowalski, & Toni, 1993).

Abductive reasoning utilizes the logic relations between propositions to make inferences of explanations, and does not usually take probabilistic causal relations into consideration. Therefore, there is a necessity to integrate abductive reasoning with probabilistic reasoning methods, such as is found in Bayesian networks.

### 2.4.2 Bayesian Networks

A Bayesian network is a directed acyclic graph that represents the conditional independencies of variables with probabilities (Pearl, 1985, 1988). A Bayesian network for a set of variables \( X = \{X_1, ..., X_n\} \) is composed of (1) a network structure \( S \) that encodes a set of conditional independencies of variables in \( X \), and (2) a set of local probability distributions \( P \) associated with each variable (Heckerman, 1999).

Since a Bayesian network encodes variables and their conditional independencies, it can be used to answer probabilistic queries about the variables. The basic principle of Bayesian inference is based on the Bayes Theorem, expressed as

\[
P(H \mid E, c) = \frac{P(H \mid c) \times P(E \mid H, c)}{P(E \mid c)}
\]

where \( H \) is a hypothesis, \( E \) is the evidence, and \( c \) is the background context. Basically, Eq. 2-3 represents the following fact: based on the same background context \( c \), the probability of an unconfirmed hypothesis \( H \), given evidence \( E \), can be computed from the
prior probability of H, the conditional probability of evidence E on hypothesis H, and the prior probability of E.

Although Eq. 2-3 shows the basic principle of calculating the posterior probability given any hypotheses and evidence, its computational cost is extremely high. Therefore, it is necessary to utilize the features of Bayesian network to simplify computation processes. One important feature is \textit{d-separation}. Two disjoint sets of variables A and B are conditionally independent given C, if C d-separates A and B; that is, if along every undirected path between a variable in A and a variable in B there is a variable D such that: (1) D has converging arrows and neither D nor its descendents are in C, or (2) D does not have converging arrow and D is in C (Geiger, Verma, & Pearl, 1990b, 1990a). D-separation can be explained in this way: given the observation of any variable in C, if the conditional probability on a node in A does not affect the conditional probability on a node in B, then C d-separates A and B. D-separation is used to distinguish the relevant variables from irrelevant variables for a given variable, and thus critical to Bayesian inferences.

Based on d-separation, many algorithms have been developed for making inferences about Bayesian networks such as variable elimination (Zhang & Poole, 1994), clique tree propagation (Zhang & Yan, 1997), and recursive conditioning (Zhang & Poole, 1996).
2.4.3 Combining Logic with Probability

Rule-based methods (such as predicate logic) and probability-based methods (such as Bayesian networks) have been successfully applied in many areas. However, there are also areas where problems cannot be solved solely by one method. Therefore, efforts have been made to integrate the two. In this section, I will give a brief review of some of these efforts.

Probabilistic Logic (Nilsson, 1986) was an attempt to combine probability theory with deductive logic, so that the deductive logic could be used to process probabilistic inferences. Instead of using “true” and “false” as the truth values such as they are in the traditional deductive logic, probabilistic logic uses probability values as its truth values.

Independent Choice Logic (ICL) (Poole, 1993), a logic language composed of a syntax and semantics, was an attempt to integrate Bayesian networks with first-order logic. It can be seen either as adding independent stochastic input to a logic program, or as a new way to allow Bayesian networks to have rule-based specifications with logic variables (Poole, 2008). An ICL theory consists of: F - the facts, an acyclic logic program; C - a choice space; P₀ - a probability distribution over the alternatives in C. Its semantics are defined in terms of possible worlds, where each is governed by independent choices. What is true in each possible world is determined by both the choices made in that possible world and the logic program.

Probabilistic Inductive Logic Programming (Raedt & Kersting, 2004) extended inductive logic programming, which generates hypotheses based on positive examples, negative examples, and background knowledge, with two essential changes: (1) clauses
are labeled with probability values, and (2) the covers relation will be probabilistic. The task is to find the hypothesis $H$ that maximizes the probability $P(E|H, B)$, where $E$ is the set of examples, $H$ is the set of hypotheses, and $B$ is the set of knowledge.

_Bayesian Logic Programming_ (Kersting & Raedt, 2007) unifies Bayesian networks with logic programming to allow propositional character of Bayesian networks. A Bayesian logic program consists of a set of clauses, where each is an expression of form $A|A_1, \ldots, A_n$, meaning $A$ is possibly dependent on $A_1, \ldots, \text{and } A_n$. Each clause is associated with a conditional probability distribution. Integrating logic programming with Bayesian networks allows the expression of the dependent relationships not only for random variables, but also for relations that are represented in predicates.

_Relational Bayesian Networks_ (Jaeger, 1997) were designed to solve the problem of a standard Bayesian network’s inability to express the conditional probabilistic relationships for _relations_, since a single random variable in the standard Bayesian network represents _attributes_ instead of relations. Therefore, relational Bayesian networks extend the standard to allow the expression of probabilistic relationships among relations, giving a class of probability formulas over the relational vocabulary.

_Probabilistic Relational Models (PRM)_ (Pfeffer, 2000) combines frame-based logic representation using probabilistic semantics within Bayesian networks. It extends Bayesian networks through concepts of objects, their properties, and the relations between them. It is also known as the Object-Oriented Bayesian Network (OOBN) (Koller & Pfeffer, 1997), providing the semantics for modeling objects and organizational structure. Within an OOBN, nodes are represented in complex classes. Classes or attributes may be linked by causal relations. OOBN encapsulates attributes in
objects or classes and thus can represent complex structures. However, the OOBN representation limits the possible causal relations among the attributes from the same class and does not differentiate the probabilities for different values of an attribute.

*Markov Logic Networks* (Domingos & Richardson, 2004) was an approach to integrate first-order logic and probabilistic graphic models. It attempted to attach a weight to each clause of the first-order knowledge base, where the weight is learned from relational databases by iteratively optimizing a pseudo-likelihood measure. The underlying philosophy of Markov logic networks is that when a world violates one formula in the knowledge base, it will be less probable, instead of being impossible as specified in traditional first-order logic. A world is more probable if it violates fewer formulas. Therefore, the strength of each formula can be represented by a weight.

*Multi-Entity Bayesian Networks* (MEBN) (Laskey, 2006) is another attempt to allow Bayesian inferences on first-order predicates. Similar to first-order logic which extends proposition logic to provide an inner structure for sentences, the MEBN logic extends ordinary Bayesian networks to provide an inner structure for random variables. However, it does not differentiate the possibly different probabilities for different values of parameters of a Bayesian node.

* Lifted First-Order Probabilistic Inference* (Braz, Amir, & Roth, 2005) was proposed to overcome the problem that many algorithms accepting first-order specifications were still performing proposition-like inferences because of the instantiation of first-order constructs. Therefore, the first-order inferences need to be “lifted”, so that inferences can be conducted not only on individuals but also groups of individuals. This process will greatly speed up the inferences.
2.4.4 Value of Information from Decision Theory

As modeled by principles of decision analysis, a decision-maker tends to take the action that can produce maximum expected utility. The expected utility takes into account both the utility and the probability of each consequence. Since the probability of a consequence is dependent on information availability, it is important to measure the value of information. The value of information can be defined as the difference in expected value between best actions before and after information is obtained (Howard, 1966).

The behavior of an information seeker can be regarded as an activity to obtain the value of a random variable $E_j$, which is yet unknown. Therefore, according to decision theory, the expected utility that can be achieved given current information may be calculated as

$$EU(\alpha \mid E) = \max_A \sum_i U(Result_i(A))P(Result_i(A) \mid E, Do(A))$$

(2-4)

where $E$ is the currently known information, $A$ is an possible action, and $\alpha$ is the best action among all actions. Therefore, Eq. 2-4 indicates that the expected utility of a best action under current information is the maximum expected utility among all those brought by all possible actions.

Suppose we have the new obtained information of $E_j$. The expected utility may now have changed. The new expected utility can be calculated by adding new information to it, such that

$$EU(\alpha_{E_j} \mid E, E_j) = \max_A \sum_i U(Result_i(A))P(Result_i(A) \mid E, E_j, Do(A)).$$

(2-5)
However, before obtaining the real value of the random variable $E_j$, we only know its distribution on possible values. Therefore, the value that knowing $E_j$ can bring is calculated as the expected maximum utility over the distribution of $E_j$ given $E$ on different values, shown as

$$VPI_E(E_j) = \left( \sum_k P(E_j = e_{jk} \mid E)EU(\alpha_{e_{jk}}) \right) - EU(\alpha \mid E). \quad (2-6)$$

The information value theory was proposed by Howard (Howard, 1966) and has been applied in real application domains for information seeking, such as medical decision-making (Yakota & Thompson, 2004).

The definition of value of information focuses on the expected benefit that knowing the information can add in terms of utility. However, for other problems, the utility-based approach is not appropriate. This is because first, the utility of each action is difficult to estimate for decision-makers, and second, decision-makers may not have anticipated all possible consequences, let alone their utilities. For these problems, it is necessary to explore other definitions of value of information.

### 2.4.5 Information Theory

Information theory was initially developed in 1948 by Claude E. Shannon to determine the minimal number of required bits on average for reliable storing and communication of data. Its key concept is entropy.
2.4.5.1 Information Entropy

Information entropy is a measurement of the amount of uncertainty of a random variable. Suppose a discrete random variable has $n$ possible values $\{x_1, x_2, ..., x_n\}$, the entropy $H$ measures the expected information content over all possible values, such that

$$H(X) = E(I(X)) = \sum_{i=1}^{n} p(x_i) I(x_i) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$  \hspace{1cm} (2-7)

where $b$ is the base of the logarithm, and $I(x_i)$ is the information content of the $i$-th value for $X$.

The information content of a value measures how much “surprise” it will deliver when the random variable is finally observed with its value. If the probability of a value is 1, it is certain that the random variable will finally be observed with its value and there is no surprise; i.e., information content is 0. However, if the probability of a value is 0, it is certain that the random variable will never be observed with the value. Therefore, the surprise will be infinitely large when it is observed with the value.

The entropy of a random variable, therefore, measures the expected surprise over all possible values. Since the probabilities of a random variable over all the possible values add up to 1, the entropy can thereby be used to measure the distribution of probabilities. The entropy value of a random variable reaches its maximum if the probabilities are equally distributed over all possible values, and reaches its minimum if one of the values has the probability of 1.
2.4.5.2 Information Gain

Information gain, also known as Kullback-Leibler divergence, is a measure of the difference between two probability distributions of a discrete random variable (Kullback & Leibler, 1951). Whenever the probability distribution is changed over possible values, its entropy is also changed. Since entropy is typically used to measure the probability distribution, the information gain is therefore calculated by the difference between two entropies such that

\[ IG = H(X') - H(X) \]  

(2-8)

where \( H(X') \) is the entropy after the change of probability distribution, and \( H(X) \) is the entropy before the change of probability distribution. Therefore, information gain is also known as entropy loss.

The change of probability distribution of a random variable may be caused by an event. An event may have multiple outcomes, and each may change that probability distribution. However, the outcomes of the event may still be uncertain and probabilistic. Therefore, the expected information gain or expected entropy loss is used to measure the effectiveness of the event such that

\[ \overline{IG} = \sum_{i=1}^{n} p(c_i)IG(c_i) \]  

(2-9)

where \( p(c_i) \) is the probability of the possible outcome \( c_i \) of the event, and \( IG(c_i) \) is the information gain caused by the \( c_i \).

Expected entropy loss has been used to select events based on their contributions to the change of probability distribution. For example, expected entropy loss has been
used to the problem of feature selection for information retrieval (Etzkorn & Davis, 1997; Glover et al., 2001; Ugurel et al., 2002)

2.5 Summary

This chapter focused on four directions of the related literature.

The first direction is human errors and cognitive biases. Several types of cognitive biases and their influence to the quality of clinical decision-making were reviewed. Chapter 7 will focus on two specific types of cognitive biases (anchoring heuristics and confirmation biases) and show evidence that these biases can be counteracted by a decision support tool with this feature designed in the underlying framework.

The second direction is the medical decision-making and clinical decision support systems. Factual evidence of medical errors and misdiagnosis in the U.S. and types of misdiagnosis were presented. The roles and purposes of clinical decision support systems were reviewed, and a series of clinical decision support systems focused on making diagnosis recommendations were presented. The remainder of this dissertation will focus on the design of a decision support system that will help improve diagnosis accuracy and cost efficiency with an emphasis on possible ways to counteract cognitive biases.

The third direction is the recognition-primed decision-making theory and story building, which inspired my Hypothesis-Driven Story Building framework.

The fourth direction is the related techniques for building a clinical decision support system. These include abductive reasoning, Bayesian networks and their
extensions with logic presentations, value of information in decision theory, and information theory.

Chapter 3 describes the details of the techniques which I used to develop a clinical decision support prototype.
Chapter 3

MULTI-LAYER BAYESIAN NETWORK: MODELING PROBABILISTIC CAUSALITY WITH VARIABLE BINDING

3.1 Introduction

The capability of inferring unknown information by a decision support system significantly relies on how the causal relations are represented and how the inferences are made. Logic-based knowledge representation and probability-based knowledge representation are two major methods for knowledge representation, each focusing on a different nature of the causal relationship.

The logic-based knowledge representation takes the assumption of “truth maintenance” (Doyle, 1979; Kleer, 1986), which aims to maintain the consistency of the knowledge base whenever information is added, revised, or retracted. A logic sentence is semantically labeled either true or false in a logic-based system. First-order logic (predicate logic) is the basic form in many knowledge representation systems. In the first-order logic, inferences are made through predicate clauses. A predicate clause formally represents a sentence that expresses an n-ary ($1 \leq n < \infty$) relationship (or an attribute when $n=1$) among entities or objects. For example, $R(X,Y)$ is a binary predicate representing the fact that there is a relation $R$ between $X$ and $Y$, where $X$ and $Y$ can be substituted by either a variable (e.g. $?x$) or a constant (e.g. $a$). Its truth value is determined by the semantic relationship between the real entities of $X$ and $Y$. Therefore, a first-order logic
based knowledge representation system is able to investigate the relationships between entities and objects, which was not able to be done within propositional logic.

A query on a predicate clause is to find out the values for the variables (arguments) that make the predicate clause true. For instance, \{?x = a, ?y = b\} is the query result for \(R(?x, ?y)\), if it has been semantically defined that \(<a, b>\) satisfies the relation \(R\). In the query result \(?x = a, ?y = b\)\}, the variable \(?x\) is said bound to value \(a\), and the variable \(?y\) is said bound to value \(b\). Accordingly, a query on a collection of predicate clauses is to find out the values for the variables (arguments) that will make all the predicate clauses true. Apparently, within each single query, the variables that appear multiple times across predicates clauses must refer to same entities (bound to same values). This is known as **variable-bound constraint**, which constrains the entity-level inferences.

*Negation as Failure* (NAF) has been an important feature in logic programming (Clark, 1978). It takes the approach that the negation of a proposition is true if and only if the proposition itself fails to be derived. The underlying assumption behind NAF is the **closed world assumption**, which presumes that the world and all its facts can be completely known: anything that has not been known to be true must be false (Cadoli & Lenzerini, 1994). The closed world assumption has brought the systematic simplicity to logic programming and has been successfully applied in many systems. However, it has left out the possibility that in more cases a world can not be completely known. The new **open world assumption** was then proposed (Giacomo & Levesque, 2000), assuming that a world is oftentimes open: a fact that has not been known true may not have to be false; it could be labeled as “unknown”. The open world assumption is especially true in medical
diagnosis. A not-yet-discovered symptom does not necessarily mean that it does not exist. Therefore, under the open world assumption, the NAF approach is no longer appropriate for solving real world problems.

Logic programming usually deals with certainty, such as “true” and “false”. Although in multi-valued logic (Beziau, 1997) the concept of uncertainty has been introduced, such as “unknown”, there is no description of the degree of uncertainty which is usually described in probabilities.

A Bayesian network is able to describe the probabilistic causal relationships between variables. It is a directed acyclic graph where each node represents a variable and each edge represents the conditional dependency between variables (Pearl, 1985). A variable delivers information when it is assigned with a value. Each node is associated with a conditional probability table, which contains the conditional probability of each single value-assigned variable given any combinations of values from parent variables. Given any number of observed variables, the posterior probability of any other node with a particular value can be calculated, using a variety of inference algorithms (Zhang & Poole, 1994, 1996; Zhang & Yan, 1997). Because of their capability of probabilistic inferences, Bayesian networks have been widely applied in many fields, such as medical diagnosis, business management, and industrial engineering.

Despite the successful applications in many fields, the classical model is still insufficient for making probabilistic causal inferences for logic sentences, especially predicate clauses with variable-bound constraints.

1 The variable in Bayesian network is different from the variable in predicate logic, where a variable refers to an argument in a predicate clause.
In a classical Bayesian network, each node represents a variable which can be assigned with one of its possible values that are mutually exclusive. For instance, if a node in Bayesian network represents a proposition, it can have “true” and “false” as values, but can only have one value at a time. When a Bayesian node is used to represent a predicate, it is oftentimes a unary predicate in which there is only one parameter. Therefore, the only parameter can be used to represent the node. An n-ary predicate can also possibly be represented by a Bayesian node only if the n-1 parameters are constants instead of variables, so that only one variable is left in the predicate and can represent the predicate itself. However, the Bayesian network has difficulties in processing the relationships of predicates that have same variable names across one another.

The difficulties of Bayesian inference for predicate clauses in such situations is caused by the variable-bound constraints from two aspects: First, nodes are connected not only by the conditional dependency, but also by the crossing variables. In a predicate Bayesian network, a query on a node with a particular variable binding will limit the observed values of other nodes. The variable bindings will be propagated to all the conditionally dependent nodes. With different variable bindings, a node in the Bayesian network may have different observations (true, false, or unknown). It is important to apply the right observation for each variable binding. Different variable bindings may have different effect on the initial query. Second, the conditional probability table alters with different variable bindings. Under classical models of Bayesian networks, the prior probabilities or the conditional probabilities of one node are unique. These probabilities will no longer be fixed because the conditional dependency may change with variable bindings. Each variable binding may have its own influence on the degree of conditional
dependency between nodes. For example, the probability of getting a particular disease on pregnant women may be significantly distinct from others. This feature enables a more accurate distinction of probabilities among different categories of entities.

Therefore, to allow probabilistic causal inferences for predicate clauses, Bayesian networks need to be extended to address these two issues. Multi-Layer Bayesian Network (MLBN) is such an attempt.

### 3.2 Multi-Layer Bayesian Network: Definitions

Although Bayesian networks are often regarded to represent the causal relationships between variables and are often used for making causal inferences, it is not exactly the case. The conditional probabilities encoded in a Bayesian network only reflect the conditional dependency relationship between variables, rather than their actual causal relationships. For example, Figure 3-1 shows a Bayesian network that encodes the conditional probabilistic relationships between pulmonary embolism and its symptoms - chest pains and dyspnea. It also shows that the occurrence of dyspnea may be affected by the patient’s age and sex, because a young active man may have lower probability to present dyspnea than an older woman.

There are limitations for the Bayesian network in Figure 3-1 to make inference for the following two requirements:

1. Abductive reasoning: In clinical diagnosis, a physician wants to find out the clinical reasons why the patient has a symptom. In Figure 3-1, pulmonary embolism should be the only clinical reason of dyspnea. However, in the
presentation of the Bayesian network, there is no syntactic difference between the node of pulmonary embolism and the node of age and sex, since they all contribute to the conditional dependency of dyspnea as depicted by the Bayesian network. It is hard for an inference engine to distinguish the real causes and other influential factors (or constrains) without domain knowledge. Therefore, we need a different representation for the causes and the constraints, so that an inference engine can find the correct explanations for the observed information.

2. Reasoning about relations: In some diagnosis situation, we need to know not only the existence of chest pain, but also the location (pleuritic or back), type (dull or acute), or severity (low, medium, or high). These are the properties of chest pain that together describe the chest pain. Therefore, we need a predicate ChestPain(?id, ?location, ?type, ?level) to describe the chest pain, instead of the simplistic representation of a variable.

---

Figure 3-1: A Classical Bayesian Network

Figure 3-2 shows a revised Bayesian network that differentiates causes and constraints and also enables conditional probabilities for relations. In this Bayesian network, the main structure reflects the real causal relationships between pulmonary
embolism, chest pain, and dyspnea. Age and Sex now became two factors that constrain the probabilistic causal relations between pulmonary embolism and dyspnea. For each special combination of the constraints, a special conditional probability table can be defined. For example, we can define a condition probability table on dyspnea for “women older than 50” and another for “men younger than 20”. These two conditional probability tables can be totally different.

No matter the constraints or the relations, the revised Bayesian network needs to process the information and inference at the parameter level inside the nodes, since the value of the node itself will now depend on the semantic relations defined for the possible values of the parameters. The parameters assigned with values can be named as variable binding.

**Definition 3-1: Variable Binding**

A \( n \)-ary variable binding is a collection of variable-value pairs \( \{ \text{var}_1 = \text{value}_1, \text{var}_2 = \text{value}_2, \ldots, \text{var}_n = \text{value}_n \} \), where \( \text{var}_i \) is a variable name, and \( \text{value}_i \) is a value assigned to the variable. If the variables are known in a certain context, a variable binding can be abbreviated as \( \{ \text{value}_1, \text{value}_2, \ldots, \text{value}_n \} \) with default sequence.
For example, \( \{\text{Age} = 50+, \text{Sex} = \text{male}\} \) is a variable binding that specifies age and sex. In this variable binding, the variable \( \text{Age} \) is said to be bound to value “50+”, and the variable \( \text{Sex} \) is said to be bound to value “male”.

A meaningful variable binding needs to satisfy two conditions: (1) each variable has at most one occurrence; and (2) the value assigned to each variable belongs to a predefined value set.

A variable binding defines a group of objects that satisfy all the conditions defined in it. For example, the variable binding \( \{\text{Age} = 50+, \text{Sex} = \text{male}\} \) defines a group of people whose age is older than 50 and whose gender is male. Therefore, we can compare two variable bindings by comparing the two object sets that are defined by the variable bindings. For example, it is reasonable to say that the variable binding \( \{\text{Age} = 50+, \text{Sex} = \text{male}\} \) is smaller than the variable binding \( \{\text{Sex} = \text{male}\} \), because the object set defined by the former is a subset of the object set defined by the latter, i.e. the former is subsumed by the latter. We can also define a partial relationship between variable bindings using the covering relationship.

For any variable binding \( B \), let \( \text{Var}(B) \) represent the set of all the variables contained in \( B \), and \( \text{value}(\text{var}, B) \) represent the value assigned to variable \( \text{var} \) in \( B \). Then the subsumption relationship between variable bindings can be defined as:

**Definition 3-2: Subsumption Relationship of Variable Bindings**

For any two variable bindings \( B_1 \) and \( B_2 \):

\[ B_1 \leq B_2, \text{ if and only if } \text{Var}(B_2) \subseteq \text{Var}(B_1), \text{ and for each variable } \text{var} \text{ in } B_2, \text{value}(\text{var}, B_2) = \text{value}(\text{var}, B_1). \]

\[ B_1 = B_2, \text{ if and only if } B_1 \leq B_2 \text{ and } B_2 \leq B_1. \]
As defined by the subsumption relationship, a greater variable binding has a greater object set, and thus has smaller number of constraints on variables. Be noted that the subsumption relationship is not based on the number of constraints in the variable binding, but based on the set of objects that the constraints have defined. Also, because the type of objects that a variable binding defines is not specified, the subsumption relationship can only compare two variable bindings that one contains all the variables from the other. Therefore, the empty variable binding (\{\} ) is greater than any other variable bindings.

The definitions of variable binding and its subsumption relationship are particularly useful for Bayesian networks making probabilistic inferences between predicates. If each node in a Bayesian network represents a predicate, its truth value is determined by the variable bindings whose variables are the predicate’s arguments.

To be general, we can relax a Bayesian node from representing a predicate to any function, which is a mapping from combinations of values to values.

**Definition 3-3: Function**

A \(n\)-ary function \(f_n\) is a mapping:

\[ f_n : X_1 \times X_2 \times \ldots \times X_n \rightarrow X_{n+1}, \]

where \(X_i\) is any category of objects or entities.

A \(n\)-ary function \(f_n\) actually defines a \((n+1)\)-ary relation: \(f_n \subseteq X_1 \times X_2 \times \ldots \times X_n \times X_{n+1} \).

A \(n\)-ary predicate is a special case of \(n\)-ary function, where \(X_{n+1}\) has only two values \{true, false\}. For simplicity, a predicate is redefined.

**Definition 3-4: Predicate**
A *predicate* \( p \) is a \( n \)-ary relation among objects from \( X_1, X_2, \ldots, \) and \( X_n \), i.e.,
\[
 p \subseteq X_1 \times X_2 \times \ldots \times X_n,
\]
where \( X_i \) is any category of objects.

\( p(x_1, x_2, \ldots, x_n) \) is a *predicate clause*, where \( x_i \) is a variable or a constant.

An instantiated predicate clause \( p(a_1, a_2, \ldots, a_n) \) is true, if and only if \( a_1 \in X_1 \), \( a_2 \in X_2 \), ..., and \( a_n \in X_n \), and \( (a_1, a_2, \ldots, a_n) \in p \).

A query on a predicate is to find out all variable bindings that satisfy the relationship defined by the predicate. More generally, a query on an \( n \)-ary function is to find out all \((n+1)\)-ary variable bindings that satisfy the relationship.

Since function is a general format of a predicate, the following definition for Multi-Layer Bayesian Network will be built based on function.

**Definition 3-5: Multi-Layer Bayesian Network (MLBN)**

A Multi-Layer Bayesian Network is an extended Bayesian network. Specifically, a MLBN = \( \{N, O, V, U, F, E, P\} \), where:

- \( N = \{N_1, N_2, \ldots, N_n\} \): a set of nodes, each node represents a function;
- \( O = \{O_1, O_2, \ldots, O_n\} \), where \( O_i \) is a set of possible observations of the node \( N_i \);
- \( V = \{var_1, var_2, \ldots, var_m\} \): a set of random variables;
- \( U = \{U_1, U_2, \ldots, U_m\} \), where \( U_i \) is a set of possible values for the variable \( var_i \); \( U_i = \{u_{i1}, u_{i2}, \ldots, u_{ik}\} \);
- \( D = \{D_1, D_2, \ldots, D_m\} \): where \( D_i \) is the probability distribution of values from \( U_i \) on the population; \( D_i = \{p_{i1}, p_{i2}, \ldots, p_{ik}\} \); \( p_{ij} \) is the probability for value \( u_{ij} \);
- \( F = \{F_1, F_2, \ldots, F_n\} \), where \( F_i \subseteq U_{i1} \times U_{i2} \times \ldots \times U_{ik} \rightarrow O_i \cup \{unknown\} \); each \( F_i \) is represented in a node \( N_i \).
\[ E = \{e_i | e_i = \langle N_i^j, N_i^k \rangle \} \] : a set of directed edges connecting nodes in \( N \), without forming any loops;

\[ P = \{P_1, P_2, \ldots, P_n\} | U_1 \times U_2 \times \ldots \times U_m \], where

\[ P_i = \{P_{[u_1, u_2, \ldots, u_m]} (*O_i | *O_i^1, *O_i^2, \ldots, *O_i^k) | \{u_1, u_2, \ldots, u_m\} \in U_1 \times U_2 \times \ldots \times U_m\},\] and \(*O_i^j\) represents any observation of the node: the conditional probability distribution function of node \( N_i \) given a combination of its parents’ observations \( < *N_i^1, *N_i^2, \ldots, *N_i^k > \) and a variable binding \( \{u_1, u_2, \ldots, u_m\} \).

The MLBN definition is further explained in the following paragraphs.

\( N \) represents the collection of all \( n \) nodes of a MLBN. Like classical Bayesian networks, a node represents a variable that could be observed with any one of several possible values. Here \( O \) represents the collection of \( O_i \)'s, each \( O_i \) represents the collection of possible observation values of a node. For example, if a node represents body temperature, the possible observation values of this node could include in \( O_i = \{\text{low, normal, low fever, high fever}\} \).

The difference between a classical Bayesian node and a MLBN node is that an MLBN node can be further decomposed to its attributes. Therefore, a MLBN node is able to represent a result value from a function with \( k \) inputs rather than just a single variable. For example, a MLBN node could be like: \( bt = \text{bodyTemperature}(?who, ?date, ?time) \), which is a function of getting the body temperature of a person on a specific date at a specific time. Therefore, each node in an MLBN is associated with a function. Each function defines a mapping from a collection of values to an observation value or an unknown value. For example, for the node that represents the body temperature, \( bt \) could be “high fever” if the body temperature for the specific person on the specific date at the
specific time has been observed as “high fever” or “unknown” if body temperature has not been tested.

In the MLBN, \( V \) represents a complete set of variables, which is a union set of variables from every node. For each variable \( \text{var}_i \) in \( V \), \( U_i \) defines the set of possible values that the variable can be assigned with, and \( D_i \) defines the probability distribution of \( \text{var}_i \) across all the values in \( U_i \).

Same as classical Bayesian networks, nodes in MLBN are also connected by edges without forming any loop. \( E \) represents a collection of edges of a MLBN.

\( P \) aims to define the multiple layers of conditional probability tables. First, a conditional probability table in a classical Bayesian network contains all the conditional probabilities of the node given any combination of values from parent nodes, which is represented as \( P(*O_i|*O_{i1},*O_{i2},...,*O_{ik}) \). Second, the conditional probability table may vary according to different variable bindings. Therefore, an abbreviated variable binding \( \{u_1, u_2, ..., u_m\} \) is used to define the specific layer of the conditional probability table.

A Multi-Layer Bayesian Network is an extended Bayesian network. Therefore, it is still a directed acyclic graph, but with each node representing a function, instead of a variable. In the rest of the paper, I may use “function” and “node” interchangeably, since function is the content that a node represents. The term “variable” is used to refer to an argument of a function, rather than a node in classical Bayesian networks.

The conditional probability tables are now more complicated in MLBNs than in classical Bayesian networks. A probability value is not only determined by the causal relations among different nodes, but also determined by variable-bound constraints. Since
the structure of the network remains relatively stable, a new variable-bound constraint
adds one more layer of conditional probability table to the node. Therefore, for each node,
there will be multiple probability tables associated with it. Each table is defined by a
particular variable binding, and each variable binding forms one layer of a Bayesian node.
Each node should include the empty variable binding ({} ) as its default layer for default
probability definition.

As discussed earlier, a variable binding actually defines a particular group of
objects from population. Each variable-value pair in the variable binding specifies a
condition that this particular group of objects needs to satisfy. Each condition narrows
down the group of objects by applying the constraint. Since the probability distribution of
each category of objects is defined ($D_i$ in the definition), the probability of a variable
binding can be obtained. Suppose the occurrences of objects from different categories are
independent, the probability of a variable binding then can be calculated by Eq. 3-1.

\[
p(? v_{1}, ? v_{2}, ..., ? v_{k} = v_{k})
= \prod_{i=1}^{k} p(? v_{i} = v_{i}) \tag{3-1}
\]
Figure 3-3 shows a sample MLBN with four nodes. The conditional probabilities for each node are defined for several variable bindings. For example, the probability for node $A$ to be true under variable binding $\{?a = a_1\}$ is 0.9, while the same probability becomes 1.0 under variable binding $\{?a = a_2\}$; for any other variable bindings, the probability for $A$ to be true is 0.8, retrieved from the default layer.

When making probabilistic inferences with a MLBN, it is the first step to determine which layer to use given a particular variable binding $B$. If $B$ has been explicitly represented in a probability table, the required probability value can be directly retrieved from the tables. For any other variable bindings, the agent needs to find out the table whose defining variable binding is the closest to the requesting variable binding. This is defined by the closest covering variable binding.

**Definition 3-6: Closest Covering Variable Binding**

For any variable binding $B$, a variable binding $B_1$ is considered the closest covering variable binding of $B$, if and only if: (1) $B \leq B_1$; and (2) there is no other variable binding $B_2$, such that $B \leq B_2$, and $B_2 \leq B_1$. •
The closest covering variable binding works when there is no exact match of variable bindings. It helps to find out the closest conditional probability table. For example, if an agent needs to find the proper conditional probability table for variable binding \( \{a=a_1, b=b_2\} \) in node \( C(a, b) \), its closest covering variable binding is the empty variable binding \( \{\} \). Then the agent needs to find the probability from the first column in node \( C(a, b) \).

The closest covering variable binding also works as an interface between logic-based knowledge base query and the MLBN inference. More details of the MLBN inference and the use of closest covering relationship will be discussed in the next section.

### 3.3 Inferences with MLBN

Inferences with Bayesian network are the processes of reasoning about the probability of an unknown variable, given some other variables that have been observed as evidence. Similarly, inferences with MLBN are to reason about the probability of an unknown MLBN node, given other MLBN nodes that have been observed as evidence. However, inferences with MLBN are more complicated, since an MLBN node is not simply either observed or not. The probability of an MLBN node depends on variable bindings. Therefore, the inference algorithm on an MLBN has to deal with the propagation, the aggregation, and the selection of variable bindings.

In the rest of this section, I describe algorithms to support these probabilistic inferences. More specifically, I will describe algorithms that (1) merge variable bindings of observation nodes in a multi-layer Bayesian network, (2) retrieve conditional
probabilities from multi-layer Bayesian network, and (3) infer the posterior probability of a Bayesian node for a specified variable binding of the node, given a set of the observation nodes. Together, these algorithms enable the core inference capabilities of the Multi-Layer Bayesian Network.

As discussed earlier, a variable binding of a node in a MLBN specifies objects that satisfy the constraints defined by the variable binding. Therefore, a merging of two variable bindings combines the constraints from both variable bindings. Hence, conflicting constraints need to be resolved during the merge process. For example, the merging of variable bindings \{?a = a_1\} and \{?b = b_1\} produces \{?a = a_1, ?b = b_1\}, which further narrows down the group of objects specified by \{?a = a_1\} and \{?b = b_1\}. If the two variable bindings contain the same variable with different values, the two variable bindings are said to be conflicting, and can not be merged together. For example, \{?a = a_1\} and \{?a = a_2\} are conflicting and can not be merged, since the variable \(?a\) refers to different values \(a_1\) and \(a_2\). Algorithm 3-1 shows a preprocessing of merging two collections of variable bindings. Each variable binding in the first collection will be compared with each variable binding in the second collection. For a variable binding \(v_{bi}\) from the first collection, and a variable binding \(v_{bj}\) from the second collection, they can be merged only if all the variables they share have the same values. In the algorithm, \(Var(vb)\) represents all the variables that have appeared in the variable binding \(vb\), and \(value(var, vb)\) represents the specific value for variable \(var\) defined in \(vb\).

Algorithm 3-1: MERGE(\(vbs_1[0..n]\), \(vbs_2[0..m]\))

Input:
\(vbs_1[0..n]\): an array of variable bindings;
\(vbs_2[0..m]\): an array of variable bindings;

Output:
bindings[: an array of variable bindings;

Steps:
for each \( vb_i \in vbs_1[0..n] \), do
for each \( vb_j \in vbs_2[0..m] \), do
  if (for \( \forall \text{var} \in Var(vb_i) \): \( \text{var} \notin Var(vb_j) \) or \( \text{value}(\text{var}, vb_i) = \text{value}(\text{var}, vb_j) \)), do
    bindings[#] \( \leftarrow vb_i \cup vb_j \);
  end if
end for
end for
return bindings[:].

The MERGE method attempts to create global variable bindings from partial variable bindings, while the partial variable bindings are obtained through observations on a node. Algorithm 3-2 shows the method of obtaining the partial variable bindings for a given node with the initial variable binding \( vb \) and the initial observation value \( o \). It is actually the step of querying the knowledge base to get the variable binding.

Algorithm 3-2: OBSERVE(\( net, node, vb, o \))

Input:
\( net \): a MLBN;
\( node \): a MLBN node;
\( vb \): an initial variable binding;
\( o \): the given value of the node;

Output:
\( bindings[: \) : an array of variable bindings;

Steps:
for each \( vb' \leq vb \), do
  if \( F(vb') = o \), do
    bindings[#] \( \leftarrow vb' \);
  end if
end for
return bindings[:].

For each \( vb' \leq vb \), since \( \leq \) is defined as the partial relationship between variable bindings, the variable binding \( vb' \) defines no less entities than \( vb \). Therefore, \( vb' \) has
more constraints introduced by more variable-value pairs. If the function $F$ defined on the node satisfies $F(vb') = o$, then $vb'$ will be considered the actual variable binding obtained from the observation value $o$.

**Algorithm 3-3: INFER(net, node, vb, o)**

**Input:**
- $net$: a MLBN;
- $node$: a MLBN node;
- $vb$: an initial global variable binding obtained through observations from all nodes;

**Output:**
- $prob$: the probability of the $node$ to be observed $o$ given variable binding $vb$ in MLBN $net$.

**Steps:**
1. $vbs[1..n] = \{\}$
2. for each node $node_i \neq node$ in net, do
   1. $vbs[i] = \{\}$;
   2. for each observation $o_j$ of $node_i$, do
      1. $vbs_{ij} = \text{OBSERVE}(node_i, vb, o_j)$;
      2. $vbs[i] \leftarrow vbs_{ij}$;
   end for
end for
3. $\text{global}_vbs = \text{MERGE}(vbs[1..n])$
4. $Layer_i[\cdot] = \{\}$ //will contain the conditional probability tables of all nodes with $i$-th variable binding.
5. $Obs_i[\cdot] = \{\}$ //will contain the observed values for all observed nodes.
6. for each $vb_i \in \text{global}_vbs$, do
   1. for each node $node_j \neq node$ in net, do
      1. if $o_k = F_j(vb_i)$
      2. $Obs_i[\cdot] \leftarrow (node_j, o_k)$;
      3. $Layer_i[\cdot] \leftarrow \text{SELECT_PROB}(node_j, vb_i)$;
   end for
   2. $\text{probability}[i] = B\_\text{INFER}(Obs_i[\cdot], Layer_i[\cdot])$;
end for
7. $prob = \sum_i \text{probability}[i] \times p(vb_i | vb)$

**return** $prob$;

The key novelty of the algorithm is its utilization of the suitable “layers” of conditional probability distribution for computing the posterior probabilities of a node.
given a set of observations. The layer chosen by the algorithm is one that is associated with the variable binding that is the closest covering variable binding of the variable bindings generated by merging observation nodes. For instance, to compute the posterior probability of a male patient with age 65, the layer associated with “senior male” may be chosen. During this process, the algorithm automatically identifies potential variables that may be relevant (e.g., whether the patient is senior) from those associated with different layers of a Bayesian node in the MLBN. This enables the Bayesian inferences to use the most suitable probabilistic knowledge (i.e., the most suitable “layers” of conditional probability distribution) for a specific patient, as well as to capture these knowledge at the most general level (e.g., senior citizen) for their ease of reuse.

Algorithm 3-4 shows the procedure of computing the probability of a node with variable binding vb. For each query node nodei (e.g., a possible diagnosis), if it is not currently labeled with an observation value (e.g., true or false), there are potential variable bindings that could be identified from the structure of the network (e.g., lab tests that can be performed). Hence, each query may be linked to a series of potential variable bindings (an array vbsij, generated by OBSERVE function). vbsi then is the array that contains multiple vbsij’s for different observation oj. Then, vbs[1..n] is the array that contains vbsi for different nodei. The MERGE(vbs[1..n]) algorithm produces the global variable bindings where each variable binding is a full combination of variables from all nodes with specific observation. It is implemented in an iterative fashion (Algorithm 3-4).

**Algorithm 3-4: MERGE(vbs[1..n])**

*Input:*

vbs[1..n]: an array that contains n arrays, where each
array contains a number of vectors where each vector contains a series of variable bindings corresponding to each observation;

Output:
\( bindings[ \] \): an array of variable bindings;

Steps:
\( bindings\_rest[ \] = MERGE(vbs[2..n]); \)
\( bindings[ \] = MERGE(vbs[1], bindings\_rest[ \]); \)
\( \text{return} \ bindings[ \]. \)

For each global variable binding \( vb_i \), a node\( j \) can be decided whether observed with a specific observation \( o_k \) or not. At the same time, the specific layer of conditional probability table can be obtained by using the SELECT_PROB algorithm, which selects the conditional probability table associated with the “layer” whose variable binding is the closest covering the input of \( vb \).

**Algorithm 3-5: SELECT_PROB(node, vb)**

Input:
- \( net \): a MLBN;
- \( node \): a MLBN node;
- \( vb \): an initial global variable binding obtained through observations from all nodes;
- \( probability[ ] \): an array that saves probabilities for each variable binding;

Output:
- \( CPT \): a conditional probability table.

Steps:
\[ \text{If } \{u_1, u_2, ..., u_m\} \text{ is the closest covering variable binding of } vb, \text{ then} \]
\[ \text{ } \]
\[ CPT = P_{\{u_1, u_2, ..., u_m\}} (*O_1, *O_{i1}, *O_{i2}, ..., *O_{ik}) \]
\[ \text{return } CPT; \]

The Bayesian inference algorithm will be invoked to make Bayesian inferences. \( \text{B\_INFER(Obs[ ], Layer[ ])} \) returns the probability of the initial node given specific observation set and a specific layer of conditional probability table, at a specific variable
binding. It can be implemented by any Bayesian inference algorithm. The ultimate probability of the node to be $o$ is an expected probability on all global variable bindings. The probability of a global variable binding is a conditional probability given the initial variable binding, obtained from the probability distribution $D$ in MLBN definition.

The critical steps in the inference algorithm are to decide the variable bindings that are related to the inference, the observation value of each node given each variable binding, and the layer of conditional probability tables for each node. For example, in Figure 3-3, if $A(?a,?b)$ is queried for variable bindings $\{?a = a1,?b = b1\}$ and $\{?a = a3,?b = b1\}$, and $B(?a,?b)$ is queried for variable bindings $\{?a = a1,?b = b1\}$ and $\{?a = a1,?b = b2\}$. Then there are four different variable bindings that will affect the probabilities of the node $C$, which are $\{?a = a1,?b = b1\}$, $\{?a = a1,?b = b2\}$, $\{?a = a3,?b = b1\}$, and $\{?a = a3,?b = b2\}$. These four variable bindings make the nodes $A/B$ true/true, false/true, true/false, and false/false and produce four probabilities for node $C$: 0.7, 0.75, 0.76, and 0.556.

I briefly show how 0.76 was calculated here. Based on variable binding $\{?a = a3,?b = b1\}$, node $A$ is true, but node $B$ is unknown. For node $A$, the closest covering variable binding is the default variable binding at layer 1. For node $B$, the closest covering variable binding is the variable binding $\{?b = b1\}$ at layer 2. For node $C$, the closest covering variable binding is also the default variable binding at layer 1. Since node $A$ is the only evidential node, the probability for node $C$ is in fact the conditional probability based on $A$. By taking the proper probabilities, the probability for $C$ is $0.8 * 0.9 + 0.4 * 0.1 = 0.76$. 
3.4 Summary

MLBN was proposed based on two requirements: abductive reasoning and reasoning for relations. The abductive reasoning requires the classical Bayesian nodes that represent constraints to collapse into the nodes that represent causes, so that the Bayesian network will reflect the causal relationship. The reasoning for relations requires the Bayesian network to process the internal relations between parameters. These two requirements can both be realized with the proper representations of variable bindings. Variable bindings are used to define the layers of a node in the MLBN, so that the conditional probability tables are indexed by the variable bindings and the searching for the proper conditional probability for abductive reasoning will be efficient.
Chapter 4

HYPOTHESIS-DRIVEN STORY BUILDING

4.1 The Framework

Before a decision can be made, the situation in which the decision is made needs to be fully recognized. For example, before a treatment plan can be developed for a patient, his condition must be correctly diagnosed and his real disease must be correctly identified. This is why the RPD model focuses on recognizing a situation using previous experiences of the decision-maker.

A situation is described by a set of features, where each feature characterizes one attribute of the situation. From a knowledge representation perspective, a piece of information can be represented in a logic sentence, such as a proposition or a predicate. Therefore, a situation can be described with a set of logic sentences, either in propositions or predicates. A piece of uncertain information can be represented by a probability statement.

Each piece of information is directly observed or indirectly inferred. In a patient care situation, for example, the symptoms, the physical examination results, the lab test results, and the disease of a patient compose a description of the situation. The goal of diagnosis is to determine the disease that best explain the observed symptoms, physical examination results, and lab test results.

In many cases, however, the situation may not be fully described at the beginning
of a decision-making process. The initial description may contain only limited partial information, which requires a strategy to determine how the missing information should be identified and collected. The choice of what missing information to gather not only has direct implications on the quality and the timeliness of decision-making, but also effects of the overall cost of the decision-making process.

The construction of an explanation of the situation is called “story building” in RPD, which is an iterative process involving (1) synthesizing hypotheses (i.e., elements of the story) from current situation description, (2) identifying the relevant information that is missing, and (3) deciding what missing information to gather. For instance, a patient diagnosis process involves (1) generating hypotheses about possible diagnoses, (2) identifying relevant lab test results, and (3) ordering specific lab tests results for the patient.

However, there might be more than one possible set of missing types of information that, together with current existing information, might potentially compose a story. For example, in a patient care situation, the initial symptoms and physical examination results might be caused by any one of several possible diseases. Each disease along with the current symptoms and physical examination results can potentially compose a story. All potential diseases are considered during the initial phase of differential medical diagnoses. With additional lab test results, certain diagnoses can be ruled out, until a final diagnosis is reached.

In general, the concept hypothesis can refer to any suspected story that provides an explanation of observed information. For instance, a differential diagnosis is a hypothesis.
The process of story building, which involves finding out what types of information are missing and what their values are, depends heavily on what hypotheses are being generated. Therefore, story building is essentially a hypothesis-driven procedure. Hence, I use the term *Hypothesis-Driven Story Building* (HDSB) to refer to the process of generating a satisfactory explanation of a situation from limited initial information, with an emphasis on the key role of hypothesis during the process. Figure 4-1 illustrates the Hypothesis-Driven Story Building process in a formal way.

At the beginning of a story building process, the decision-maker is presented with a set of *observations* (e.g., a collection of observed facts). Based on the observation set, the decision-maker begins the *hypothesis generation* procedure and produces a *hypothesis space*, which consists of hypotheses that might explain the initial observations. The hypotheses generated are then evaluated, which produces the *ranked hypothesis space* in terms of the likelihood of each hypothesis. The ranked hypotheses are then used as guidance for seeking missing information.

The *information seeking* procedure aims to answer three questions: (1) what types of information are useful but missing; (2) if resources for information seeking are constrained, what types of information have a higher priority to seek; and (3) where to
obtain the missing information. The newly collected information triggers the hypothesis revision process. It determines whether or not to simply adjust the ranking of the hypotheses, or whether to rebuild the entire hypothesis space if a significant change in the situation has occurred. This process may iterate for several times until it finally generates a confirmed story with a satisfactory explanation of the situation.

4.2 The Formal Definitions of HDSB

Before introducing methods and algorithms of HDSB, I formally define the concepts of HDSB. Since the HDSB will be built on the basis of MLBN, these concepts are defined using MLBN terminologies.

A story building process starts when information is observed, but it is limited and insufficient to describe the situation.

Definition 4-7: Observation

An Observation $o$ is a MLBN node that satisfies:

1. the variables of the node have all been bound to values by a variable binding; and
2. the function $F$ defined on the node can produce a non-unknown value for the node based on the variable binding.

An observation describes a relationship between the values of variables and the value of a node. For example, $\text{bodyTemp}(\text{Tom, 05/10/2010, 9:00am}) = 100^\circ F$ is an observation which describes the fact that the body temperature of Tom on May 10, 2010 at 9:00 a.m. is $100^\circ F$. 

Suppose we have an MLBN as shown in Figure 4-2. The nodes D, E, and F have been observed with specific values, given specific variable bindings, and are considered observations in this situation. (An observation here refers to an observed fact, rather than the observation activity.)

**Definition 4-8: Observation Space**

An *Observation Space*, noted in \( O \), is a collection of observations.

In Figure 4-3, the area that covers D, E, and F is considered the observation space in the current situation.
An observation space may contain two or more different types of observations. Some observations need further explanation of why they are present. For example, in a patient care situation, “chest pain” is an observation and requires further explanation of what might have caused it. Other observations do not require further explanation, but are accepted as true facts. For example, “The patient has been to Africa” is an observation that requires no explanation. However, this information may still be useful for filtering the differential diagnoses.

An observation that needs further explanation may be considered an explanandum, while an observation that requires no explanation may be considered to be evidence. The distinction between “explanandum” and “evidence” is important for story building. The “explanandum” actively drives the story building process, whereas the “evidence” can only be passively drawn into story building on an if-needed basis.
A key challenge in identifying explanandum is to identify causal relationships in Bayesian network. As discussed in Chapter 3, the proposed Multi-Layer Bayesian Network collapses probabilistic dependencies whose causes we are not interested in into probabilistic dependencies whose causes we are interested in to form multi-dimensional edges in MLBN. Hence, we use these edges to identify explanandum. More specifically, a node with an incoming edge in MLBN (i.e. a node with a parent node in MLBN) is an explanandum if the node is observed. More formally, we have the following definition.

**Definition 4-9: Explanandum**

An *Explanandum*, noted as $em$, is an observation that an external factor has caused something to happen which requires further explanation. In a MLBN, an explanandum is a node if it satisfies two conditions:

1. it is an observation; and
2. it has parent nodes in MLBN.

**Definition 4-10: Explanandum Space**

An *Explanandum Space*, noted as $EM$, is a collection of explananda that contains all the explananda in a given situation.

Since each explanandum is an observation, an $EM$ will be a subset of $O$, i.e. $EM \subseteq O$.

**Definition 4-11: Evidence**

An *Evidence* node, noted as $ev$, is an observation whose cause is of no interest and therefore needs no further explanation. A piece of evidence is only accepted as true fact.

In a MLBN, a piece of evidence is a node if it satisfies two conditions:
(1) it is an observation;

(2) it has no parent nodes.

That an observation is regarded as an evidence node does not imply that this observation has no causes at all. It only indicates that the actual cause of the observation plays no role in the current story building. The recognition of the scope of causal relationships of interest to a story building process is highly significant. Because a story building process attempts to construct explanations to given observations by following “causal relationships” in the reverse direction, it can involve a very large number of causal relationships, if the scope of interest is not defined.

**Definition 4-12: Evidence Space**

An *Evidence Space*, noted as $EV$, is a collection of evidence nodes that contains all the evidence nodes in a given situation.

An observation is either an explanandum node or an evidence node; therefore, $EV \cup EX = O$, and $EV \cap EX = \emptyset$. Figure 4-4 shows the explanandum space and the evidence space of an observation. Nodes E and F compose the explanandum space, since the observation of E and F might be explained by the parent nodes, A, B, or C. Node D itself composes the evidence space, since no other nodes can explain the observation of D.
Among nodes in MLBN that do not have direct observation, those that are ancestor nodes to explanandum are the “hypotheses” to explain the explanandum. I formally define it as follows.

**Definition 4-13: Explanans**

An *explanans*, noted as $ex$, is a MLBN node that satisfies two conditions:

1. it is not an observation node;
2. it is an ancestor node of an explanandum node.

An explanans is an unobserved MLBN node that could be potentially used to explain an explanandum. Since MLBN is a structure that reflects the causal relationships between nodes, an explanans must be an ancestor of an explanandum node.

**Definition 4-14: Explanans Space**

An Explanans Space, noted as $EN$, is a collection of all the explanans of a MLBN for a given situation.
Figure 4-5 shows the explanans space which covers A, B, and C, given that E and F are the explananda. Each explanans in the explanans space covers at least one explanandum in the explanandum space.

In their work about the logic of explanation, Hempel and Oppenheim (1948) defined *explanandum* and *explanans*:

“By the *explanandum*, we understand the sentence describing the phenomenon to be explained (not that phenomenon itself); by the *explanans*, the class of those sentences which are adduced to account for the phenomenon” (p. 152)

The definitions offered in this section formally define explanandum and explanans within the framework of MLBN.

In many cases, only one explanans is not sufficient to account for all the explananda, but the entire explanans space may be more than enough to justify the explananda. It is expected to find out subsets of the explanans space, where each subset can sufficiently explain all the explananda. Each subset will be called an explanation.

**Definition 4-15: Explanation**
Given an MLBN, an *Explanation* to an observation set \( O \) (composed of \( EV \) and \( EM \)) is a collection of MLBN nodes, noted as \( EX \), if it satisfies:

1. \( EX \subseteq EN \);
2. For any \( em \in EM \), there is an \( ex \in EX \), such that \( ex \) is a parent node of \( em \);
3. For any \( ex \in EX \), there is an \( ex' \in EX \), such that \( ex' \) is a parent node of \( ex \), or \( ex \) has no parent node.

An explanation is a subset of the explanans space, by which not only the explananda can be explained, but also each explanans can be potentially explained by other explanans. Figure 4-6 (a) shows an explanation that consists of nodes A and B. Not only are the explanandum nodes E and F explained by B and A, but also the explanans node B can be further explained by A. Figure 4-6 (b) shows that the explanans space could be another explanation, where A, B, and C are all used to explain E and F.

![Figure 4-6: Explanation](image)

An observation space and its explanation compose a story which describes a given situation. I formally define the notion of Story below.
**Definition 4-16: Story**

Given an MLBN, a Story, noted as $S$, is a triple $<EV, EM, EX>$ such that $EX$ is an explanation of $EM$.

Figure 4-7 depicts the relationship of observation space $O$ (consisting of the evidence space $EV$ and the explanandum space $EM$), the Explanation $EX$, and the story $S$.

---

Figure 4-7: Relations of the Components of a Story

Figure 4-8 shows two different stories built from the same observation space with different explanations.

---

Figure 4-8: Story
Figure 4-9 shows a sample Multi-Layer Bayesian Network with causal relationships between medical conditions. Each node in the graph represents a medical condition. The grey nodes represent observed information, the non-shaded grey nodes represent symptoms that need to be explained, and the shaded node represents a known fact that is a potential cause of shortness of breath.

Figure 4-9 describes the situation of a patient with id “123”, who has a shortness of breath with a “level 2 severity”. To explain the symptoms, posterior probabilities of the potential causes (i.e., hypotheses) are calculated using the suitable variable bindings in the observation (e.g., ?patient_id will be bound to “123”). The bindings of age-category and sex in the MLBN are determined from the information of patient “123” stored in the knowledge base.
4.3 The Methods

The HDSB framework has four major components: hypothesis generation, hypothesis evaluation, hypothesis-driven information seeking, and hypothesis revision. A hypothesis can be any set of unobserved nodes that are suspected to have specific observation values. The detailed mathematic and algorithmic methods for each component are discussed in this section.

4.3.1 Hypothesis Generation

In general, hypothesis generation is the process of generating hypothetical *causes* that are able to *explain* the *observed facts*, according to the *causal relationships*. In tandem with MLBN and the definitions earlier in this chapter, hypothesis generation is more specifically referred to the process of generating hypothetical *explanations* that are able to explain the *explananda* in *observations*, using the causal relationships and the variable binding constraints defined in a Multi-Layer Bayesian Network.

**Algorithm 4-1: GenerateEN(em)**

*Input:*

*em* : an explanandum

*Given:*

*net* : a MLBN;

*EM* : the explanandum space

*EV* : the evidence space

*Output:*

*EN* : the explanans nodes of the node *em*.

*Steps:*

1. \( vb = \text{GetVariableBinding}(EM, EV) \);
2. \( ancestors = \text{GetParentNodes}(em) \);
3. \( i = 0 \);
4. \( \text{while} \ (i < \text{ancestors.length}) \{
   \text{if} \ (\text{ancestor}[i] \ \text{has parent})
   \}
\)
Algorithm 4-1 describes the algorithm to obtain the set of explanans nodes for a given explanandum \( em \) by traversing the causal links in MLBN. A variable binding \( vb \) is first obtained from \( EM \) and \( EV \). Next the \( em \)’s parent nodes are added to the \( ancestors \) array. Then for each node in the \( ancestors \) array, its parent nodes are again added to the \( ancestors \) array, until all the ancestors nodes are included. Finally, for each node in the \( ancestors \) array, those nodes that have been observed with the initial variable binding or have repetitive occurrences are removed. Those nodes left in the \( ancestors \) array are the explanans nodes for \( em \). The explanans nodes for \( em \) only contain those unknown nodes which will be further set as hypotheses.

**Algorithm 4-2: GenerateEN(\( EM, EV \))**

Input:
- \( EM \): an explanandum space
- \( EV \): an evidence space

Given:
- \( net \): a MLBN;

Output:
- \( EN \): a collection of explanans;

Steps:
- for each \( em_i \) in \( EM \),
  - \( EN_i \)=GenerateEN(\( em_i \));
- end for
Algorithm 4-2 generates the explanans space for a set of explananda $EM$ by combining the explanans nodes from each explanandum in EM.

Algorithm 4-3 shows a recursive process for generating explanations for $EM$ and $EV$. For each explanandum $em$ in the $EM$, each of its parent nodes in MLBN will be considered a possible cause. For each parent node that is considered a cause of the $em$, becomes a temporary explanandum and needs further explanation. Therefore, this parent node is added to the $EM$, with the original $em$ temporarily removed from the MLBN and the $EM$. An iterative process with reduced explanandum space and MLBN is used for generating the partial explanations. Each explanation that combines a partial explanation with the parent node will compose a complete explanation for the original $EM$.

\begin{algorithm}
\hspace{1cm} \textbf{Algorithm 4-3: GenerateEXs}(EM, EV) \\
\textbf{Input:} \\
\hspace{2.5cm} EM: an explanandum space \\
\hspace{2.5cm} EV: an evidence space \\
\textbf{Given:} \\
\hspace{2.5cm} net: a MLBN; \\
\textbf{Output:} \\
\hspace{2.5cm} EXs: a collection of explanations. \\
\textbf{Steps:} \\
\hspace{2.5cm} EN = GenerateEN(EM, EV); \\
\hspace{2.5cm} EM' = copy EM; \\
\hspace{2.5cm} \textbf{for each} em in EM', \textbf{do} \\
\hspace{3.5cm} parents = em.parents; \\
\hspace{3.5cm} Remove em from EM'; \\
\hspace{2.5cm} \textbf{for each parent node}_i \text{ in parents, do} \\
\hspace{4.5cm} \text{Add node}_i \text{ into EM'}; \\
\hspace{4.5cm} \text{Delete em from net and EM'}; \\
\hspace{2.5cm} EXs' = GenerateEXs(EM', EV); \\
\hspace{2.5cm} \textbf{for each} EX_j \text{ in EXs', do} \\
\end{algorithm}
\[
EXs \leftarrow EX_j \cup \{\text{node}_i\}; \\
\text{end for} \\
\text{Remove repetitive occurrence of } EX \text{ in } EXs; \\
\text{end for} \\
\text{end for} \\
\text{Return } EXs;
\]

Stories can be built now with explanations generated from the observation space. Each story will be a combination of the observation space and one of its explanations. A story is a complete set of nodes with observation values, where each node is either explained by other nodes through the causal relationship in the MLBN or accepted as a true fact.

### 4.3.2 Hypothesis Evaluation

Among all explanations, some may be more plausible than others. The plausibility of an explanation is its posterior probability given the observation space, composed of the explanandum space (\(EM\)) and the evidence space (\(EV\)).

The nodes in an explanation space can be divided into two types: (1) the nodes that are accepted as the ultimate reasons and need no further explanations (Type-1 explanans), and (2) the nodes that are not accepted as the ultimate reasons and need further explanations (Type-2 explanans). A specific classification of nodes depends on the specific problem of a specific situation and the specific interest of the decision-maker. For example, in regular medical diagnosis situations, the nodes that represent diseases are usually considered the ultimate reasons for patient conditions. The ultimate goal of a diagnosis process is to find out the actual disease that can explain the observed symptoms.
However, in other scenarios such as an epidemiologic study, a decision-maker may need to further know what has caused the disease. Therefore, other factors such as environmental changes or personal contacts may be of greater interest to the decision-maker than the disease itself. Hence, diseases can become Type-2 explanans in these studies.

Another type of nodes that is important for story building is the nodes that represent further evidence that can be gathered. One of the key decisions during story building is to decide which piece of additional evidence needs to be gathered. In many domains (such as medical diagnosis), costs associated with the gathering of different evidence differ. Hence it is desirable to assist decision-makers in their choice of the additional evidence to gather.

In the HDSB framework based on MLBN, a Type-1 explanans is an MLBN node that represents a function with one or more arguments and one outcome. Given a variable binding, each Type-1 explanans has a posterior probability distribution on its outcome values.

In general, a Type-1 explanans that is more certain on one value will get more attention from decision-makers than those which are uncertain on each value. Intuitively, a Type-1 explanans whose probabilities of outcome values are more unevenly distributed should be ranked higher. This intuition is formally reflected by the measure of information entropy (Eq. 4-1)

\[
H(X) = -\sum_{i=0}^{n-1} p_i \log p_i
\]  

(4-1)
where \( p_i \) is the posterior probability of a node \( X \) to have the \( i \)-th outcome value given the observations (Eq. 4-2):

\[
p_i = p(X = o_i | EM, EV) \quad (4-2)
\]

Because an entropy plot looks like Figure 4-1, where the entropy value \((H(X))\) gets maximized when the probabilities are evenly distributed, I use the reciprocal of the entropy to rank the hypotheses.

For a hypothetical Type-1 explanans node that has \( n \) outcome values with a posterior probability distribution \(< p_0, p_1, \ldots, p_{n-1}>\), its ranking index (RI) can be calculated as in Eq. 4-3.

\[
RI(H) = 1 - H(X) \quad (4-3)
\]
Figure 4-11 depicts the plot of RI(H). The maximum value of the ranking index is 1 when one of the probabilities reaches 1.

This general measurement may not work perfectly to evaluate medical conditions, where the outcome values are usually composed of one normal value and one or more abnormal values. For example, a normal value could be *normal, negative, or false*; an abnormal value could be *high, low, positive, or true*. In a medical diagnosis, a physician will usually focus only on those conditions which have high probabilities on abnormal values rather than normal values. Based on this specific feature of medical conditions, two criteria are applied to determine the ranking of a hypothesis condition:

1. A condition that has lower probability on the normal value should be ranked higher;

2. A condition whose probabilities of abnormal values are more unevenly distributed should be ranked higher.

For a Type-1 explanans node that has n values (including 1 normal value and n-1 abnormal values), if its probability distribution over its values is \(< p_0, p_1, \ldots, p_{n-1} \rangle\), where
\( p_0 \) is the probability for the normal value, and \( p_1, \ldots, p_{n-1} \) are the probabilities for abnormal values, then the ranking index (RI) for each condition can be measured using Eq. 4-4.

\[
RI(H) = (1 - p_0)(1 - H(X))
\]  

(4-4)

**Figure 4-12** shows the plot of the revised RI(H). If \( p_0 \) is 1, the ranking index of the hypothesis will be 0; if \( p_0 \) is 0, the ranking value of the hypothesis will depend on how the probabilities are distributed among the abnormal values. If one \( p_i \) is 1, then the ranking index will be 1. Therefore, this formula successfully synthesizes the two criteria.

![Plot of Revised RI(H)](image)

Figure 4-12: Plot of Revised RI(H)

### 4.3.3 Hypothesis-Driven Information Seeking

The ultimate goal of story building is to construct a complete description of a situation, where the root causes are identified and used to explain the observed
phenomena. Because of the incompleteness of the initial information, story building is actually a process of approaching the true description of the situation with information being accumulated during the process. Some of the information may arrive automatically, such as a newly developed symptom in a patient. But other information may need to be requested, instead of being automatically observed, such as lab test results.

Information seeking is not a trivial activity but rather a challenging one, and a decision-maker needs to answer three questions during the information seeking process:

(1) What information is useful?

A piece of information is considered useful if it can help diagnose the situation. For example, in a patient care situation, if the D-dimer is 600 ng/mL, it is pretty safe to say that the patient has Pulmonary Embolism. Therefore, the information “D-dimer = 600 ng/mL” is very useful in making a diagnosis of Pulmonary Embolism. However, the same information is not useful for diagnosing pneumothorax, since the D-dimer level is not related to pneumothorax.

(2) How is the information collected?

Information can be obtained in different ways. It may be obtained from a database, a person, or even an activity that collects the needed information directly from the environment or the objects in the environment, such as clinical lab tests. Although the type of information to collect may have been decided, there is still uncertainty about what results will be returned for this type of information. In a patient care situation, for example, although it may be decided to administer a CT scan for a patient, the result is still uncertain before the outcome is returned. Therefore, it is also important to measure the usefulness of a test that may return different information.
(3) How is the resource use for collecting information prioritized?

Collecting the needed information is oftentimes constrained by resource availability. The resources for collecting information may be costly or tightly scheduled. In hospitals, for example, CT scans and MRIs are usually more expensive than other tests. Therefore, when selecting the resources for collecting a particular type of information, not only must the usefulness of the information needs be considered, but also the cost of the method for collecting the information.

This section aims to answer the above three questions by offering an quantitative mechanism for measuring the usefulness of a piece of information or an information seeking activity via an information-theoretical method. Within the framework of MLBN, a missing piece of information may be a Type-2 explanans or a node that can be potentially an explanandum node or an evidence node. I will not differentiate the types of the nodes, since the method remains same for all unobserved nodes.

Before describing the evaluation methods in detail, it is important to create a common understanding of two basic concepts.

**Definition 4-17: Information**

A piece of *information* is a sentence that states a fact about a n-ary relationship of n entities or values. In MLBN, a piece of information $I$ is a statement that states the fact that a node defined by a function $F$ is observed $o$ given variable binding $vb$: $F(vb) = o$.

For example, the sentence “The patient with ID-PA0001 has a D-dimer value of 500 ng/mL on 05/06/2010” states a fact about one feature of the patient. In predicate format, the sentence can be represented in “D-dimer(PA0001, 05/06/2010, 500 ng/mL)”,
which is a 3-ary relationship. In function format, the sentence can be represented in “D-dimer(PA0001, 05/06/2010) = 500 ng/mL”.

**Definition 4-18: Test**

A *test* is an activity that attempts to obtain the answer to a question by applying certain resources. The result of a test could be a set of pieces of information. ¶

For example, the D-dimer test aims to answer the question “What is the D-dimer value for the patient with ID-PA0001 on 05/06/2010?” by using lab resources. Therefore, information is the result of a test, and a test will return information.

A piece of information is considered useful if it can help diagnose a situation. In other words, a piece of information is considered useful if the probability distribution of a hypothesized Type-1 explanans on its values will be changed when the information changes from unknown to known. The probability distribution of a node is usually measured by the entropy, and therefore the change of entropy can be used to measure the value of information.

If the information $I$ states such a fact: $F_j(vb) = o_i$, where $F_j$ is the function defined on the $j$-th node of an MLBN, $vb$ is the current variable binding, and $o_i$ is the $i$-th value in $O_j$, and if the function defined on the hypothesis node $H$ is $F_H$, which can produce one of the $n$ possible values \{$h_0, h_1, ..., h_{n-1}$\} with arguments constrained by $vb$, then the entropy of the hypothesis node $H$ with knowing $I$ based on $vb$ can be represented as

$$H_I(F_H(vb)) = -\sum_{k=0}^{n-1} p_{rk} \log(p_{rk}),$$

where $p_{rk}$ is the posterior probability of the hypothesis node $H$ to be observed with $k$-th value $h_k$, given existing observation space $O$, and the information $I$: $p_{rk} = p(F_H(vb) = h_k \mid O, I)$. Similarly, the entropy of the hypothesis node
$H$ without knowing $I$ can be represented in $H_0(F_H(vb)) = -\sum_{k=0}^{n-1} p_{0k} \log(p_{0k})$, where $p_{0k}$ is the posterior probability of the hypothesis node $H$ to be observed with $k$-th value $h_k$, given existing observation space $O$: $p_{0k} = p(F_H(vb) = h_k | O)$.

**Definition 4-19: Value of Information**

The value of the information $I$ to the hypothesis $H$ is the reduction of entropy of $H$:

\[
\text{Value}(I, H) = H_I(F_H(vb)) - H_0(F_H(vb))
\]

\[
= -\sum_{k=0}^{n-1} p_{hk} \log(p_{hk}) - \sum_{k=0}^{n-1} p_{0k} \log(p_{0k})
\]

\[
= -\sum_{k=0}^{n-1} p(F_H(vb) = h_k | O, I) \log(p(F_H(vb) = h_k | O, I))
\]

\[
- \sum_{k=0}^{n-1} p(F_H(vb) = h_k | O) \log(p(F_H(vb) = h_k | O))
\]

\[
= -\sum_{k=0}^{n-1} p(F_H(vb) = h_k | EM, EV, I) \log(p(F_H(vb) = h_k | EM, EV, I))
\]

\[
- \sum_{k=0}^{n-1} p(F_H(vb) = h_k | EM, EV) \log(p(F_H(vb) = h_k | EM, EV))
\]

For a hypothesis, the higher information value indicates a greater change in the probability distribution of the hypothesis. Since information entropy is always greater than 0 and smaller than 1, the information gain will belong to $[-1, 1]$.

The value of information reflects its capability for changing the probability distribution of the values of a hypothesis. Therefore, it is important for a decision-maker to decide what information is good to know. However, it is more important to measure the value of a test, since that is the activity used to obtain the information.
Definition 4-20: Value of Test

The value of a test is the expected value of different information that the test can possibly obtain. If a test can have \( m \) possible results, then

\[
Value(t, H) = \sum_{i=0}^{m-1} p(t = I_i) \times Value(I_i, H)
\]

where \( p(t = I_i) \) is the probability of getting the \( i \)-th result, and \( Value(I_i, H) \) is the value of the information \( I_i \).

The probability of getting the \( i \)-th result for the test \( t \) is determined by the current variable binding \( vb \) and the current explanandum space and evidence space. Therefore, \( p(t = I_i) = p(F(vb) = o_i | EM, EV) \), which can be obtained by Algorithm 3-3.

When making the choice of which test should be administered, not only should the value of a test be considered, but also the cost of the test. Both the value and the cost contribute to the effectiveness of the test.

Definition 4-21: Effectiveness of Test

The effectiveness of a test is the value of test per cost unit:

\[
E(t) = \frac{Value(t, H)}{c}
\]

where \( c \) is the cost of the test.

In general, a decision-maker will prefer a test that yields greater value with less cost. Definition 4-16 offered one criterion for measuring the effectiveness of a test. There could be other criteria, depending on how much a decision-maker cares about the value and the cost of the test.
In some situations, more than one condition can be obtained from one test. For example, in a medical diagnosis, an ABG test (considered “large test”) can return results for pH, pO2, pCO2, and A-a gradient (each considered a “small test”). In those cases, the value of a large test $T$ (a group of small tests $i$) should be the value of the small test that has the maximal value among all tests in $T$:

$$\text{Value}(T, H) = \max_{j=0}^{n-1} \left( \sum_{i=0}^{m-1} p(t_j = I_i) \times \text{Value}(I_i, H) \right)$$

(4-8)

Similarly, the effectiveness of a large test $T$ should be:

$$\text{Value}(T, H) = \frac{\max_{j=0}^{n-1} \left( \sum_{i=0}^{m-1} p(t_j = I_i) \times \text{Value}(I_i, H) \right)}{C}$$

(4-9)

where $C$ is the cost of the large test.

### 4.3.4 Hypothesis Revision

Due to the uncertain and dynamic nature of decision-making environments, hypotheses should be revised accordingly to include new evidence. For instance, during a medical diagnosis process, if a patient develops a new symptom or a new lab result is returned, the physician should be able to revise the differential diagnoses to explain the new condition. He needs to decide whether this new symptom is relevant to a previous diagnosis. If so, how significant is it to the diagnosis? If not, does it indicate another disease that may have caused this new symptom?

Therefore, it is important to keep monitoring the environment for any changes, and just as important to differentiate the changes according to their relevance and
significance to previous hypotheses. In general, new information may have following effects on a previous hypothesis:

1. The new information is irrelevant to any previous hypothesis, and is caused by another factor. In this case, a new hypothesis should be developed to explain the new information.

2. The new information might be caused by an existing hypothesis or by a new factor. In this case, a new hypothesis should be made and the ranking of the hypotheses should be adjusted.

Particularly in medical diagnosis, each condition has three states: unknown, normal, and abnormal. Among these three states, there are five types of changes. Different strategies for hypothesis revision should be applied to these changes.

1. unknown to normal: After a test, a condition may change from unknown to normal. In this case, the new observation will not serve as an explanandum, but will serve as negative evidence. This negative evidence may remove the explanation path that may have assumed an abnormal value of the condition. Therefore, the hypothesis space will be reduced. Figure 4-13 shows a situation where a node changed from unknown to normal removes one hypothesis from the current hypothesis space. In the situation, \( H \) was at first the only explanandum. The two areas that cover \( \{A, D, H\} \) and \( \{B, E, H\} \) were two stories with two different explanations. If at a later point D was observed to be normal, the explanation that covers A and D was no longer the explanation. This hypothesis was then removed from the current hypothesis space.
2. unknown to abnormal: If a condition changes from unknown to abnormal, the new observation will serve as an explanandum and need to be explained. This new explanandum may require a new hypothesis. Therefore, the hypothesis space may be increased. Figure 4-14 shows a situation where a new explanation was added by the new observation of an abnormal value on node J. This new observation requires C and F to be the new explanations.

3. normal to abnormal: If a condition changes from the normal value to an abnormal value, it will need to be explained. It may require a new hypothesis.
to explain it. Therefore, the hypothesis space may be increased. For example, Figure 4-15 shows a situation where D was changed from the normal value to an abnormal value. The new explanation that covers A is added to the hypothesis space to explain D.

4. abnormal to normal: A previously abnormal condition changes to normal. In this case, this normal condition no longer needs to be explained. The previous hypothesis used to explain the abnormal condition may be cut. Therefore, the hypothesis space may be reduced. For example, as shown in Figure 4-15, node D may be changed from abnormal to normal. Therefore, H will only be explained by B and E. The area that covers A and D will no longer serve as an explanation of H.

5. abnormal to abnormal: A previously abnormal condition changes to another abnormal state. In this case, this condition still needs to be explained. However, the hypothesis that was used to explain the previous abnormal state may still explain the new abnormal state because there is still the probabilistic
causal relationship between the hypothesis node and the condition node. Therefore, the hypothesis space does not change. For example, as shown in Figure 4-15, if D changes from one abnormal value to another abnormal value, node A will still be the explanation for D, although its probability distribution will change.

Table 4-1 summarizes the revision strategy for the different types of changes of a condition. It shows that no matter how the hypothesis space is revised, the ranking of the hypotheses will always be adjusted, since any change of one condition will influence the probability distributions of the other nodes according to the probabilistic dependence.

<table>
<thead>
<tr>
<th>Change of a condition</th>
<th>Revise Space?</th>
<th>Revise Ranking?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown → Normal</td>
<td>Reduce</td>
<td>Yes</td>
</tr>
<tr>
<td>Unknown → Abnormal</td>
<td>Add</td>
<td>Yes</td>
</tr>
<tr>
<td>Normal → Abnormal</td>
<td>Add</td>
<td>Yes</td>
</tr>
<tr>
<td>Abnormal → Normal</td>
<td>Reduce</td>
<td>Yes</td>
</tr>
<tr>
<td>Abnormal → Abnormal</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Chapter 5
SYSTEM IMPLEMENTATION

5.1 System Architecture

The Hypothesis-Driven Story Building component is newly implemented based on R-CAST, an architecture of RPD-enabled Collaborative Agents for Simulating Teamwork (Fan et al., 2006; Fan & Yen, 2007). R-CAST is a collaborative agent architecture built on certain cognitive models—primarily the Recognition-Primed Decision-making (RPD) (Klein, 1989) model and Shared Mental Models (SMM) (Cannon-Bowers, Salas, & Converse, 1990). RPD describes how experienced decision-makers make decisions under time pressure in real situations. It stresses that human experts usually make decisions based on their past experiences by selecting an experience that worked previously in a similar situation, instead of calculating and comparing expected utility for each decision choice. SMM is a hypothetical cognitive construct that refers to a common understanding among team members regarding their objectives, roles, knowledge, and so forth. It attempts to explain many of the human behaviors in high performance teams.

The new version of R-CAST with HDSB feature is renamed SR-CAST. Figure 5-1 illustrates the architecture of an SR-CAST agent. It is composed of several functional modules and knowledge repositories. The SR-CAST modules include Decision-Making Manager, Story-Building Manager, Knowledge Base Manager, MLBN Manager, Information Manager, Communication Manager, and Process Manager. These modules
play important roles in SR-CAST. They cover the critical functions of the architecture: decision-making, story building, logical knowledge inference, probabilistic knowledge inference, identifying missing information, and communication. The SR-CAST knowledge repositories include *Knowledge Base, Experience Base, Plan Library*, and *MLBN*. They store different types of knowledge. There are control flows between modules (as marked with black arrows) and data flows between modules and repositories (as marked with blue arrows).

Each of the functional modules plays a particular role in the SR-CAST agent.

1) *Decision-Making Manager*

The Decision-Making Manager is the module for making RPD-style decisions. As
described in the original RPD model, it compares the current situation (described in the Knowledge Base) with the past experiences (saved in the Experience Base), and selects an experience that is similar enough to the current situation. The course of actions that were involved in the selected experience will be sent to the Process Manager for execution.

2) Story Building Manager

The Story Building Manager implements the major steps of HDSB. After being triggered by the Decision-Making Manager, it starts the process of generating hypotheses, evaluating hypotheses, prioritizing information seeking if there is missing information, and revising hypotheses if there are any significant changes of information. The result will be used by the Decision-Making Manager for RPD-style decision-making.

3) Knowledge Base Manager

The Knowledge Base Manager manages the knowledge base and makes forward-chaining inferences when new information is observed. It deals with the logical relationships among predicates. New information can be implied based on the logic rules and the inference methods, as well as the variable binding relationship among predicates.

4) MLBN Manager

The MLBN Manager manages the Multi-Layer Bayesian Network and makes probabilistic inferences. It is triggered by the Story Building Manager to provide the posterior probability of a predicate, given other information that has already been known. It gets information from the Knowledge Base as evidence before making probabilistic
inferences. The results will be returned to the Story Building Manager.

5) Information Manager:

The Information Manager identifies the missing information and specifies the information sources for the missing information. A request message will be sent to the Communication Manager when a particular piece of information is needed. The returned information will be written back to the Knowledge Base.

6) Communication Manager:

The Communication Manager is responsible for exchanging information with other agents, databases, data collection equipments, or user interfaces for human user to input. Once it receives the queried information, it will send the information back to the Information Manager.

7) Process Manager:

The Process Manager executes the corresponding course of actions that are decided by the Decision-Making Manager. It will access the Plan Library for the predefined steps. In many applications, such as clinical decision support systems, these actions include making certain kind of recommendations to human users.

At the beginning of each decision-making cycle, when new information is captured by the Observer, the Knowledge Base Manager and the Decision-Making Manager are triggered. The Knowledge Base Manager will check the new information against the current content of the Knowledge Base and update the Knowledge Base.
according to the new information. (The Knowledge Base is always kept updated.) The Decision-Making Manager will assess whether or not the current information can be matched to a past experience. If a similar situation is matched in the Experience Base, the Decision-Making Manager will identify what actions were taken and request the Process Manager to execute the same actions. Otherwise, if a situation is not similar enough to any past experience, the Decision-Making Manager will request the Story Building Manager to derive the most plausible explanation of the current information.

Then the newly implemented Story Building Manager generates the hypothetical information that may explain the current information by analyzing the structure of the logical inferential knowledge and the Bayesian network. Then the Story Building Manager will request the probability for each hypothesis from the MLBN Manager which, in turn, will calculate the posterior probability for the hypothesis given the available information in the Knowledge Base. When all the probabilities of the hypotheses are returned to the Story Building Manager, it will evaluate the hypotheses, and identify the value of each activity for seeking particular pieces of information. If it is finally determined to locate a missing piece of information, the Story Building Manager will request the Information Manager to identify an information source. The Information Manager then requests the Communication Manager to query the particular missing piece of information from identified information sources.

When new information arrives or is returned by the Communication Manager, both the Decision-Making Manager and the Story Building Manager will check whether or not the new information is significant enough to revise the previous decision or hypotheses. If it is, then the Story Building Manager will start a hypothesis revision
process. The revision will affect the decision-making process of the Decision-Making Manager. This “matching – hypothesizing – seeking – matching” process runs iteratively until a decision is made with relative certainty.

5.2 Knowledge Repositories

The activities performed by each module rely on knowledge that has been predefined for the agent. In SR-CAST, there are four different knowledge repositories for storing four different types of knowledge: logical inferential knowledge, probabilistic inferential knowledge, experiential knowledge, and procedural knowledge. We identify the bases as follows.

1) Knowledge Base

The Knowledge Base stores the logical inferential knowledge. It is comprised of three major logical components: FactType, Rule, and Fact.

A FactType defines the types of facts that the agent is able to understand. It regulates the format of each single fact. Any fact that is not written according to known FactType parameters will not be recognized by agent. The definition of FactType follows the predicate format, which describes an n-ary relationship of objects. A FactType definition primarily includes name and arguments.

A Rule signifies the logical inference relationship between the premises and the consequences. If all the premises are satisfied, the consequences will be true.
A Fact defines each single fact of the current situation. Each fact is actually a predicate. In each fact, all the arguments have been substituted with particular entities, objects, or values.

2) \textit{MLBN}:

The MLBN stores knowledge that represents probabilistic causal relations between predicates. It is composed of definitions for nodes and their causal relationships, written in XML format.

A node definition includes its name, arguments, outcome, and position. The name and the arguments must be consistent with the FactType definition in the Knowledge Base; otherwise, data exchange between the Knowledge Base and the MLBN will be impossible. The outcome identifies the possible values of the node, such as “true” or “false” and “normal,” “low,” or “high.”

A causal relationship is defined for each node. For those nodes which have no parent nodes, a table of prior probabilities is given. For those nodes with parent nodes, a table of conditional probabilities with all combinations of conditions is given. Figure 5-2 shows the syntax of defining the causal relationship between a parent node PTX and a child node SubEm.

\begin{verbatim}
<DEFINITION>
  <FOR>SubEm</FOR>
  <GIVEN>PTX</GIVEN>
  <TABLES>
    <TABLE BINDINGS="DEFAULT">0.9 0.1 0.0 1.0</TABLE>
  </TABLES>
</DEFINITION>
\end{verbatim}

Figure 5-2: Syntax of Causal Relationship Definition
Because of the variable-binding constraints, more than one table may be defined for a node. For example, if the causal relationship between the parent node PTX and the child node SubEm depends on the age of the patient, then one more table might be added to the definition of the causal relationship. Figure 5-3 shows the multiple-layer definition of a causal relationship.

```
<DEFINITION>
  <FOR>SubEm</FOR>
  <GIVEN>PTX</GIVEN>
  <TABLES>
    <TABLE BINDINGS="DEFAULT">0.9 0.1 0.0 1.0</TABLE>
    <TABLE BINDINGS="?age=old">0.99 0.01 0.0 1.0</TABLE>
  </TABLES>
</DEFINITION>
```

Figure 5-3: A Multi-Layer Definition of Causal Relationship

3) Experience Base

The Experience Base contains experiences that reflect how experts take actions toward different situations. Each experience is basically a mapping from the description of a situation to a course of actions.

4) Plan Library

The Plan Library contains a collection of plans which define the steps of a procedure. A plan is comprised of a sequence of actions, each of which can be either an elementary action or another plan.
5.3 SR-CAST HDSB Interface

**Figure 5-4** shows the HDSB interface of an SR-CAST agent. The interface can be used to edit the Multi-Layer Bayesian Network as well as to display its current status.

The left side of the interface displays the structure of a Multi-Layer Bayesian Network. Each node is a predicate that represents a patient condition in a particular case. The attributes of a node can be checked and edited through the editing window (**Figure 5-5**), where we can see and edit the Name, Arguments, and Values of a predicate. For example, **Figure 5-5** represents the predicate Hypovolemia(?id, ?Hypovolemia), whose value could be true or false. After clicking “Apply”, this information will be saved in an XML format.

Nodes are connected by edges with probabilities, which represent the probabilistic causal relationships between them. **Figure 5-6** is the editing window of the conditional probabilities between its parent nodes and the child node. For example, as shown in this
figure, the value of a predicate CO is co-affected by the values of HR and StrokeVolume. One can also identify the conditional probabilities for a particular variable binding. After clicking “Apply”, the conditional probability tables will be saved in the XML format that is defined for the MLBN syntax.

Figure 5-5: Editing Window for a Predicate Node

Figure 5-6: Editing Window for Conditional Probabilities
On the right side of Figure 5-4 is a listing of hypotheses, ranked in a specific order. The criteria for ranking them are:

1. The condition which has a lower probability on the normal (false, negative, ...) value should be ranked higher. This is because physicians will care more about the patient’s abnormal condition;

2. If there is more than one abnormal (true, positive, and other non-normal) value, the condition whose probabilities on the abnormal values are more evenly distributed should be ranked higher. This is because physicians will pay more attention to the condition that is highly probable.

For example, based on the current situation, the node InsuffAnesth (Figure 5-4) is ranked the highest, because it has the highest probability on its true value.

Each node may have one of four states in SR-CAST: unknown, observed, hypothesized, and recommended.

An unknown node is a one whose value is not yet known. It is colored in grey. An unknown node can potentially change to another other state.

An observed node is a one whose value is already known. It is colored in blue. An observed node will not be considered as hypothesized or recommended in this analysis.

A hypothesized node is a one that is considered the hypothetical cause of all the observed nodes. Since the MLBN reflects causal relationships, only the top nodes will be considered for the hypothesized nodes. A hypothesized node is colored in red.

A recommended node is a one that is recommended to observe based on the hypothesized node. The recommended node is colored in orange.
The process by which the HDSB component works in a multi-agent environment is as follows.

There are two agents involved in a given setting. Agent_A plays the role of the story building manager, and Agent_B plays the role of the information provider. Agent_A has a predefined knowledge base, an MLBN, an experience base, and a plan library. Agent_B has a complete knowledge base.

Both agents are activated as the system starts. Agent_A checks its knowledge base to see which nodes have been observed and colors them blue. Then the system lists the nodes which are considered the hypotheses along the right side of the display panel, using the ranking criteria. To obtain more information about a suspected node, the user simply clicks on that node in the list. The corresponding node in the network will be colored red, indicating that the node is currently hypothesized. Simultaneously, another node will be colored orange, indicating that this node contains the information most valuable to know (recommended) to further confirm or exclude the hypothesized node.

When any node is clicked, a pop-up menu will appear. To learn the current probability distribution of a node, the user simply accesses the “Probability Distribution” option, and a listing of all relevant values will be shown in the lower right area of the window. To learn the actual value of the node, the “Order” option is accessed. The agent may then try to identify the sources which might have more information for the node. In this case, Agent_A sends a request message to Agent_B and queries the value of the accessed node. Agent_A receives the query results back from Agent_B, and inserts them into its knowledge base. This changes the state of the accessed node and updates the hypotheses list. The user may also ask the system to recommend another node that will
generate the most valuable knowledge for the next step by selecting the “Recommend Test” option. Finally, when the user is certain about the hypothesis, the diagnosis can be made simply by accessing the “Diagnose” option. The HDSB process is then ended.

5.4 SRCAST-Diagnosis: A Web-based Clinical Diagnostic Decision Support Tool

Based on SR-CAST, a web-based clinical diagnostic decision support tool SRCAST-Diagnosis has been developed. It allows a user to access the tool from a web browser. For users who are not familiar with MLBN and probabilistic causal relationships between nodes, the tool shows only the information and recommendation, but hides the process of producing the recommendations at the back end.

Figure 5-7 depicts the architecture of SRCAST-Diagnosis. When the user initiates the webpage, a request is sent to the servlet that runs on the remote server. The servlet then starts and maintains an SR-CAST agent instance. The agent will then request and receive data from the data sources. During the decision support process, requests and responses are sent and returned one layer at a time, layer by layer.

Figure 5-7: SRCAST-Diagnosis Architecture
Figure 5-8: SRCAST-Diagnosis Interface

### Diagnosis Decision Support

**00:00:37**

**Patient ID:** PA01, 74 female

**Patient Complaint:**
A 74-year-old female arrives at the Emergency Department, with shortness of breath, a 3 day history of cough, productive of yellow sputum. She notes occasional sharp, stabbing chest pain, and is slightly cyanotic.

#### Patient Conditions

<table>
<thead>
<tr>
<th>Symptoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cough; Shortness of breath; Chest pain; Cyanosis;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Physical Examination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current vital signs:</td>
</tr>
<tr>
<td>Body temperature = 38.1 C [Low Fever]; Blood pressure = 128 mmHg [Normal]; Heart rate = 100 times/min [Normal]; Respiratory Rate = 28 times/min [High]; Saturated of Oxygen = 92 % [Low];</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Physical Exam Results:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rales; Rhonchi;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lab Results</th>
</tr>
</thead>
</table>

#### Differential Diagnoses (Agent Recommendations)

Please select your tentative diagnosis:
- Pulmonary Embolism [77%]

Supporting Evidence for Agent Recommendation
- Cough = Present;
- Dyspnea = Present;
- Chest Pain = Present;
- Cyanosis = Present;
- BMI = Low-Fever;
- BP = Normal;
- HR = Normal;
- RR = High;
- SaO2 = Low;
- Rales = Present;

**Agent Recommended Lab Tests**

- D-Dimer: Increased - Pulmonary Embolism = 100%, Normal - Pulmonary Embolism = 0%

- VQScan: Positive - Pulmonary Embolism = 100%, Negative - Pulmonary Embolism = 26%

- Ultrasound: [cost: $542]
- CT: [cost: $835]
- ABG: [cost: $58]
- BNP: [cost: $79]
- PeakFlow: [cost: $87]
- Cardiac Panel: [cost: $126]
- Basic Metabolic Panel: [cost: $25]
- X-ray: [cost: $117]
- EKG: [cost: $89]
- CBC: [cost: $50]

### Please Select Your Final Diagnosis

- Viral Pneumonia
- Pulmonary Embolism
- Cardiogenic Pulmonary Edema
- Asthma
- Acute Bronchitis
- Pneumothorax
- Myocardial Infarction
- Dissecting Thoracic Aneurysm

[Continue diagnosis]
Figure 5-8 shows a snapshot of SRCAST-Diagnosis. The top of the webpage shows the time, the patient id, and the patient’s initial complaints. The major component of the webpage is a table that consists of patient conditions, differential diagnoses recommended by the agent, and a list of diagnoses for the user to select.

The patient conditions include symptoms, physical examinations, and lab test results. They are normally displayed in blue. However, if a piece of information indicates a significant change of the patient condition, which will be displayed in red.

The second column shows the differential diagnoses, which are ranked from higher probability to lower probability. If the user accesses a listed diagnosis, more information can be seen. This will include supporting evidence that explains how the diagnosis is derived, consistent with patient conditions listed in the left column. Next a list of lab tests useful for further diagnosis may be accessed. These are also ranked based on their usefulness in further differentiating the differential diagnoses. Accessing a recommended lab test will generate a pop-up window with an explanation of how useful this test is. It lists the possible outcomes of the suspected diagnosis for all the possible values returned from the test. The test that has higher information gain will be highest on the list.

Based on this information, a user can select and order the lab tests that are the most useful. When the lab results are returned, and the agent recommendations about differential diagnoses and their corresponding lab tests will be updated.
Chapter 6

EXPERIMENT DESIGN

6.1 Experiment Objectives

The new HDSB framework has been added to R-CAST architecture and configured as a prototype for supporting medical decision-making. However, it is yet unknown whether or not the prototype can actually help a human decision-maker to improve a decision-making performance. Therefore, this chapter investigates the design of an experiment to evaluate the usefulness of the tool. In particular, it is expected to demonstrate: (1) whether the system can reduce wrong decisions; and (2) whether the system can help reduce the resource use for information gathering without compromising decision quality.

To evaluate the HDSB framework and answer the questions, I chose to evaluate the tool in the medical domain of clinical diagnosis. In a typical medical environment, a physician is required to make a diagnostic decision about a patient’s condition. The physician oftentimes faces a situation where patient symptoms and lab results arrive only partially during the diagnostic process, where hospital resources are expensive and tightly scheduled, and where patient conditions may constantly change over time. All of these factors make diagnostic decision-making difficult and a decision support tool necessary. In particular, the experiment will answer: (1) whether the system can reduce misdiagnoses; and (2) whether the system can reduce redundant use of lab resources.
6.2 Experiment Hypotheses

The experiment is expected to show an improvement in a decision-makers’ performance when using the decision support tool in the following ways.

(1) When supported by the tool, a decision-maker is expected to reduce the possibility of making a wrong decision. In the medical experiment, it is expected that the rate of misdiagnosis will be reduced. Partial and incomplete information is one of the key reasons why a misdiagnosis is made. Therefore, by helping generate and rank differential diagnoses based on partial information using quantitative probabilistic reasoning, the system should help a decision-maker correctly identify the diagnosis.

(2) When supported by the tool, a decision-maker is expected to optimize the resource use for information seeking. Particularly, a decision-maker supported by the system is expected to expend fewer lab resources for collecting information about patient conditions. Without decision support, a clinical decision-maker may order unnecessary lab tests while uncertain of the accurate usefulness of each lab test to the hypotheses. Therefore, by recommending lab tests based on their usefulness in differentiating differential diagnosis, the system should help a decision-maker to use the information seeking resources more efficiently.

(3) When supported by the tool, a decision-maker is expected to be more likely to change a suspected hypothesis to the correct one with fewer resources. The decision-maker’s biases that tend to stick to previous information and
previous hypotheses may be overcome due to the recommendations provided by the system, which may lead to a quicker identification of the correct hypothesis.

6.3 Experiment Design

To evaluate the system, a controlled experiment was conducted. Clinical decision-makers were recruited and divided into a control group and an experimental group.

Control Group and Experimental Group

A control group in an experiment is the group for which no changes are made. Therefore, results from control group are usually considered the “normal” or “base” results. An experimental group is the group to which specific conditions have been applied. Therefore, the results from the experimental group are expected to differ from those of the control group, if these conditions are true factors.

In this experiment, the experimental group was a group of medical decision-makers whose decision-making processes were supported by the HDSB decision support prototype. The tool (Error! Not a valid link.) helped each participant go through the clinical decision-making process. Each participant was presented with ten patient scenarios. At the beginning of each scenario, the participant received the initial symptoms and physical exam results from a patient. During the course of the scenario, the participant was asked to select a suspected diagnosis and order labs from a set of available lab tests. Once the participant believed that a diagnosis was certain, he submitted his decision by selecting the diagnosis listed.
The decision support system was intended to help the participant in the following ways:

(1) rank the differential diagnoses based on their probabilities (posterior probability);
(2) rank the lab tests based on their usefulness to confirm or exclude a suspected diagnosis;
(3) adjust the ranking automatically when new but significant information arrived.

Each participant was able to use the information and recommendations offered by the tool to help him through the decision-making process. However, as the clinical decision-maker, each was able to choose to make a decision based on self judgment.

Because of the emergent condition of the patients, participants were required to make the final diagnosis as accurately as possible, and to try to minimize the time expense and optimize their resource use.

The **control group** in the experiment was a group of medical decision-makers whose decisions completely depended on their own judgment. Each participant was presented the same ten patient diagnosis scenarios, and viewed same initial symptoms and physical exam results. They were asked to take the same actions but without recommendations from the decision support tool. Participants from the control group used the following interface (Figure 6-1) where only the decision-making situation was presented without agent recommendations.
Independent Variables and Dependent Variables

In experiment design, independent variables and dependant variables are the two critical elements. Their relationship is studied through the experiment. An independent variable is one that can be changed or controlled independently from other factors. A dependent variable is a one that will change accordingly based on the change of the independent variable. By controlling the change of independent variables and observing the changes of dependent variables during the experiment, the relationship between them can be tested.

The independent variables of the experiment include:
(1) Experiment group vs. control group: The experiment tested whether medical decision-makers had different outcomes with or without the support from the tool.

(2) Single-diagnosis scenario vs. double-diagnosis scenario: The experiment tested whether the improvement brought by the tool differed between cases with one diagnosis and those with two diagnoses.

(3) Role of participants (nurses, residents, and physicians): The experiment tested whether different role players had different levels of improvements due to use of the tool.

The dependent variables of the experiment included:

(1) Diagnosis Accuracy: This is measured as the rate of misdiagnosis. There are two ways to measure the rate. From the participants’ perspectives, each participant can be measured by the number of misdiagnosed scenarios out of the ten scenarios. From the scenarios’ perspective, each scenario can be measured by the number of participants who misdiagnosed the scenario.

(2) Resource efficiency: In each scenario, the total cost of lab resources that a participant has ordered before the final diagnosis is made was recorded.

(3) Relationship with cognitive biases: In the experiment, two cognitive biases (anchoring heuristics and confirmation biases) were particularly focused on. The anchoring heuristics was measured by the percentage of participants who finally changed their diagnosis to the correct one from initial incorrect diagnosis. The confirmation biases were measured by the accumulative cost
of resources spent when a diagnosis switched from an incorrect one to the correct one.

A MySQL database was built to record the data from each participant.

6.4 Experiment Process

6.4.1 Knowledge Acquisition

In SR-CAST, the hypothesis reasoning of HDSB is conducted based on the MLBN. Therefore, a critical step in the process is to build the MLBN before the system can be evaluated. Building a medical MLBN followed these requirements:

1) The MLBN should reflect some of the common problems of ED patients, so that the average ED decision-maker can capably make a diagnosis.

2) The MLBN should cover problems of ED patients that may have common symptoms, common physical exam results, and common lab results, since helping discriminate the differential diagnoses is the value of the HDSB-enabled decision support tool.

3) For the purpose of experiment, the MLBN should cover a controlled number of diseases and their related symptoms, physical exam results, and lab results. Although building a complete MLBN that covers all patient diseases is desirable, it is infeasible and unnecessary for the experimental study.

4) For each disease listed, the MLBN should also include:
   a) symptoms related to the problem;
b) related physical examinations and their possible results;

c) related lab tests and their possible results.

(5) The MLBN should reflect the causal relationships and the probabilities between nodes that represent diseases, symptoms, physical exam results, and lab test results.

Following consultation with a physician at Hershey Medical Center, the following ten diseases were selected for representation in the MLBN:

(1) Viral pneumonia (VP)

(2) Bacterial pneumonia (BaPn)

(3) Pulmonary Embolism (PE)

(4) Cardiogenic pulmonary edema (CPE)

(5) Asthma

(6) Pneumothorax (PTX)

(7) Myocardial Infarction (MI)

(8) Acute Bronchitis (AB)

(9) Dissecting thoracic aneurysm (DTA)

(10) Aortic Stenosis (AS)

Based on this selection, I studied the related medical literature and developed a document that organized the relevant content for each disease. For instance, Appendix Error! Not a valid link. shows the content of viral pneumonia. It includes Name, Description, Risk Factors, Possible Symptoms, Physical Examinations, and Lab Tests.

The ten diseases and their relevant symptoms, physical examinations, and lab tests are summarized in three tables (Appendix B.1, B.2, and B.3).
Clinically, a *symptom* is a departure from normal function, sensation, or appearance which is noticed by a patient, indicating the presence of disease or disorder. It is subjective, observed by the patient, and not measured (Devroede, 1992). Symptoms are usually the initial concern that brings a patient to a doctor.

*Physical examination* is the process by which a doctor examines the body of a patient for signs of disease. Unlike symptoms which can be experienced by the patient, a physical examination is conducted by a professional and will disclose signs that are not directly experienced by the patient. Physical examination usually covers vital signs and chest exams for a patient presenting chest pain and shortness of breath.

The *vital signs* of a patient typically consist of body temperature, blood pressure, heart rate, respiratory rate, and saturation of oxygen. Different diseases will have different effects on vital signs. Although there are criteria for classifying patients at different levels of severity, the following are commonly accepted:

*Body Temperature:*

Lower than normal: $< 36.1 \, ^\circ C$

Normal: $36.1 \, ^\circ C – 37.5 \, ^\circ C$

Lower fever: $37.5 \, ^\circ C – 39.5 \, ^\circ C$

High fever: $> 39.5 \, ^\circ C$

*Blood Pressure (systolic):*

Low: $< 90$

Normal: 90-140

High: $> 140$

*Heart Rate:*


Low: < 50
Normal: 50 – 110
High: > 110

Respiratory Rate:
Low: < 6
Normal: 6 – 20
High: > 20

Saturation of Oxygen:
Very low: < 92%
Low: 92 – 97%
Normal: > 97%

A chest exam covers the cardiovascular system and the respiratory system. There are four typical ways to conduct the chest exam:

Perception: the process by which a doctor observes a patient from his appearance. For example, the doctor may have observed that the patient has rapid breathing.

Auscultation: the process by which a doctor listens to the internal sounds of the patient’s body, usually using stethoscope. For example, in a patient with viral pneumonia, rales or other abnormal breathing sounds may be heard.

Palpation: the process by which a doctor touches a patient’s body to determine an object’s size, shape, firmness, and location.

Percussion: the method by which a doctor taps on the surface of a patient’s body to determine the underlying structure. A particular sound can be heard with percussion,
such as resonant, hyper-resonant, dull, and stony dull. For example, from a patient with viral pneumonia, a dull sound may be heard.

A medical lab test is a medical procedure performed to help a doctor to detect, diagnosis, or evaluate a disease, usually in a physical, chemical, biological, or radiological way using advanced technologies and equipment. The lab tests that are typically applicable to diagnose chest diseases include:

**Arterial Blood Gas (ABG):** typically consists of the values of pH, pO2, pCO2, and A-a Gradient.

**Complete Blood Count (CBC):** typically consists of White Blood Count (CBC), Hemoglobin, Hematocrit, Platelet Count.

**Basic Metabolic Panel:** returns values of Blood Sodium, Potassium, Chloride, Bicarbonate, BUN, and Creatinine.

**Cardiac Panel:** checks the values of CK-MB fraction, and Troponin.

Other labs:

**C-reactive Protein**

**D-Dimer**

**BNP**

**Chest X-Ray**

**CT Chest**

**EKG**

**Echocardiogram**

**V/Q Scan**

**Ultrasound on Lower Extremities**
The causal relationship between a disease and a symptom (or a physical examination result, or a lab test result) is represented by a number, which is the conditional probability of the symptom (or the physical examination result, or the lab test result) to be true (or another possible value), given that a patient has the disease. A higher probability indicates a stronger causal binding.

The probabilities were obtained in three ways:

1. Literature study

Some of the causal relationships have been studied in previous medical research. These probabilities can be obtained from medical literature. For example, it was reported that in patients with recognized pulmonary embolism, 34% were found to have S3 and S4 gallop (Sutherland, 2009). Those probabilities that can be obtained from the literature are included in the conditional probability table.

2. Expert estimate

Most of the conditional probabilities are not available in medical literature. However, they can be estimated by experienced physicians. An experienced physician may have seen thousands of patients and may have mentally built a network of causal relationships. A knowledge elicitation process was conducted with three experienced physicians at Hershey Medical Center and at Mount Nittany Medical Center. Each was asked to estimate the probabilities for each causal relationship. The results for average probability were calculated and are included in the probability table.

3. Data Mining
A more scientific and accurate way to obtain probability is to retrieve it from an existing database, such as the Electronic Medical Record (EMR) system. With the rapid development of computer technologies, more patient data are saved in electronic media than in paper media. At the Hershey Medical Center, a data warehouse (PowerInsight) has been established to hold patient data since 2000. Due to limitations on access authorization, probabilities from the system were not obtainable. However, securing access to a database may become an important way to build the solid knowledge base of a decision support system.

In the experiment, the total expense and the time that a participant spent on ordering lab tests was measured. Table 6-1 lists the money and time costs for each lab test. The numbers in the table may vary by hospital. However, they reflect the levels of cost.
The MLBN was built based on the results from knowledge elicitation and integrated into the system.

6.4.2 Scenario Design

A *scenario* is a postulated situation that outlines a sequence of possible events. In this experiment, each scenario represented a patient case where the patient conditions were predefined and presented to the decision-maker at specific times based on their interactions with the scenario. When designing scenarios, I particularly followed four criteria in order to test the relationship between the independent variables and dependent variables:

1. Each scenario describes one patient case.

<table>
<thead>
<tr>
<th>Lab Test</th>
<th>Cost of Money ($)</th>
<th>Cost of Time (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABG</td>
<td>85</td>
<td>20</td>
</tr>
<tr>
<td>CBC</td>
<td>36</td>
<td>15</td>
</tr>
<tr>
<td>Basic Metabolic Panel</td>
<td>25</td>
<td>15</td>
</tr>
<tr>
<td>CK-MB Fraction</td>
<td>57</td>
<td>15</td>
</tr>
<tr>
<td>Troponin</td>
<td>69</td>
<td>15</td>
</tr>
<tr>
<td>C-Reactive Protein</td>
<td>65</td>
<td>15</td>
</tr>
<tr>
<td>D-Dimer</td>
<td>90</td>
<td>15</td>
</tr>
<tr>
<td>BNP</td>
<td>79</td>
<td>15</td>
</tr>
<tr>
<td>Chest X-Ray</td>
<td>117</td>
<td>10</td>
</tr>
<tr>
<td>CT Chest</td>
<td>835</td>
<td>20</td>
</tr>
<tr>
<td>EKG</td>
<td>89</td>
<td>10</td>
</tr>
<tr>
<td>Echocardiogram</td>
<td>849</td>
<td>20</td>
</tr>
<tr>
<td>V/Q Scan</td>
<td>638</td>
<td>12</td>
</tr>
<tr>
<td>Ultrasound on lower extremities</td>
<td>342</td>
<td>20</td>
</tr>
<tr>
<td>Peak Flow</td>
<td>87</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 6-1: Price and Time of Lab Tests
2. Symptoms do not directly disclose the underlying diseases.

3. Each scenario contains the value for each possible test.

4. Some scenarios include symptoms that are developed at a later time.

Each scenario is defined with a specific xml format.

A scenario is composed of the following components:

1. **Initial description**

   The description is a short sentence that describes the initial conditions of a patient. It highlights the chief complaints that have brought the patient to the hospital. For instance, an initial description of a patient case can be:

   ```xml
   <description>
   A 65 y/o male patient presents to the Emergency Department with the complaint of progressive breathlessness, increased with exertion for last 3 days. He also has complaints of cough with sputum, chest pain, and diaphoresis. </description>
   ```

2. **Symptoms**

   Symptoms are defined by name, value, and appearing time. Because of the dynamic feature of patient conditions, symptoms may appear at different times. For example, a symptom may be defined as:

   ```xml
   <symptom>
   <name>Dyspnea</name>
   <value>Present</value>
   <appear_time>0</appear_time>
   </symptom>
   ```

3. **Physical Examination Results**

   Physical examination results consist of vital signs and other observations. Vital signs are usually ordered by default for every patient in Emergency Department, and may be repeated periodically. The scenario should specify the value for each vital sign at each
specific time. For example, the definition of vital signs of a specific time may look like this:

```xml
<vitalsign>
  <time>0</time>
  <BT>37.0</BT>
  <BP>100</BP>
  <HR>100</HR>
  <RR>20</RR>
  <SaO2>95</SaO2>
</vitalsign>
```

Other physical examination results are listed with the vital signs, and may be presented to the user when the scenario starts. For example, a physical examination result may appear as

```xml
<exam>
  <value>Rales</value>
  <obs>Bilateral diffuse rales/crackles</obs>
</exam>
```

where `value` indicates the value for computer processing, and `obs` explains the meaning of the value to the participants.

4. Lab Test Results

Lab test results contain the values for all lab tests. For example, the *ABG* results can be defined as follows:

```xml
<testresult>
  <name>ABG</name>
  <result>
    <name>pH</name>
    <value>7.2</value>
  </result>
  <result>
    <name>pO2</name>
    <value>70</value>
  </result>
  <result>
    <name>pCO2</name>
    <value>48</value>
  </result>
  <result>
    <name>Aa</name>
```

5. Diagnoses

Each scenario may have one or two diagnoses (i.e., one or two diseases that have caused the symptoms, physical examination results, and lab test results.) This section defines the true diseases that the patient is having.

6. Disposition

This entry defines the disposition of the patient after he is correctly diagnosed. For example, a disposition of a patient could be:

<disposition>Patient was admitted and treated. </disposition>

7. Consequence

This element identifies the consequence of a patient left untreated in the worst case scenario (if he is misdiagnosed or delayed in diagnosis.) The most severe consequence is death.

For the experiment, 12 scenarios were designed, 2 for training and 10 for testing. Among the 10 testing scenarios, there are 6 with a single diagnosis, and 4 with double diagnoses. These scenarios were designed and verified by experienced physicians at the Hershey Medical Center.

6.4.3 Participant Recruitment

SRCAST-Diagnosis can be evaluated by comparing the performances of participants from the experimental group and the control group. To enhance its evaluation,
participants were required to have a certain level of medical knowledge so that they were able to understand the clinical diagnosis process and the recommendations it provided.

Based on this criterion, the following groups of healthcare practitioners were eligible to participate in the experiment:

1. Nurses
2. Residents
3. Physicians

A recruitment letter was distributed to potential participants in the Emergency Department at Hershey Medical Center. It stated the purpose and the procedure of the experiment, and indicated the compensation for a participating participant. It also stated that participation was voluntary, and that refusal to take part in or withdrawing from the experiment would involve no penalty or loss of benefits that a participant would receive otherwise. Approximately 40 participants were scheduled for recruitment.

6.4.4 Data Collection

Prior to data collection, the study was reviewed and approved by the IRB at Penn State College of Medicine and Hershey Medical Center (30756EM).

The experiment was conducted in a training center (LionReach) at Hershey Medical Center using 14 laptop computers.

Each participant was asked to read and sign a consent form before logging onto a computer. Each was then randomly assigned to either the experimental group or the control group and assigned a laptop. Each was asked to use a given username and
password. Demographic data about the participant was collected, including age, gender, current position, and number of years of medical experience.

The participants were led to the instruction page which stated the purpose, features, and capabilities of SRCAST-Diagnosis, and a tutorial video was provided. Participants in each group were shown different instruction pages. Participants in the experimental group were informed that they could use a decision support prototype that would prioritize the information and recommendations based on ranking criterion, while participants in the control group were informed that they were using a simulation tool that listed optional choices for them to select based on their own judgment.

Participants had two training sessions to familiarize themselves with the system. Then they proceeded to the ten testing sessions, one by one, in a random order. Each session presented one predefined patient case scenario.

The primary task of the participant in each session was to make choices based on lab tests and select a final diagnosis. Prior to ordering lab tests, participants were asked to select their tentative diagnosis at that specific moment (enabling later analysis of how each hypothesis changed during the process.) Submitted data with the following data types were recorded in a MySQL database:

1. username
2. the group that the participant was assigned to (experimental vs. control)
3. the scenario name from s0 to s9
4. the decision type including lab ordering or diagnosis confirming
5. the time stamp when the decision was made
6. the lab test that was ordered
7. the value of this lab test for the suspected diagnosis
8. the cost of the lab test
9. the accumulative cost of all lab tests
10. the suspected diagnosis
11. the correctness of the diagnosis

After completing the ten scenarios, the participants in the control group had finished the experiment. However, the participants in the experimental group were asked to fill out a survey form. Each was asked to mark whether or not the tool was useful in making the diagnosis and to provide any general comments. In specific, each participant was asked to indicate a degree of agreement (1 – completely disagree, 10 – completely agree) to each of the following statements:

1. It reduced the time needed to make a diagnosis.
2. The recommendations of lab ordering activities were useful.
3. It reduced redundant resource use and improved resource efficiency in lab ordering activities.
4. The ranking of the differential diagnosis were useful for me to quickly identify the true diagnosis.
5. It helped me to quickly update differential diagnosis when new symptoms / lab results arrive.
6. It explained how each recommended diagnosis was made.
7. It organized the information well.
8. It was useful to highlight the relevant information.
9. It made wrong recommendations.
10. It displayed too much information.

11. The recommended diagnosis does not make sense.

12. It updated recommendations too frequently.

13. The visual effect distracted me.

Each participant in the experimental group was also asked to answer the following two questions with a grade from 1 (never) to 10 (always):

1. How frequently did you follow the system’s recommendations?

2. How likely are you to trust the recommendations made by the system?

The experiment ended when the participant submitted the survey. The participants in each group received twenty dollars for compensation and signed confirmations of receipt.

Altogether, 37 participants were involved in the experiment, including nurses, residents, and physicians. Throughout the experiment, sufficient amount of data were collected. Chapter 7 provides the analysis of the collected data.

6.5 Summary

In this section, I first discussed the purpose of the experiment: it was aimed to evaluate the usefulness of the HDSB-enabled decision support tool for clinical diagnosis, mainly focusing on three questions: whether the tool can improve diagnosis accuracy; whether the tool can reduce the resource cost for information seeking without compromising diagnosis accuracy; and what have caused these improvements. Then the experimental hypotheses were given: the tool is expected to improve the diagnosis
accuracy, and at the same time reduce the resource cost for information seeking; these improvements are caused by the counteraction of the cognitive biases (anchoring heuristics and confirmation biases). Then I discussed the design of experiment. It will be a controlled experiment. The independent variables and dependent variables were given. Finally in the section of experiment process, I discussed the knowledge acquisition, scenario design, participant recruitment, and data collection, each of which is an important step of the experiment.
Chapter 7

DATA ANALYSIS AND RESULTS

A total of 37 medical personnel participated in the experiment. Figure 7-1 shows their demographic data. Overall, 20 participants were assigned to the experimental group and 17 were assigned to the control group (Figure 7-1 (a)). There were 19 male participants and 18 female participants (Figure 7-1 (b)). Within the entire group, 20 had less than 5 years of medical experience, 10 had 5-10 years, and 7 had more than 10 years (Figure 7-1 (c)). They represented various roles within the Emergency Department. The majority was residents (18), nurses (10), and physicians and their assistants (6) (Figure 7-1 (d)).

The collected data may suggest an association between the participants’ performance of diagnosis and their demographic group. A participant’s performance can be measured by the diagnosis accuracy and the resource expense. Therefore, the primary goal of this chapter is to analyze the data to test whether the HDSB-enabled diagnostic decision support prototype improved the performance of a participant, especially by increasing diagnosis accuracy and reducing resource expense. A further analysis investigates how these performance improvements occurred. In addition, it may be interesting to observe whether improvement differed among the groups of different roles.
7.1 Diagnosis Accuracy

Each participant was tested with ten scenarios, where six scenarios had only one diagnosis and four scenarios had two co-existing diagnoses. For each scenario, there were six possible results:

1. correct: all diagnoses of the scenario were correctly diagnosed either in single-diagnosis scenarios or in double-diagnosis scenarios;
2. wrong: the scenario was totally misdiagnosed;
3. expired: no diagnosis was given within a limited time;

Figure 7-1: (a) Assigned Group; (b) Gender; (c) Years of medical experience; (d) Position in Emergency Department
4. correct+: the scenario was diagnosed with one actual diagnosis and another extra incorrect diagnosis;

5. correct-: only one correct diagnosis was given in double-diagnosis scenarios;

6. correct*: one diagnosis was correct, but the other was incorrect.

Based on the types of results, there were two ways to measure the diagnosis accuracy:

**Exactly Accurate:** counts exactly diagnosed scenarios.

**Partially Accurate:** counts those scenarios which have at least one correct diagnosis diagnosed.

Clinically, although a patient might not be diagnosed completely with multiple co-existing diseases, it still makes a significant difference to the patient if the diseases can be partially diagnosed. This is why partial accuracy must be analyzed.

The diagnosis accuracy can also be viewed from different perspectives:

**Participant’s perspective:** The diagnosis accuracy can be measured by the number of correctly diagnosed scenarios from each participant.

**Scenario’s perspective:** The diagnosis accuracy can be measured by the number of participants in each group who correctly diagnosed the scenario. Since the total number of participants in each group was different, results can be measured by the percentage of participants who correctly diagnosed the scenario compared to all participants of the group.

The remainder of this section will first show the analysis from participants’ perspective and then from the scenarios’ perspective.
7.1.1 Participant’s Perspective of Diagnosis Accuracy

Figure 7-2 shows the histograms of the number of correctly diagnosed scenarios from participants in both groups. The average number for control group is 4.93, while the average number for experimental group is 5.94. The maximum number from control
group is 7, while the maximum number from experimental group is 9. From observing both histograms together, it is clear that the distribution of the experimental group is on the right side of the control group.

Figure 7-2 shows the general histograms from both groups. To further test the significant difference between the two groups, it is necessary to check whether the two groups followed normal distributions. Figure 7-3 suggests that the sample data from both groups followed a normal distribution, since all the data points fall into the region of a 95% confidence level.

A two-sample Student’s T-test was conducted to test the significant difference of exact accuracy between the control group and the experimental group (Table 7-1). The p-value (0.063) suggests strong evidence that the decision support system used by experimental group greatly helped decision-makers improve their diagnosis accuracy. The improvement has reached 18%.
As mentioned earlier, the participants included nurses, residents, and physicians and their assistants. The improvement level was suspected to differ among the different groups. Table 7-2, Table 7-3, and Table 7-4 show the two-sample Student’s T-test for exact accuracy based on roles.

Table 7-1: Two-sample Student’s T-test for Exact Accuracy (All Participants)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Size</th>
<th>Mean</th>
<th>StDev</th>
<th>T-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>16</td>
<td>4.9375</td>
<td>1.236595</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment</td>
<td>19</td>
<td>5.842105</td>
<td>1.537066</td>
<td>-1.93</td>
<td>0.063</td>
</tr>
</tbody>
</table>

As mentioned earlier, the participants included nurses, residents, and physicians and their assistants. The improvement level was suspected to differ among the different groups. Table 7-2, Table 7-3, and Table 7-4 show the two-sample Student’s T-test for exact accuracy based on roles.

Table 7-2: Two-sample Student’s T-test for Exact Accuracy (Nurses)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Size</th>
<th>Mean</th>
<th>StDev</th>
<th>T-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>control</td>
<td>5</td>
<td>5</td>
<td>1.581139</td>
<td></td>
<td></td>
</tr>
<tr>
<td>experimental</td>
<td>5</td>
<td>5</td>
<td>1.581139</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

The nurses’ diagnosis accuracies from the control group and the experimental group were exactly the same, with mean value 5.

Table 7-3: Two-sample Student’s T-test for Exact Accuracy (Residents)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Size</th>
<th>Mean</th>
<th>StDev</th>
<th>T-Value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>control</td>
<td>7</td>
<td>5</td>
<td>0.57735</td>
<td></td>
<td></td>
</tr>
<tr>
<td>experimental</td>
<td>11</td>
<td>6.545455</td>
<td>1.29334</td>
<td>-3.46</td>
<td>0.004</td>
</tr>
</tbody>
</table>
The diagnosis accuracies from resident group showed a significant improvement when using the tool (p = 0.004). The improvement reached 31%.

Table 7-4: Two-sample Student’s T-test for Exact Accuracy (Physicians)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Size</th>
<th>Mean</th>
<th>StDev</th>
<th>T-value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>control</td>
<td>3</td>
<td>4.333333</td>
<td>2.081666</td>
<td></td>
<td></td>
</tr>
<tr>
<td>experimental</td>
<td>3</td>
<td>4.666667</td>
<td>1.154701</td>
<td>-0.24</td>
<td>0.824</td>
</tr>
</tbody>
</table>

The diagnosis accuracies from physician group showed some improvement, but not a significant one.

The above three tables suggest that the resident group received the greatest advantage and benefitted most from using the tool. It is interesting to note that the improvements of nurse group and the physician group were not as significant as that of the resident group. There are two reasons that could possibly explain this phenomenon:

(1) The level of medical knowledge may have affected a participant’s decision to take advantage of the tool. A nurse’s knowledge set is significantly different from that of a resident or a physician. Nurses are not always trained to make diagnostic decisions. They typically take actions requested by a resident or a physician. They may know the diagnostic decision process at a certain level, but not at a level sufficient to enable them to take full advantage of the tool. This theory may be supported by data from an ED Technician (further away from diagnostic decision-making than a nurse) who was assigned to the experimental group. He had only three scenarios correctly diagnosed with the help of the tool, a number much lower than the average number from other
groups. Therefore, an adequate level of medical knowledge may be an important factor that determines the usefulness of the tool.

(2) The value of the system may be affected by participants’ trust in the system. Although residents and physicians are trained to make diagnostic decisions, they differ significantly in medical experience. The demographic data show that the average number of years of medical experiences was 14.4 for physicians and 4.2 for residents. A physician may trust the system less than a resident did because he was more confident in his own judgment. The survey data also disclosed trust levels. The physicians had an average trust level of 3.67, while the residents had a trust level of 4.44. Therefore, a better collaboration between decision-makers and the decision support system requires a high degree of trust.

7.1.2 Scenarios’ Perspective on Diagnosis Accuracy

The previous section analyzed the data from the participant perspective. This section aims to disclose more details of the accuracy improvement from the scenario perspective—that is, how the tool helped decision-makers in each scenario.

Table 7-5 shows the numbers of participants who identified each specific result in each group for each scenario. A percentage is used to represent the partial accuracy and exact accuracy for each group and each scenario.
In Table 7-5, several observations are worthy of note:

1. There is a significant decrease of correct numbers (column 3) from single-diagnosis scenarios (s0-s5) to double-diagnosis scenarios (s6-s9). This indicates that it was much more difficult to exactly recognize both diagnoses if there were two. However, if counting partial accuracy, the numbers for
single-diagnosis scenarios and double-diagnosis scenarios were similar (column 9).

2. There is no significant improvement from the control group to the experimental group for s2. This may be because the patient case was relatively unusual. The true diagnosis for this scenario was pneumothorax. One symptom uniquely associated with pneumothorax was included in this scenario. Therefore, it was easily diagnosed by participants from both groups.

3. There is a drop from the control group to the experimental group in exact accuracy for s3. The true diagnosis for this scenario is myocardial infarction (MI). However, MI has similar symptoms and appearance with cardiogenic pulmonary edema (CPE) which is clinically associated with MI. The system may have not helped clearly differentiate between MI and CPE. There were five participants in the experimental group who selected more than the true diagnosis.

Paired T-tests were applied to test the significance level of improvement from the control group to the experimental group. Table 7-6 shows the p-values for all scenarios, for only single-diagnosis scenarios, and for double-diagnosis scenarios, considering partial accuracy and exact accuracy. The p-values suggest:

1. The improvement in all scenarios for either partial accuracy or exact accuracy is significant (0.035 and 0.019).

2. The improvements in single-diagnosis for both partial accuracy and exact accuracy are not significant (0.062 and 0.163). This is due to the fact that s2
did not contribute much to the improvement and that there was a drop in s3 as noted earlier.

3. The improvements in double-diagnosis for both partial accuracy and exact accuracy are significant (0.043) or close to significant (0.073). The decision support tool was helpful when identifying either two exact diagnoses or recognizing one diagnosis.

Table 7-6: P-Values Based on Scenarios (All Participants)

<table>
<thead>
<tr>
<th>Criterion of Accuracy</th>
<th>All Scenarios</th>
<th>Single-Diagnosis Scenarios (s0 – s5)</th>
<th>Double-Diagnosis Scenarios (s6 - s9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partial Accuracy</td>
<td>0.035</td>
<td>0.062</td>
<td>0.073</td>
</tr>
<tr>
<td>Exact Accuracy</td>
<td>0.019</td>
<td>0.163</td>
<td>0.043</td>
</tr>
</tbody>
</table>

Table 7-7 contains all the p-values for specific groups. It suggests the following:

1. Consistent with the analysis from the participants’ perspective, the significant improvement in all scenarios with exact accuracy was mostly contributed by residents (0.032).

2. Both for single-diagnosis scenarios and double-diagnosis scenarios, the nurse group did not show improvement in the exact accuracy when using the tool. However, the improvement is still obvious for partial accuracy (0.092). This suggests that the tool was still useful for nurses to partially find the diagnoses, even without sufficient knowledge to clinically differentiate single-diagnosis scenarios and double-diagnosis scenarios.

3. For residents, the improvement in single-diagnosis scenarios is less significant than the improvement in double-diagnosis scenarios for either partial accuracy or exact accuracy. This is due to the fact that there were two
special single-diagnosis scenarios (s2 and s3) that did not contribute to the improvement. However, the decision support tool was very useful for helping residents with double-diagnosis scenarios (0.075 and 0.042).

4. It is interesting to note that physicians’ improvement in single-diagnosis scenarios is more significant by partial accuracy (0.012) than by exact accuracy (1). This is possibly due to the existence of over-reminding: the system listed all possible diagnoses, which led to a significant improvement in partial accuracy but also over-reminded the decision-maker with a false diagnosis. The over-reminding may explain why the improvement in single-diagnosis scenarios is not significant for all participants.

5. The improvement of physicians in double-diagnosis scenarios with both partial accuracy (0.761) and exact accuracy (0.391) are not as significant as the same improvements of residents. The accuracy of these p-values was

<table>
<thead>
<tr>
<th>Role</th>
<th>Criterion of Accuracy</th>
<th>All Scenarios</th>
<th>Single-Diagnosis Scenarios (s0 – s5)</th>
<th>Double-Diagnosis Scenarios (s6 – s9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nurses</td>
<td>Partial Accuracy</td>
<td>0.223</td>
<td>1</td>
<td>0.092</td>
</tr>
<tr>
<td></td>
<td>Exact Accuracy</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Residents</td>
<td>Partial Accuracy</td>
<td>0.135</td>
<td>0.344</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>Exact Accuracy</td>
<td>0.032</td>
<td>0.275</td>
<td>0.042</td>
</tr>
<tr>
<td>Physicians</td>
<td>Partial Accuracy</td>
<td>0.066</td>
<td>0.012</td>
<td>0.761</td>
</tr>
<tr>
<td></td>
<td>Exact Accuracy</td>
<td>0.678</td>
<td>1</td>
<td>0.391</td>
</tr>
</tbody>
</table>
affected by the sample size of the group. Only three physicians participated in
the group. With a small sample size, one outlier may have a significant effect
on the p-value.

7.1.3 Summary

Based on the above analysis, this section summarizes the current findings on
diagnosis accuracy:

1. The decision support tool significantly improved the decision accuracy either
   with partial accuracy and exact accuracy in general.
2. Among the three roles, residents benefited the most from the tool, and
   physicians benefited more than nurses.
3. The decision support tool introduced over-reminding to some scenarios,
   which made improvements in single-diagnosis scenarios not as significant.
4. The improvements in double-diagnosis scenarios are more significant than
   the improvements in single-diagnosis scenarios. The system either helped
   identify the two exact diagnoses or helped identify at least one diagnosis.

The system was proved to be very useful for improving diagnosis accuracy.
However, to produce better performance, an adequate level of medical knowledge is
required, and a certain level of trust is preferred.
7.2 Cost Efficiency

The HDSB framework depicted the abstract process of clinical diagnostic decision-making, and developed an information-theoretic method to calculate the value of an information gathering activity. The web-based implementation of the HDSB framework for clinical diagnostic decision support has applied this method to rank the related lab tests for a particular suspected diagnosis. A lab test that is more capable of distinguishing the differential diagnoses is listed higher. It is expected that lab test resources can be optimized by showing participants the usefulness of each lab test in diagnosing a patient situation.

Through the experiment, the data about each participant’s lab ordering activities were collected.

7.2.1 Analyses on Resource Cost

Figure 7-4 shows the histograms of resource cost by each participant in both groups. It suggests that the control group has more participants at the high cost end, and that the experimental group has more participants at the low cost end.
Table 7-8 lists the results from a two-sample Student’s T-test. The p-value suggests that there is a significant difference between the resource costs of the two groups. The resources saved were 64% on average.

Figure 7-4: Histogram of Resource Cost in Control Group and Experimental Group
Table 7-9 shows the results from a two-sample Student’s T-test for each scenario. The t-values and p-values for each scenario are listed. It shows that the p-values of single-diagnosis scenarios are much smaller than those of double-diagnosis scenarios, which means that the decrease of the resource cost in double-diagnosis scenarios is not as significant as in single-diagnosis scenarios. In particular, scenario s8 caused the experimental group to expend more resources in terms of cost.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Size</th>
<th>Mean</th>
<th>StDev</th>
<th>T-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>168</td>
<td>757</td>
<td>592</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>200</td>
<td>461</td>
<td>377</td>
<td>5.61</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 7-9: P-Value of Resource Cost
It can be reasonably explained by the level of scenario complexities. In single-diagnosis scenarios, a differential diagnosis can be made prominent by using the tool, as opposed to other differential diagnoses after conducting several lab tests. However, in double-diagnosis scenarios, the system does not rule out either of the two diagnoses. This will cause participants to order more lab tests, attempting to rule out one of the diagnoses until they suspect that two diagnoses may co-exist. Therefore, it is natural for a participant to expend more resource costs in complex scenarios.

Table 7-9 shows the average cost a participant spent for each scenario. There are significant decreases of resource cost among different roles and for different types of

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Group</th>
<th>Total Cost</th>
<th>Size</th>
<th>Mean</th>
<th>StDev</th>
<th>T-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>s0</td>
<td>control</td>
<td>14056</td>
<td>17</td>
<td>826.8235294</td>
<td>599.7076616</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s0</td>
<td>experiment</td>
<td>7367</td>
<td>20</td>
<td>368.35</td>
<td>280.1979141</td>
<td>2.894933032</td>
<td>0.009</td>
</tr>
<tr>
<td>s1</td>
<td>control</td>
<td>13156</td>
<td>17</td>
<td>773.8823529</td>
<td>614.9141081</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s1</td>
<td>experiment</td>
<td>7858</td>
<td>20</td>
<td>392.9</td>
<td>274.4045995</td>
<td>2.362424737</td>
<td>0.028</td>
</tr>
<tr>
<td>s2</td>
<td>control</td>
<td>8545</td>
<td>17</td>
<td>502.6470588</td>
<td>654.8342291</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s2</td>
<td>experiment</td>
<td>1635</td>
<td>20</td>
<td>81.75</td>
<td>96.55043678</td>
<td>2.625989982</td>
<td>0.018</td>
</tr>
<tr>
<td>s3</td>
<td>control</td>
<td>13508</td>
<td>16</td>
<td>844.25</td>
<td>634.8469107</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s3</td>
<td>experiment</td>
<td>10329</td>
<td>20</td>
<td>516.45</td>
<td>345.149773</td>
<td>1.857415175</td>
<td>0.077</td>
</tr>
<tr>
<td>s4</td>
<td>control</td>
<td>10003</td>
<td>16</td>
<td>625.1875</td>
<td>550.360151</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s4</td>
<td>experiment</td>
<td>5487</td>
<td>20</td>
<td>274.35</td>
<td>211.4753376</td>
<td>2.411433169</td>
<td>0.027</td>
</tr>
<tr>
<td>s5</td>
<td>control</td>
<td>18941</td>
<td>17</td>
<td>1114.176471</td>
<td>472.1958327</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s5</td>
<td>experiment</td>
<td>13355</td>
<td>20</td>
<td>667.75</td>
<td>444.6950315</td>
<td>2.943424934</td>
<td>0.006</td>
</tr>
<tr>
<td>s6</td>
<td>control</td>
<td>11732</td>
<td>17</td>
<td>690.1176471</td>
<td>527.6806423</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s6</td>
<td>experiment</td>
<td>7374</td>
<td>17</td>
<td>433.765</td>
<td>119.862004</td>
<td>1.953289098</td>
<td>0.067</td>
</tr>
<tr>
<td>s7</td>
<td>control</td>
<td>11634</td>
<td>17</td>
<td>684.3529412</td>
<td>613.6781466</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s7</td>
<td>experiment</td>
<td>8623</td>
<td>20</td>
<td>431.15</td>
<td>424.2361898</td>
<td>1.434587714</td>
<td>0.163</td>
</tr>
<tr>
<td>s8</td>
<td>control</td>
<td>10341</td>
<td>17</td>
<td>608.2941176</td>
<td>402.3261682</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s8</td>
<td>experiment</td>
<td>14592</td>
<td>20</td>
<td>729.6</td>
<td>411.7800128</td>
<td>-0.904168482</td>
<td>0.372</td>
</tr>
<tr>
<td>s9</td>
<td>control</td>
<td>15333</td>
<td>17</td>
<td>901.9411765</td>
<td>704.6501322</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s9</td>
<td>experiment</td>
<td>11885</td>
<td>20</td>
<td>594.25</td>
<td>408.5385861</td>
<td>1.587789982</td>
<td>0.125</td>
</tr>
</tbody>
</table>
scenarios between the groups. The percentages indicate the ratio of cost savings from the control group to the experimental group.

Table 7-10: Means of Resource Cost

<table>
<thead>
<tr>
<th></th>
<th>Group</th>
<th>All Scenarios</th>
<th>Single-Diagnosis Scenarios</th>
<th>Double-Diagnosis Scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All Participants</strong></td>
<td>Control</td>
<td>757</td>
<td>781</td>
<td>721</td>
</tr>
<tr>
<td></td>
<td>Experiment</td>
<td>449 (41%)</td>
<td>383 (51%)</td>
<td>547 (24%)</td>
</tr>
<tr>
<td><strong>Nurses</strong></td>
<td>Control</td>
<td>741</td>
<td>744</td>
<td>736</td>
</tr>
<tr>
<td></td>
<td>Experiment</td>
<td>366 (51%)</td>
<td>322 (57%)</td>
<td>432 (41%)</td>
</tr>
<tr>
<td><strong>Residents</strong></td>
<td>Control</td>
<td>769</td>
<td>773</td>
<td>763</td>
</tr>
<tr>
<td></td>
<td>Experiment</td>
<td>464 (40%)</td>
<td>385 (50%)</td>
<td>582 (24%)</td>
</tr>
<tr>
<td><strong>Physicians</strong></td>
<td>Control</td>
<td>882</td>
<td>990</td>
<td>719</td>
</tr>
<tr>
<td></td>
<td>Experiment</td>
<td>540 (39%)</td>
<td>467 (53%)</td>
<td>651 (9%)</td>
</tr>
</tbody>
</table>

Table 7-11 lists the p-values comparing the resource costs between the control group and the experimental group for different roles and different types of scenarios. The levels of significance (p-values for different types of scenarios) are consistent among different roles (nurses, residents, and physicians). The decreases of resource cost in single-diagnosis scenarios by the three roles are all significant (0.001, 0.001, 0.007), while the decreases in double-diagnosis scenarios are not as significant (0.222, 0.227, 0.751). However, these decreases in double-diagnosis scenarios are still apparent. The degrees of saving are 41%, 24%, and 9% respectively for nurses, residents, and physicians.
The fact that the saving of resource cost among physicians is not as significant as that for nurses and residents is possibly due to two reasons:

1. The sample size for the physician group is small. Therefore, one outlier may have a large influence on the test results.

2. The physicians did not trust the diagnosis recommended by the system as much as nurses and residents did. Therefore, they ordered more lab tests to confirm their own judgment.

Table 7-12 lists the number of times a lab test was ordered for each scenario in each group. The last column shows the average number of lab tests ordered for each scenario in each group. There is an apparent reduction of number of labs test from the control group to the experimental group in single-diagnosis scenarios (s0-s5). However, the number did not diminish for double-diagnosis scenarios (s6-s9), which is consistent with the analysis of cost saving discussed earlier. Therefore, we can conclude that the saving of cost was due to a smaller number of lab tests used for diagnosis.

Table 7-11: P-Values for Resource Cost Comparison

<table>
<thead>
<tr>
<th></th>
<th>All Scenarios</th>
<th>Single-Diagnosis Scenarios</th>
<th>Double-Diagnosis Scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Participants</td>
<td>0</td>
<td>0</td>
<td>0.178</td>
</tr>
<tr>
<td>Nurses</td>
<td>0.001</td>
<td>0.001</td>
<td>0.222</td>
</tr>
<tr>
<td>Residents</td>
<td>0.001</td>
<td>0.001</td>
<td>0.227</td>
</tr>
<tr>
<td>Physicians</td>
<td>0.023</td>
<td>0.007</td>
<td>0.751</td>
</tr>
</tbody>
</table>
Table 7-12: Number of Labs

<table>
<thead>
<tr>
<th>scenario</th>
<th>group</th>
<th>CBC</th>
<th>D-Dimer</th>
<th>EKG</th>
<th>Cardiac Panel</th>
<th>Echo</th>
<th>BNP</th>
<th>XRay</th>
<th>CT</th>
<th>Peak Flow</th>
<th>Basic Metabolic Panel</th>
<th>ABG</th>
<th>VQ Scan</th>
<th>CPro</th>
<th>Ultra Sound</th>
<th>total</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>s0</td>
<td>control</td>
<td>11</td>
<td>6</td>
<td>16</td>
<td>14</td>
<td>4</td>
<td>9</td>
<td>14</td>
<td>4</td>
<td>0</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>89</td>
<td>5.235</td>
</tr>
<tr>
<td>s0</td>
<td>experiment</td>
<td>9</td>
<td>2</td>
<td>14</td>
<td>12</td>
<td>1</td>
<td>16</td>
<td>15</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>76</td>
<td>3.8</td>
</tr>
<tr>
<td>s1</td>
<td>control</td>
<td>10</td>
<td>6</td>
<td>8</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>11</td>
<td>7</td>
<td>10</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>72</td>
<td>4.235</td>
</tr>
<tr>
<td>s1</td>
<td>experiment</td>
<td>9</td>
<td>12</td>
<td>10</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>2</td>
<td>10</td>
<td>8</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>76</td>
<td>3.8</td>
</tr>
<tr>
<td>s2</td>
<td>control</td>
<td>6</td>
<td>2</td>
<td>9</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>15</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>48</td>
<td>2.823</td>
</tr>
<tr>
<td>s2</td>
<td>experiment</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>0.85</td>
</tr>
<tr>
<td>s3</td>
<td>control</td>
<td>12</td>
<td>2</td>
<td>16</td>
<td>14</td>
<td>4</td>
<td>2</td>
<td>13</td>
<td>4</td>
<td>1</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>77</td>
<td>4.529</td>
</tr>
<tr>
<td>s3</td>
<td>experiment</td>
<td>12</td>
<td>6</td>
<td>19</td>
<td>19</td>
<td>2</td>
<td>4</td>
<td>19</td>
<td>1</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>90</td>
<td>4.5</td>
</tr>
<tr>
<td>s4</td>
<td>control</td>
<td>15</td>
<td>4</td>
<td>14</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>15</td>
<td>4</td>
<td>6</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>78</td>
<td>4.588</td>
</tr>
<tr>
<td>s4</td>
<td>experiment</td>
<td>13</td>
<td>5</td>
<td>10</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>19</td>
<td>0</td>
<td>9</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>71</td>
<td>3.55</td>
</tr>
<tr>
<td>s5</td>
<td>control</td>
<td>11</td>
<td>4</td>
<td>14</td>
<td>15</td>
<td>2</td>
<td>0</td>
<td>10</td>
<td>13</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>77</td>
<td>4.529</td>
</tr>
<tr>
<td>s5</td>
<td>experiment</td>
<td>8</td>
<td>2</td>
<td>13</td>
<td>13</td>
<td>0</td>
<td>5</td>
<td>13</td>
<td>10</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>71</td>
<td>3.55</td>
</tr>
<tr>
<td>s6</td>
<td>control</td>
<td>12</td>
<td>5</td>
<td>14</td>
<td>15</td>
<td>3</td>
<td>7</td>
<td>16</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>82</td>
<td>4.823</td>
</tr>
<tr>
<td>s6</td>
<td>experiment</td>
<td>13</td>
<td>8</td>
<td>19</td>
<td>13</td>
<td>0</td>
<td>18</td>
<td>19</td>
<td>3</td>
<td>0</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>103</td>
<td>5.15</td>
</tr>
<tr>
<td>s7</td>
<td>control</td>
<td>11</td>
<td>3</td>
<td>8</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>14</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>62</td>
<td>3.647</td>
</tr>
<tr>
<td>s7</td>
<td>experiment</td>
<td>13</td>
<td>6</td>
<td>14</td>
<td>8</td>
<td>0</td>
<td>6</td>
<td>16</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>81</td>
<td>4.05</td>
</tr>
<tr>
<td>s8</td>
<td>control</td>
<td>14</td>
<td>4</td>
<td>12</td>
<td>10</td>
<td>0</td>
<td>3</td>
<td>14</td>
<td>5</td>
<td>3</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>74</td>
<td>4.352</td>
</tr>
<tr>
<td>s8</td>
<td>experiment</td>
<td>15</td>
<td>12</td>
<td>18</td>
<td>10</td>
<td>0</td>
<td>3</td>
<td>18</td>
<td>8</td>
<td>6</td>
<td>9</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>103</td>
<td>5.15</td>
</tr>
<tr>
<td>s9</td>
<td>control</td>
<td>12</td>
<td>4</td>
<td>15</td>
<td>13</td>
<td>6</td>
<td>7</td>
<td>13</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>84</td>
<td>4.941</td>
</tr>
<tr>
<td>s9</td>
<td>experiment</td>
<td>10</td>
<td>6</td>
<td>15</td>
<td>12</td>
<td>4</td>
<td>16</td>
<td>17</td>
<td>1</td>
<td>2</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>95</td>
<td>4.75</td>
</tr>
</tbody>
</table>

Table 7-12 shows another reason why the control group spent more on resource cost in diagnosis. The column labeled CT (Column 10) shows that there was a great decrease in the use of CT scans from the control group to the experimental group for almost all scenarios. However, a CT scan is several times more expensive than other lab
tests. Therefore, the tool successfully prioritized the use of resources that significantly contributed to the cost saving.

7.2.2 Summary

This section summarizes the current findings on cost efficiency.

1. The tool helped decision-makers significantly reduce the cost of lab test resources. The average savings of resource costs that a participant spent for a scenario was 41%.

2. The reduction of resource expense was much more significant in single-diagnosis scenarios (51%) than in double-diagnosis scenarios (24%). This is because a double-diagnosis scenario requires a decision-maker to order more lab tests, attempting to differentiate the two diagnoses.

3. Decision-makers from all three roles reduced the resource cost with the help of the decision support tool. The nurse group benefited the most from the tool. The average savings for the nurse group was 51%. However, as noted earlier, the nurse group performed the worst in improving the diagnosis accuracy. Putting these two findings together, it may be implied that nurses followed the lab tests recommended by the tool but lacked sufficient clinical knowledge and experience to recognize when the lab tests were sufficient to make a correct diagnosis.

4. The reduction of resource cost for the physician group in double-diagnosis scenarios is much lower (9%) than for other groups. This is consistent with
the finding that the physicians did not improve the diagnosis accuracy significantly ($p=0.391$). Both findings are the consequence of the fact that a physician trusted the tool at a low level and hence spent more resources to verify a hypothesis.

5. The cost savings were brought about by the decision support tool in two ways: reducing the overall number of lab tests, and minimizing the more expensive lab tests.

7.3 Counteracting Cognitive Biases

Five types of human cognitive biases related to hypothesis making and information seeking were discussed in Chapter 2. It is believed that the performance improvement from the control group and the experimental group was possibly brought about by the HDSB-enabled decision support tool’s ability to counteract some of the negative effects of the cognitive biases. Therefore, this section aims to investigate: (1) the particular biases of a decision-maker during the diagnosis-making process in the experiment, (2) how the biases affected performance, and (3) how the tool improved the performance of diagnosis-making by counteracting the negative effects of the cognitive biases.

A diagnosis-making process is generally composed of two critical decisions: (1) making a differential diagnosis, and (2) deciding which lab tests to order. Cognitive biases will influence both decisions: they may affect the generation of differential
diagnoses, or affect the selection of lab tests. Furthermore, the result of one decision may affect the result of the other.

Among the five types of the cognitive biases, the anchoring heuristics and confirmation biases are specifically related to the two decisions of making diagnosis and selecting lab tests. In clinical diagnostic situations, the anchoring heuristics is the tendency that a decision-maker may be reluctant to change his initial diagnosis, unless the evidence is strong enough to persuade him to change the diagnosis. This reluctance will make the decision-maker to stick to the wrong diagnosis and affect the diagnosis accuracy. A confirmation bias is the tendency that a decision-maker may favor the labs that he believes to confirm the suspected diagnosis but may be actually less useful to differentiate the differential diagnosis. It will make the decision-maker to spend more lab resources before he modifies his diagnosis. Therefore, the following two sub-sections will discuss the effects of anchoring heuristics and confirmation biases on diagnosis accuracy and cost efficiency in respect.

7.3.1 Effect of Anchoring Heuristics on Diagnosis Accuracy

Table 7-13 shows the effect of anchoring heuristics on diagnosis accuracy. For each scenario in each group, the number of participants who made a wrong tentative diagnosis on a first attempt is listed in Column 3, and the number of participants who finally corrected a diagnosis is listed in Column 4. The last column lists the percentage change from wrong to correct.
Anchoring heuristics can be observed from the table. For all scenarios, only limited number of participants revised their initial (wrong) diagnosis and reached the correct diagnosis. The rate of switching to a correct diagnosis is as low as 33%, which means two-thirds of the participants failed to modify their initial (wrong) diagnosis, even with additional information obtained from lab tests.

Section 7.1 showed the usefulness of the HDSB-enabled decision support tool to significantly improve the diagnosis accuracy, but did not disclose the fundamental
reasons why the tool can improve the diagnosis accuracy. However, these fundamental reasons can be seen in Table 7-13:

(1) The tool significantly helped the decision-makers determine the correct diagnosis at their first attempts. The number of participants in the experimental group who made a wrong tentative diagnosis at their first attempt is significantly smaller than that of control group, as shown in Column 3 of Table 7-13. The p-value is 0.006.

(2) The tool significantly helped the decision-maker select the correct diagnosis when provided with lab tests. Among those participants who had made the wrong tentative diagnosis, the experimental group had a larger percentage that made the final diagnosis correct, as shown in last column of Table 7-13. The p-value is 0.031. The tool helped more participants switch from an initially wrong diagnosis to a correct final diagnosis. Therefore, the tool is proved to be able to counteract the negative effects of anchoring heuristics. This improvement is possible because the tool offers quantitative evidence to convince a decision-maker that the hypothesis needs to be changed, whereas that idea may not occur without the tool.

In summary, anchoring heuristics make a decision-maker reluctant to change an initial diagnosis. This experiment has shown that the tool is useful to counteract the negative effects of anchoring heuristics by demonstrating significant improvement in the revision of initially wrong diagnoses to correct final diagnoses. This capability of the HDSB-enabled decision support tool, as well as its ability to help decision-makers find the correct diagnosis at their first attempts, has substantially contributed to the significant improvements of the diagnosis accuracy.
7.3.2 Effect of Confirmation Biases on Cost Efficiency

Confirmation biases affect the cost of resources during a decision-making process. It is logically natural that a decision-maker will attempt to use more resources to collect the information believed to be useful to support a tentative diagnosis. However, because the tentative diagnosis itself may be wrong, some of the resources used for information seeking based on this tentative diagnosis are very likely to be inefficient or unnecessary. This section focuses on finding evidence from the experiment data to support the above judgment.

Table 7-14 lists the related data with diagnosis revision. Column 3 is the number of users who made the correct final diagnosis. Column 4 is the average time spent while a diagnosis was switched to the correct one. Column 5 is the average number of tests ordered when a hypothesis was switched to the correct one. Column 6 is the average cost of lab resources spent when a diagnosis was switched to a correct one.

Table 7-14 shows the data from all participants. Similar tables that contain data from nurses, residents, and physicians respectively are also provided (Table 7-15, Table 7-16, and Table 7-17).
Table 7-14: Resources Used When Initial Diagnosis Revised (All Participants)

<table>
<thead>
<tr>
<th>scenario</th>
<th>group</th>
<th># of users who had made the final diagnosis correct</th>
<th>Average time spent when a diagnosis was correctly revised</th>
<th>Average # of tests being ordered when a diagnosis was correctly revised</th>
<th>Average cost of lab resources being spent when a diagnosis was correctly switched</th>
</tr>
</thead>
<tbody>
<tr>
<td>s0</td>
<td>control</td>
<td>15</td>
<td>101415.4667</td>
<td>2.26666667</td>
<td>458.2666667</td>
</tr>
<tr>
<td>s0</td>
<td>experiment</td>
<td>20</td>
<td>62627.45</td>
<td>0.3</td>
<td>24.15</td>
</tr>
<tr>
<td>s1</td>
<td>control</td>
<td>8</td>
<td>108573.75</td>
<td>1.75</td>
<td>549.625</td>
</tr>
<tr>
<td>s1</td>
<td>experiment</td>
<td>17</td>
<td>85645.82353</td>
<td>1.411764706</td>
<td>211.7058824</td>
</tr>
<tr>
<td>s2</td>
<td>control</td>
<td>16</td>
<td>47138.1875</td>
<td>0.75</td>
<td>136.75</td>
</tr>
<tr>
<td>s2</td>
<td>experiment</td>
<td>19</td>
<td>47481.42105</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s3</td>
<td>control</td>
<td>15</td>
<td>54745.4</td>
<td>0.66666667</td>
<td>427.733333</td>
</tr>
<tr>
<td>s3</td>
<td>experiment</td>
<td>18</td>
<td>51891.38889</td>
<td>0.055555556</td>
<td>76.2222222</td>
</tr>
<tr>
<td>s4</td>
<td>control</td>
<td>13</td>
<td>131678.8462</td>
<td>3.769230769</td>
<td>535</td>
</tr>
<tr>
<td>s4</td>
<td>experiment</td>
<td>18</td>
<td>50471.33333</td>
<td>0.388888889</td>
<td>41.05555556</td>
</tr>
<tr>
<td>s5</td>
<td>control</td>
<td>14</td>
<td>90266.35714</td>
<td>2.571428571</td>
<td>515.0714286</td>
</tr>
<tr>
<td>s5</td>
<td>experiment</td>
<td>19</td>
<td>84621.21053</td>
<td>1.578947368</td>
<td>147.8421053</td>
</tr>
<tr>
<td>s6</td>
<td>control</td>
<td>16</td>
<td>62520.5625</td>
<td>0.625</td>
<td>136.875</td>
</tr>
<tr>
<td>s6</td>
<td>experiment</td>
<td>20</td>
<td>80680.6</td>
<td>0.85</td>
<td>108</td>
</tr>
<tr>
<td>s7</td>
<td>control</td>
<td>11</td>
<td>88777.18182</td>
<td>1.909090909</td>
<td>412.7272727</td>
</tr>
<tr>
<td>s7</td>
<td>experiment</td>
<td>14</td>
<td>99256.5</td>
<td>1.071428571</td>
<td>110.0714286</td>
</tr>
<tr>
<td>s8</td>
<td>control</td>
<td>12</td>
<td>91845.58333</td>
<td>1.583333333</td>
<td>257.75</td>
</tr>
<tr>
<td>s8</td>
<td>experiment</td>
<td>19</td>
<td>110791.1579</td>
<td>2.368421053</td>
<td>395.6842105</td>
</tr>
<tr>
<td>s9</td>
<td>control</td>
<td>14</td>
<td>93038.42857</td>
<td>3.142857143</td>
<td>486.3571429</td>
</tr>
<tr>
<td>s9</td>
<td>experiment</td>
<td>19</td>
<td>72438.36842</td>
<td>0.473684211</td>
<td>39.89473684</td>
</tr>
</tbody>
</table>
Table 7-15: Resources Used When Initial Diagnosis Revised (Nurses)

<table>
<thead>
<tr>
<th>scenario</th>
<th>group</th>
<th># of users who had made the final diagnosis correct</th>
<th>Average time spent when a diagnosis was correctly revised</th>
<th>Average # of tests being ordered when a diagnosis was correctly revised</th>
<th>Average cost of lab resources being spent when a diagnosis was correctly switched</th>
</tr>
</thead>
<tbody>
<tr>
<td>s0</td>
<td>control</td>
<td>4</td>
<td>126599.25</td>
<td>2.75</td>
<td>629.25</td>
</tr>
<tr>
<td>s0</td>
<td>experiment</td>
<td>5</td>
<td>74302.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s1</td>
<td>control</td>
<td>3</td>
<td>111065.6667</td>
<td>2.6666666667</td>
<td>178.3333333</td>
</tr>
<tr>
<td>s1</td>
<td>experiment</td>
<td>2</td>
<td>84640.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s2</td>
<td>control</td>
<td>5</td>
<td>48527.8</td>
<td>0</td>
<td>226.2</td>
</tr>
<tr>
<td>s2</td>
<td>experiment</td>
<td>4</td>
<td>62365.25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s3</td>
<td>control</td>
<td>5</td>
<td>69028.4</td>
<td>1.2</td>
<td>245.6</td>
</tr>
<tr>
<td>s3</td>
<td>experiment</td>
<td>5</td>
<td>56517.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s4</td>
<td>control</td>
<td>4</td>
<td>86014.25</td>
<td>2.25</td>
<td>196.75</td>
</tr>
<tr>
<td>s4</td>
<td>experiment</td>
<td>4</td>
<td>36003.25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s5</td>
<td>control</td>
<td>4</td>
<td>103698</td>
<td>3.25</td>
<td>728.5</td>
</tr>
<tr>
<td>s5</td>
<td>experiment</td>
<td>5</td>
<td>89256.8</td>
<td>1</td>
<td>78.6</td>
</tr>
<tr>
<td>s6</td>
<td>control</td>
<td>5</td>
<td>66192</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s6</td>
<td>experiment</td>
<td>5</td>
<td>70142</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s7</td>
<td>control</td>
<td>3</td>
<td>61084.3333333</td>
<td>0.6666666667</td>
<td>51</td>
</tr>
<tr>
<td>s7</td>
<td>experiment</td>
<td>5</td>
<td>78253.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s8</td>
<td>control</td>
<td>3</td>
<td>133329</td>
<td>3</td>
<td>511.66666667</td>
</tr>
<tr>
<td>s8</td>
<td>experiment</td>
<td>4</td>
<td>129572</td>
<td>1.75</td>
<td>140.5</td>
</tr>
<tr>
<td>s9</td>
<td>control</td>
<td>4</td>
<td>126017</td>
<td>4</td>
<td>531.25</td>
</tr>
<tr>
<td>s9</td>
<td>experiment</td>
<td>5</td>
<td>51881.8</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 7-16: Resources Used When Initial Diagnosis Revised (Residents)

<table>
<thead>
<tr>
<th>scenario</th>
<th>group</th>
<th># of users who had made the final diagnosis correct</th>
<th>Average time spent when a diagnosis was correctly revised</th>
<th>Average # of tests being ordered when a diagnosis was correctly revised</th>
<th>Average cost of lab resources being spent when a diagnosis was correctly switched</th>
</tr>
</thead>
<tbody>
<tr>
<td>s0</td>
<td>control</td>
<td>8</td>
<td>103958.25</td>
<td>2.5</td>
<td>515.75</td>
</tr>
<tr>
<td>s0</td>
<td>experiment</td>
<td>11</td>
<td>48258.81818</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s1</td>
<td>control</td>
<td>4</td>
<td>105386.75</td>
<td>0.75</td>
<td>741.5</td>
</tr>
<tr>
<td>s1</td>
<td>experiment</td>
<td>11</td>
<td>86698.09091</td>
<td>1.818181818</td>
<td>292.3636364</td>
</tr>
<tr>
<td>s2</td>
<td>control</td>
<td>8</td>
<td>44329.625</td>
<td>1</td>
<td>90.625</td>
</tr>
<tr>
<td>s2</td>
<td>experiment</td>
<td>11</td>
<td>41796.72727</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s3</td>
<td>control</td>
<td>7</td>
<td>40981.57143</td>
<td>0.571428571</td>
<td>646.2857143</td>
</tr>
<tr>
<td>s3</td>
<td>experiment</td>
<td>9</td>
<td>46158.77778</td>
<td>0.1111111111</td>
<td>152.4444444</td>
</tr>
<tr>
<td>s4</td>
<td>control</td>
<td>6</td>
<td>143842.5</td>
<td>4</td>
<td>573.6666667</td>
</tr>
<tr>
<td>s4</td>
<td>experiment</td>
<td>11</td>
<td>56927.18182</td>
<td>0.6363636366</td>
<td>67.181818188</td>
</tr>
<tr>
<td>s5</td>
<td>control</td>
<td>7</td>
<td>74774.71429</td>
<td>1.857142857</td>
<td>385.7142857</td>
</tr>
<tr>
<td>s5</td>
<td>experiment</td>
<td>11</td>
<td>72220.36364</td>
<td>1.454545455</td>
<td>140.6363636</td>
</tr>
<tr>
<td>s6</td>
<td>control</td>
<td>8</td>
<td>53180.375</td>
<td>0.75</td>
<td>232.25</td>
</tr>
<tr>
<td>s6</td>
<td>experiment</td>
<td>11</td>
<td>82038.63636</td>
<td>1</td>
<td>84.72727273</td>
</tr>
<tr>
<td>s7</td>
<td>control</td>
<td>6</td>
<td>115615.5</td>
<td>2.8333333333</td>
<td>697.1666667</td>
</tr>
<tr>
<td>s7</td>
<td>experiment</td>
<td>9</td>
<td>110924.7778</td>
<td>1.6666666667</td>
<td>171.2222222</td>
</tr>
<tr>
<td>s8</td>
<td>control</td>
<td>7</td>
<td>57560.14286</td>
<td>0.714285714</td>
<td>171.8571429</td>
</tr>
<tr>
<td>s8</td>
<td>experiment</td>
<td>11</td>
<td>103461.9091</td>
<td>2.727272727</td>
<td>500.9090909</td>
</tr>
<tr>
<td>s9</td>
<td>control</td>
<td>7</td>
<td>100968.2857</td>
<td>4</td>
<td>669.1428571</td>
</tr>
<tr>
<td>s9</td>
<td>experiment</td>
<td>11</td>
<td>82373</td>
<td>0.81818181818</td>
<td>68.90909091</td>
</tr>
</tbody>
</table>
Table 7-17: Resources Used When Initial Diagnosis Revised (Physicians)

<table>
<thead>
<tr>
<th>scenario</th>
<th>group</th>
<th># of users who had made the final diagnosis correct</th>
<th>Average time spent when a diagnosis was correctly revised</th>
<th>Average # of tests being ordered when a diagnosis was correctly revised</th>
<th>Average cost of lab resources being spent when a diagnosis was correctly switched</th>
</tr>
</thead>
<tbody>
<tr>
<td>s0</td>
<td>control</td>
<td>2</td>
<td>42751</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s0</td>
<td>experiment</td>
<td>3</td>
<td>93285.6667</td>
<td>2</td>
<td>161</td>
</tr>
<tr>
<td>s1</td>
<td>control</td>
<td>1</td>
<td>113846</td>
<td>3</td>
<td>896</td>
</tr>
<tr>
<td>s1</td>
<td>experiment</td>
<td>3</td>
<td>96539.6667</td>
<td>1.33333333333</td>
<td>127.6666667</td>
</tr>
<tr>
<td>s2</td>
<td>control</td>
<td>2</td>
<td>25023</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s2</td>
<td>experiment</td>
<td>3</td>
<td>47567.6667</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s3</td>
<td>control</td>
<td>2</td>
<td>64579.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s3</td>
<td>experiment</td>
<td>3</td>
<td>51833</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s4</td>
<td>control</td>
<td>3</td>
<td>168237.6667</td>
<td>5.33333333333</td>
<td>908.6666667</td>
</tr>
<tr>
<td>s4</td>
<td>experiment</td>
<td>3</td>
<td>46090.66667</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s5</td>
<td>control</td>
<td>2</td>
<td>95716.5</td>
<td>2.5</td>
<td>569.5</td>
</tr>
<tr>
<td>s5</td>
<td>experiment</td>
<td>3</td>
<td>122365</td>
<td>3</td>
<td>289.6666667</td>
</tr>
<tr>
<td>s6</td>
<td>control</td>
<td>2</td>
<td>53549.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s6</td>
<td>experiment</td>
<td>3</td>
<td>56115.66667</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s7</td>
<td>control</td>
<td>2</td>
<td>49801.5</td>
<td>1</td>
<td>102</td>
</tr>
<tr>
<td>s7</td>
<td>experiment</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s8</td>
<td>control</td>
<td>2</td>
<td>149619.5</td>
<td>2.5</td>
<td>177.5</td>
</tr>
<tr>
<td>s8</td>
<td>experiment</td>
<td>3</td>
<td>62428.66667</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s9</td>
<td>control</td>
<td>2</td>
<td>27052.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s9</td>
<td>experiment</td>
<td>3</td>
<td>70272.33333</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 7-18 contains the p-values derived by comparing the results from two groups for different data types, role of participants, and types of scenarios. There are several interesting findings that can be observed from the table.

### Table 7-18: P-values for Hypothesis Switch

<table>
<thead>
<tr>
<th>Participants</th>
<th>Data Type</th>
<th>All Scenarios</th>
<th>Single-Diagnosis Scenarios</th>
<th>Double-Diagnosis Scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td>All participants</td>
<td>Number of tests</td>
<td>0.029</td>
<td>0.035</td>
<td>0.472</td>
</tr>
<tr>
<td></td>
<td>Cost of lab tests</td>
<td>0.002</td>
<td>0.001</td>
<td>0.311</td>
</tr>
<tr>
<td>Nurses</td>
<td>Number of tests</td>
<td>0.003</td>
<td>0.008</td>
<td>0.191</td>
</tr>
<tr>
<td></td>
<td>Cost of lab tests</td>
<td>0.002</td>
<td>0.011</td>
<td>0.159</td>
</tr>
<tr>
<td>Residents</td>
<td>Number of tests</td>
<td>0.152</td>
<td>0.149</td>
<td>0.668</td>
</tr>
<tr>
<td></td>
<td>Cost of lab tests</td>
<td>0.006</td>
<td>0.003</td>
<td>0.348</td>
</tr>
<tr>
<td>Physicians</td>
<td>Number of tests</td>
<td>0.240</td>
<td>0.501</td>
<td>0.235</td>
</tr>
<tr>
<td></td>
<td>Cost of lab tests</td>
<td>0.097</td>
<td>0.159</td>
<td>0.204</td>
</tr>
</tbody>
</table>

1. In general, for all scenarios, participants from the experimental group had significantly less cost of resources when they switched from a wrong tentative diagnosis to a correct diagnosis (p = 0.002, 0.002, 0.006, and 0.097). This indicates that the decision support tool is able to optimize an efficient use of resources so that only those resources having the maximal capability of differentiating differential diagnoses will be selected.

2. For all scenarios, participants’ savings on the number of tests (p = 0.29, 0.003, 0.152, and 0.240) are less significant than the savings on the cost of resources.
This indicates that even though the tool may not significantly reduce the number of lab tests, it may still significantly reduce the cost of lab tests because it will recommend tests based on their level of usefulness.

3. However, the savings on the number of lab tests are still substantial, although they are not that significant. Since the number of lab tests that a patient needs critically determines his waiting time in a hospital, the tool is therefore able to potentially reduce patient waiting time. This potential capability of the tool will enable patients to receive a diagnosis and treatment as early as possible, as well as help hospitals alleviate over-crowdedness, especially in emergency departments.

4. The savings from both the number of lab tests and the cost of lab resources in single-diagnosis scenarios are more significant than those in double-diagnosis scenarios. This is again due to the fact that the double-diagnosis scenarios are more complicated and may cause decision-makers to order more tests to differentiate the two diagnoses.

5. The savings among the groups (nurses, residents, and physicians) have different levels of significance. The significance level for nurses is greater than that for residents, and the level for residents is greater than that for physicians. This observation may reflect different trust levels in the tool.

In summary, decision-makers may tend to order lab test believed to support their suspected diagnoses because of anchoring heuristics. This may lead to an inefficient use of resources since a suspected diagnosis could be wrong. The HDSB-enabled decision support tool is able to help a decision-maker determine the correct diagnosis with
significantly less cost, even though an initial diagnosis may be wrong. Therefore, the tool is potentially able to counteract the negative effects of anchoring heuristics by assisting a decision-maker in finding the correct diagnosis in a timely manner with significantly less expense.

7.4 Summary

Human biases play important roles in decision-making during medical diagnosis. They may affect the accuracy of diagnosis and the effectiveness of lab resources. The role of anchoring heuristics may be one of the reasons why misdiagnosis and excessive costs occur.

The HDSB-enabled decision support tool has been shown to have the capability of significantly improve diagnosis accuracy and reducing expense. This is enabled by its capability to counteract the negative effects of anchoring heuristics in two aspects:

1. The tool helped decision-makers identify a correct diagnosis more accurately at their first attempts. Quantitative measurements were used to rank differential diagnoses, so that a decision-maker’s initial diagnosis became more evidential than when not using a ranking. The analysis has shown that the tool significantly reduced the chance that a decision-maker made a wrong diagnosis initially.

2. Even though one’s initial diagnosis was incorrect, the tool also helped the decision-maker quickly find the correct diagnosis with significantly less use
of resources. This was enabled through the system’s ranking of lab tests based on their usefulness for distinguishing differential diagnoses.

In this chapter, the data analysis not only disclosed the potential usefulness of the tool to significantly improve diagnosis accuracy and reduce resource costs, but also disclosed how human’s anchoring heuristics and confirmation biases were counteracted, resulting in two significant improvements.
Chapter 8
DISCUSSION

SRCAST-Diagnosis has shown its capability of improving diagnosis accuracy and reducing resource cost in the experiment. The fundamental reasons for these two improvements were also disclosed through the experiment: anchoring heuristics and confirmation biases were counteracted by SRCAST-Diagnosis.

In this chapter, the implications of the experiment results will be further discussed. I will also discuss how the fundamental structure and procedure of MLBN and HDSB have made it possible to improve the diagnosis accuracy and resource efficiency. Because of the different performance of nurses, residents, and physicians, there could be other potential use of SRCAST-Diagnosis. Finally, the limitations of the study will be presented.

8.1 Implications of Experiment Results

8.1.1 SRCAST-Diagnosis and Cognitive Biases

Although they may be used interchangeably from time to time, a cognitive heuristic usually refers to a mental shortcut to the solutions of a decision problem, and a bias usually refers to the deviation from the normative rational theory caused by the underlying heuristics (Gilovich & Griffin, 2002). In other words, a heuristic may come into play if a decision problem becomes complicated and sophisticated and the decision-
maker is not capable of conducting the complicated and sophisticated calculations and inferences using principles from normative rational theory. For instance, when information is uncertain, a decision-maker may not always calculate the exact utility of each option and then choose the one with the highest expected utility. Instead, he may base his choice on his first impression (anchoring heuristics) or on his past experiences (availability heuristics) (Tversky & Kahneman, 1974), this is where a cognitive bias is from and a decision error may occur.

Medical diagnosis is a decision problem that is full of uncertainty: a symptom may or may not appear with the presence of a disease, or a symptom may be caused by many diseases. Therefore, cognitive heuristics such as anchoring heuristics and availability heuristics are applied by medical decision-makers very frequently, since a precise calculation of probabilities of each option is not only time-consuming, but also impossible for human brain to perform.

No matter what cognitive heuristics they may use in medical diagnoses, however, medical decision-makers still follow a certain procedure when diagnosing a patient: they need to generate differential diagnoses first based on the initial symptoms and physical examination results (hypothesis generation), and then come up with an impression how likely each option is (hypothesis evaluation); for the differential diagnosis that they suspect the most, they will order tests to see whether the suspected diagnosis is true (information seeking); whenever new symptoms are observed or new lab test results are obtained, they will revise the differential diagnoses accordingly (hypothesis revision), until the final diagnosis could be made.
This process was captured in the Hypothesis-Driven Story Building (HDSB) framework which generalizes the decision-making procedure with uncertain, missing, and changing information. The HDSB-featured SRCAST-Diagnosis, which was based on Multi-Layer Bayesian Network (MLBN) and probabilistic inferences, is able to support medical diagnosis. It assisted human decision-makers by offering recommendations based on precise calculations and inferences that a human brain is not always capable of.

The experiment showed that SRCAST-Diagnosis can significantly improve the diagnosis accuracy and reduce the cost of resources. It also disclosed that anchoring heuristics (one of the reasons of misdiagnosis) and confirmation biases (one of the reasons of unnecessary extra tests) were counteracted in those participants whose decisions were supported by the tool. In this section, I will further discuss the reasons why these improvements were possible.

One basic reason for the improvements was that SRCAST-Diagnosis increased decision-makers’ awareness of those important issues that may have not been highlighted as well in other systems:

**Ranked differential diagnoses**

Differential diagnoses were displayed to the users in the order of their likelihood (measured by entropy). SRCAST-Diagnosis highlighted those differential diagnoses that were highly probable so that a decision-maker would pay more attention to them, as well as listed those differential diagnoses with small probabilities so that they were not completely ignored or excluded. However, the ranking of differential diagnoses may be ignored or inaccurate without such a recommendation system. SRCAST-Diagnosis
reminded decision-makers to be aware of the differential diagnoses in terms of their likelihood.

**Timely update of differential diagnosis**

A diagnosis may be delayed or missed by a delayed or missed update of a patient situation. SRCAST-Diagnosis was able to capture the significant changes of patient conditions and update the ranked differential diagnoses automatically according to these changes. When the content or ranking of differential diagnoses was updated, the user was then reminded to think about the update and the underlying factors that had caused this update. Therefore the risk of missing the correct diagnosis was lowered.

**Ranked recommendations of lab tests**

To further confirm a suspected diagnosis, different lab tests need to be ordered. The usefulness of each lab test to the suspected diagnosis may vary according to their sensitivity and specificity. SRCAST-Diagnosis listed the possible tests in the order of their usefulness in differentiating the differential diagnoses and their costs. Therefore, a clinical decision-maker can request the lab tests in a prioritized order and reduce the extra requests of unnecessary tests, with the value and cost of the lab tests kept in mind.

**8.1.2 Other Types of Cognitive Biases in Experiment**

Although anchoring heuristics and confirmation biases were proved to be associated with the misdiagnosis and inefficient use of resources in the experiment, the possible biases of a decision-maker are not only limited to these two. The discussion in
this section is about other types of cognitive biases in diagnostic decision-making, although they were not the focus of the experiment.

8.1.2.1 Premature Closure

The premature closure bias is the tendency to offer only a single solution to a problem that logically allows multiple solutions, due to insufficient or ambiguous information (Acredolo & Horobin, 1987). The decision-maker may stop evaluating and searching for alternative solutions even if they exist. In diagnostic decision-making, a decision-maker may stop evaluating and searching for other differential diagnoses, and thus end up with the initially suspected one. It is critically important for a decision-maker to know when to stop, which is oftentimes a matter of experience.

The premature closure bias can explain the interesting phenomenon in the experiment: the nurse group performed the best in reducing lab resources but worst in improving the diagnosis accuracy. The tool did not help improve their diagnosis accuracy. This means that they stopped ordering lab tests much sooner than did the resident group and physician group, at a time when more lab tests were still needed. The premature closing of searching for additional information and evaluating other differential diagnoses was the cause that led to the nurse group’s low diagnosis accuracy performance.

The premature closure bias is highly associated with a decision-maker’s experience in making such decisions. In the experiment, this bias can be observed clearly within the nurse group, as it is usually the group with less experience in making diagnostic decisions. The nurses just simply followed what the decision support tool
recommended, but they were not completely able to tell when to stop the evaluating and searching procedure. Therefore, to overcome the premature closure biases is not something that a decision support tool such as SRCAST-Diagnosis can help. It requires the decision-maker to be well trained and knowledgeable enough to make such decisions, even with decision support tools.

**8.1.2.2 Availability Heuristics**

Availability heuristics is the tendency of a decision-maker to predict an event if an instance of it can be easily brought to mind (Tversky & Kahneman, 1974). Since it is dependent on the availability of remembered instances, it is essentially associated with one’s memory and experience in making similar decisions. RPD is one of the decision-making mechanisms that will potentially generate availability heuristics, since, according to RPD, a decision-maker may search for a workable solution to a decision problem by comparing the current situation with previous experienced situations. Therefore, what is contained in one’s experience base significantly determines the solutions one may come up with, which is the source of availability heuristics.

In the experiment, different decision-makers initially suspected different diagnosis. Each initial diagnosis was very likely associated with past experiences. Although the experiment did not focus on uncovering the causal relationship between a decision-maker’s experience and availability heuristics, it is reasonable to hypothesize their relationship for future testing.
8.1.2.3 Overchoice

Overchoice is a term that describes the phenomenon that a consumer suffers from making a selection if he faces too many choices (Simon, 1995). A consumer may have troubles making choices from among a variety of products, which will make the decision-maker indecisive, unhappy, and even end up with no decisions made at all (Settle & Golden, 1974).

Although overchoice has been primarily used to describe the phenomenon of consumer behavior, it can also be used to explain the choice-making behaviors in other domains, such as medical diagnosis. For example, a physician may hesitate to make a diagnostic decision if he faces too many differential diagnoses.

In the previous chapter, a phenomenon termed over-reminding was presented. Scenario S3 presented a situation with myocardial infarction as the only true diagnosis. The results showed that the experimental group was more likely to select two diagnoses (myocardial infarction (MI) and cardiogenic pulmonary edema (CPE)) than was the control group. Since CPE is a potential effect of MI, the decision support tool listed CPE as a differential diagnosis accompanied by MI. This reminded the decision-makers to think about the possibility of having two diseases simultaneously. Unable to eliminate the possibility of having either disease, the option was to select both diseases. If no list of differential diagnoses had been provided, other options may not have occurred.

A decision support tool is able to help decision-makers quickly differentiate the differential diagnoses by listing them with a ranking order. However, listing many other differential diagnoses may also potentially overload decision-makers with too many
choices, making their decision-making indecisive and even incorrect, while the decision may be decisive and correct if there are not so many choices to consider.

8.2 MLBN: Representing Probabilistic Causal Knowledge with Influencing Factors

SRCAST-Diagnosis has shown its usefulness in improving diagnosis accuracy and reducing resource costs. Its usefulness is based on the fundamental MLBN that represents probabilistic causal knowledge with influential factors.

*It enabled probabilistic abductive reasoning.*

In the design of SRCAST-Diagnosis, the first problem that needs to be solved is hypothesis generation, where a hypothesis is a differential diagnosis in medical situations. Classical Bayesian networks are able to conduct probabilistic inferences. However, the conditional probabilistic relationship between nodes does not differentiate the causal relationship with other influential factors, and therefore is not able to be used directly to conduct abductive reasoning, i.e. to find the causes of observed information. MLBN alters the Bayesian network structure, allowing the linkage between nodes to only reflect the causal relationship. Therefore, the causes of observed information can be traced upward in the MLBN.

*It enabled probabilistic reasoning for n-ary relations.*

When separating the causal relationships, other influencing factors are collapsed into the internal structure of each node. Therefore, a node of an MLBN can represent a complicated n-ary relation, which can be expressed in predicates or functions. This feature enables MLBN to be connected with predicate logic which is usually used for
logic knowledge representation. Therefore, the probabilistic reasoning for n-ary relations became possible.

*It enabled flexible partition of the n-dimensional space of n-ary relations.*

The MLBN collapses the influencing factors other than the causal factors into the internal structure of MLBN nodes and therefore allows multiple layers of probability tables for each node. In a classical Bayesian network, for each node, the conditional probability of the node to be a specific value needs to be given based on all possible combinations of the values from parent nodes. However, in an MLBN, the conditional probabilities coming from the influencing factors are separated into different layers. Each layer is defined by a variable binding that is a description of a set of conditions. This offers a more flexible way to represent the conditional probability and a more efficient way to find the appropriate probabilities for a specific inference.

In summary, as a primary means to conduct abductive reasoning, MLBN collapses the influencing factors other than causal factors into the internal structure of each node and makes the probabilistic inferences of n-ary relations possible. The representation of multiple layers of probability tables further makes it flexible and efficient to make the probabilistic inferences for hypothesis generation and evaluation.

### 8.3 HDSB: Modeling Human Decision-Making with Partial Information Arriving Over Time

HDSB was proposed mainly to describe the decision-making process of a decision-maker especially when information is only partially known and is arriving over
time. It is not only a model of describing the actual process of human decision-making under such environment, but also a framework for designing a decision support system for decision-makers.

As a model of describing the process of human decision-making with partial information arriving over time, HDSB captures the major steps of human decision-making. When initial information is only partially known, the hypothesis generation describes the step in which hypotheses are generated. Each hypothesis is a possible explanation of the partial information. Then the generated hypotheses are evaluated based on their likelihoods via the hypothesis evaluation step. Various criteria may be used to evaluate the hypotheses. Then based on the hypotheses and their ranking, strategies of seeking information will be determined in the hypothesis-driven information seeking step. Oftentimes, tentative decisions may be made to see the situation will change as expected. All the new information will trigger the hypothesis revision step, in which the hypotheses will be updated according to the new information, either the content of the hypothesis space or the ranking of the hypotheses, until a final decision is made.

Both human decision-makers and computer systems follow the process described in HDSB. The difference between a human decision-maker and a decision support system is the criteria used for each step. Therefore, as a framework for designing a decision support system, new criteria were developed for story building, especially for clinical diagnosis.

**Generating hypothesis (differential diagnosis)**

Hypothesis generation was based on algorithms that find the hypotheses given a set of observations. The algorithms were proposed based on the structure and inferences
of MLBN. Since in MLBN the edges represent the causal relationships between the two nodes, the goal of hypothesis generation is therefore to find out the top nodes that represent the ultimate reasons of the observations. The hypothesis generation algorithms utilized the definition of explanandum, evidence, explanans, and their causal relationships.

**Evaluating hypothesis (differential diagnosis)**

The ranking of each differential diagnosis should be based on its probability. However, because a differential diagnosis represented in an MLBN node may have multiple observation values, the ranking of the diagnosis should be the ranking of the node with the probabilities of all its values. In general, a node that is more certain on one value should get more attention from decision-makers. In other words, a node whose probabilities of values are more evenly distributed should be ranked higher. Therefore, information entropy can be used to measure the distribution of probabilities.

This measure needs to be slightly revised to rank differential diagnosis, since the values of a diagnosis node are not equally important. The values of a medical condition are always composed of one normal value and one or more abnormal values. A clinical decision-maker will only focus on those abnormal values. Therefore, the ranking method should also reflect the difference between the normal value and the abnormal values. Ultimately, two criteria were used to rank a clinical condition node: (1) a condition that has lower probability on the normal value should be ranked higher; and (2) a condition whose probabilities of abnormal values are more unevenly distributed should be ranked higher. A revised measure was then given: $RI(H) = (1 - p_0) (1 - H(X))$, where $H(X)$ is the entropy of the node $X$. 
Hypothesis-driven information seeking (ranking lab tests)

To further evaluate a differential diagnosis, lab tests need to be ordered. The ranking of lab tests should be based on their cost and usefulness. A lab test will potentially return one of the n possible values. Each returned value will change the probability distribution of the diagnosis. Therefore, the change of the probability distribution of the diagnosis can be used to measure the usefulness of a test with a returned value, which can be measured by reduction of entropy. Then because it is not certain which value will be returned, we need to consider all the possible values with probabilities. Therefore, an expected usefulness based on test values can be used to measure the usefulness of a test. If taking test cost into consideration, the value of test per cost unit can be used to measure the effectiveness of a test.

Revising hypothesis (updating differential diagnosis)

When new information arrives, the list of differential diagnoses need to be updated, either with the content or with the ranking. For each node that represents a clinical condition, its values belong to three categories: normal, abnormal, and unknown. Therefore, for each single change of a node, it is important to know the direction of the change. For example, a node may be changed from a normal value to an abnormal value. Five types of changes were identified: unknown to normal, unknown to abnormal, normal to abnormal, abnormal to normal, and abnormal to abnormal. For each of these changes, the strategy for revising the content and the ranking should be different.

These steps of HDSB have been implemented in SRCAST-Diagnosis. The experiment has shown that SRCAST-Diagnosis and the methods implemented in it can
potentially assist clinical decision-makers to improve diagnosis accuracy and resource efficiency.

### 8.4 Potential Use of SRCAST-Diagnosis

SRCAST-Diagnosis has shown its capability of improving diagnosis accuracy and cost efficiency in the experimental setting. However, the levels of these improvements were different for different role players (nurses, residents, and physicians). This finding reflected their different levels of knowledge and their different levels of interactions with the system. Therefore, SRCAST-Diagnosis may be potentially used for different purposes.

1. Diagnostic decision support

   As suggested by the experiment, SRCAST-Diagnosis can help a clinical decision-maker improve the diagnosis accuracy and cost efficiency. Among the three role players, residents got the high accuracy improvement and high cost saving, while nurses and physicians had much lower improvements on diagnosis accuracy although their savings on cost were still significant. Therefore, SRCAST-Diagnosis can be potentially used as a diagnostic decision support tool, especially for residents who have acquired sufficient medical knowledge but still need recommendations.

2. Training and education tool

   SRCAST-Diagnosis can also be used as a training tool for nurses or medical students who are still learning their medical knowledge. A user can be asked to go through the scenarios and make the diagnoses by themselves first, and then compare his decisions with the recommendations made by SRCAST-Diagnosis. It will not only show
the likelihood of each differential diagnosis by giving the probability, but also explain how the differential diagnosis was made by listing the evidence facts. For the lab tests, SRCAST-Diagnosis will rank the lab tests by showing its usefulness in changing the probability distribution of the differential diagnosis.

8.5 Study Limitations

There were a set of limitations of the study because of the constraints of time and cost.

*Participant sample size was relatively small.*

In the experiment, 37 participants were recruited and divided into two groups. These participants were further divided into nurses, residents, and physicians. Each participant was asked to go through 10 scenarios so that the findings were cross validated from scenarios’ perspective. Therefore, the findings are valid and valuable to provide insights to the research questions. However, the sample size of participant is still relatively small. Therefore, it is desirable to conduct experiment on large sample size in the future.

*Clinical knowledge base was not based on patient database.*

The MLBN used in the experiment reflects a small portion of medical knowledge. It was built based on the method of literature search and knowledge engineering. The conditional probabilities that reflect the causal relationships between diseases and symptoms, physical examination results, and lab results are obtained from experts’ estimates.
Although some verification has been done to the reliability of knowledge base, it only reflects the knowledge and the experience of a few physicians. Consequently, the knowledge base and the experiment result may be biased by these experts. Therefore, a database that contains a large amount of real patient data is always wanted, so that the probabilities can be learned through the real historical patient data. This will make the knowledge base more reliable.

*Simulated scenarios instead of actual patient cases were used.*

The scenarios used in the experiment are simpler versions of real patient cases. This is partly because the design of scenarios is constrained by the level of knowledge base. A scenario must not go beyond the level that the agent is able to process using its knowledge base. Therefore, the scenarios may look a little bit artificial to the decision-makers.

First, each scenario contains the key components of a patient case, such as symptoms, physical examinations, and lab tests. However it lacks more detailed information such as patient medical history, which may be subtle but still important to make a diagnosis.

Second, the interpretations of lab test results may look artificial. A lab test may not only return a quantitative result or a qualitative result that has been classified into a category, but also send back a textual sentence or a radiology image that needs to be interpreted. However, the interpretation may be complicated and difficult. Since the system currently does not incorporate the capability of natural language processing and image processing, the interpretations of such information are now hard coded.
With the future development of the system, scenarios will need to be more realistic until real patient cases can be used in the system.

**Clinical history or picture is needed.**

A real patient case may not simply start from a list of symptoms and physical examination results. It may include other subtle information such as past medical history, family history, alcohol/tobacco use, and drug allergy. These altogether depict an overall clinical history or picture of a patient.

The scenarios used in the experiment primarily focused on patient diagnosis starting with symptoms, but lacked enough information for a decision-maker to establish an overall picture of the patient. This would make some participants feel difficult to make the diagnosis.

For future development, scenarios will be considered to contain the patient history and picture with certain level of details.

**Need better timing to display new information.**

The scenarios used in the experiment are all dynamic, i.e. newly updated information will potentially change the content or the ranking of differential diagnoses. At the same time, the new information will be displayed to the decision-maker by a pop-up message. Since the new information will change the current differential diagnoses, the content and the ranking of the recommended lab tests also need to be updated. Therefore, a decision-maker will be forced to restart the selection of suspected diagnosis and relevant lab tests.

However, this feature made some participants feel frustrated, since their efforts may be canceled because of the new update. Whenever there is a new update, it is
expected not only that the information is displayed to the decision-maker in a timely manner, but also that the information is displayed to the decision-maker at a good timing, so that the decision-maker will have other options to integrate the new information into current thinking without having to restart it from the very beginning.

8.6 Summary

In this chapter, I further discussed the remaining issues of the system and the experiment. It is considered that the features of SRCAST-Diagnosis increased decision-makers’ awareness of those important issues that may have not be highlighted in other systems, which further counteracted the negative effect of cognitive biases and finally improved the performance of decision-makers. The features of SRCAST-Diagnosis were implemented based on MLBN and HDSB.

Except for diagnosis support, SRCAST-Diagnosis can also be potentially used as a training and education tool. This was implied from the observation that SRCAST-Diagnosis worked differently for nurses, residents, and physicians. The usefulness of the tool was significantly limited by the knowledge level of decision-makers. Therefore, it is potentially useful as a training and education tool for nurses and medical students who are learning medical knowledge.

Although the experiment results have shown its usefulness in increasing diagnosis accuracy and resource efficiency, there were still certain limitations such as sample size and reliability of knowledge base. These limitations are expected to be addressed in future work.
Chapter 9
CONCLUSION

Misdiagnosis and cost have been two of the major issues guiding healthcare reform and two of the major goals of healthcare reform. However, it is challenging to reduce misdiagnosis rate and reduce resource cost simultaneously, since reducing misdiagnosis rate requires increasing lab resources to narrow down differential diagnoses in many cases. However, the increase of lab resources may not be necessary for patient health, but may be because of other factors, such as cognitive biases. Therefore, a set of research questions were proposed to explore the possibility of improving diagnosis support by reducing misdiagnosis rate and resource cost at the same time with factors of cognitive biases taken into account.

A framework called Hypothesis-Driven Story Building (HDSB) was proposed to describe a decision-making process with partial information arriving over time. The HDSB framework was especially used to describe a clinical diagnosis process with patient conditions partially arriving to the decision-maker. When given partial information, HDSB will generate a list of hypotheses that can explain the current information and rank the hypotheses based on their likelihood. If more information is needed, HDSB will recommend a list of possible actions for seeking the missing information, which will be ranked based on their usefulness of differentiating the potential hypotheses as well as their cost. Whenever new information arrives, the HDSB framework will update the potential hypotheses accordingly.
The HDSB framework was built based on Multi-Layer Bayesian Network (MLBN), which is an extension of standard Bayesian network with relational representation on each node, allowing the probabilistic causal inferences for relations based on variable bindings. In MLBN, different conditional probability tables can be defined for different variable bindings so that the Bayesian inferences can be specialized or personalized.

A web-based clinical diagnostic decision support tool was developed based on the HDSB framework. Given a scenario, the tool will display the patient conditions, recommended differential diagnoses, and ranked lab tests to the user, and will ask the user to order lab tests or to make the final diagnosis.

The tool was evaluated by a controlled experiment conducted at the Hershey Medical Center. Participants included nurses, residents, and physicians, divided into a control group and an experimental group. Their behaviors (ordering lab tests and making diagnoses) were recorded. Also, their tentative diagnoses were traced.

The collected data showed that the tool can significantly improve the diagnosis accuracy and reduce the cost of resources overall, although the performances for different role players may vary. The data also showed that the tool helped more decision-makers who made the wrong initial diagnosis eventually find the correct diagnosis, counteracting anchoring heuristics. Moreover, the tool helped them quickly figure out the correct diagnosis with significantly less resource cost, counteracting confirmation biases.

This briefly summarizes my study in this dissertation. In the remainder of this chapter, I outline the contributions of my study and discuss future work.
9.1 Answering Research Questions

In the first chapter, I raised four research questions. These questions outlined the theme of this dissertation. They are now answered in this section.

**Question 1: What model can be developed to describe a physician’s hypothesis reasoning and information seeking process in clinical diagnosis?**

When information is partial and changing and when resources are limited in seeking information, a decision-maker will usually have several methods for dealing with these challenges. He may generate a list of hypotheses, given that the information is only partially known. Then he may evaluate the likelihood of each hypothesis with the available information. When more information is required, he will take actions to obtain the information. Whenever new information arrives, he will update the hypotheses and start the next round of evaluation, until a final choice can be made. Hypothesis-Driven Story Building was therefore developed to describe the decision-making process when information partially arrives over time, such as physicians’ making diagnostic decisions. HDSB not only describes the decision-making process of a human decision-maker, but also works as a model for developing decision support systems. A tool based on HDSB was developed to support clinical diagnostic decisions. The only difference between a human decision-maker and a decision support system is their criterion in each step. A human decision-maker may be constrained by his or her cognitive capabilities and therefore may be biased in decision-making, while a decision support tool will base its decision on sophisticated calculations and inferences that a human decision-maker may not be able to do.
Question 2: Can a decision support tool, built based on the proposed model, reduce misdiagnosis?

Data from a controlled experiment showed that the diagnosis accuracy of the experiment group is significantly higher than that of the control group. The resident group more significantly improved their diagnosis accuracy than the nurse group and the physician group. The p-value of the improvement of resident group was 0.004. The improvement of diagnosis accuracy also varied for different scenarios. Decision-makers (especially the residents) benefitted more in improving diagnosis accuracy for double-diagnosis scenarios than for single-diagnosis scenarios. This was due to the two special single-diagnosis scenarios which contributed less to the improvement of diagnosis accuracy. However in other scenarios, the improvement of diagnosis accuracy by the decision support tool was still substantial.

Question 3: Can the same tool reduce cost of lab tests while reducing misdiagnosis?

Data from the experiment showed that the cost of lab test resources was significantly reduced by use of the tool. The resource cost was saved 64% on average, and the savings were very significant with a \textit{p-value} of 0.000. Among the three clinical groups, the nurse group averaged the greatest savings, and the physician group averaged the least. Between the two types of scenarios, the degrees of savings are significantly different. The cost savings in single-diagnosis scenarios are much more significant than those in double-diagnosis scenarios. This is because the complexity of double-diagnosis scenarios required decision-makers to spend more resources and get more information to differentiate the two diagnoses. However, the cost savings on these scenarios are still
substantial, and the savings on all scenarios are significant on average. Therefore, the experiment showed that the misdiagnosis rate and the resource cost can be both reduced significantly. Therefore the misdiagnosis-cost dilemma can be potentially solved.

**Question 4: Are these improvements caused by the counteraction of cognitive biases?**

Anchoring heuristics refer to the tendency of relying too heavily on one trait and being reluctant to part with it. Anchoring heuristics are one reason for misdiagnosis, since decision-makers may be reluctant to consider a correct diagnosis even though an initial diagnosis is clearly incorrect. Confirmation bias refers to the tendency to favor information that may support a hypothesis. Confirmation biases may cause a decision-maker to spend more on lab tests which are less useful but which are favored by the decision-maker. The experiment data showed that with the support from the tool, more originally-errant decision-makers eventually made the correct diagnosis, counteracting anchoring heuristics, and spent significantly resources determining the correct diagnosis, counteracting confirmation biases. Therefore, the experiment showed the evidence that the counteraction of cognitive biases were the reasons that the decision support tool reduced the misdiagnosis rate and the cost at the same time.

These questions laid out the importance of the study to academic community. In the following section, I will discuss the contributions that this study has made to academic communities.
9.2 Contributions

This study is an interdisciplinary one that develops technologies and implements systems to support diagnostic decision-making. Therefore, there are two major communities to which this study has made contributions: the AI community and the medical informatics community.

9.2.1 Contributions to the AI Community

Artificial intelligence (AI) is an area in which automatic and intelligent methods are developed. This study has contributed to the AI community from the following perspectives:

*It integrated logic-based reasoning and probabilistic reasoning in a flexible framework.*

In the AI tradition, knowledge representation and inference is one of the key topics. Various methods have been developed to make knowledge representation and inference more powerful and more efficient. In general, there have been two major methodology streams of knowledge representation and inference: logic-based and probability-based. Logic-based methods use rule-based systems, where knowledge is represented in rules, and inference is conducted based on logic relationships. Probability-based methods deal with uncertainties, where knowledge and its inference are both probabilistic. The Bayesian network is among the core probabilistic methods used to represent uncertain knowledge and make probabilistic inferences.
However, both methods have limitations in solving real problems. The logic-based methods lack the flexibility to handle uncertain and probabilistic knowledge, and the standard Bayesian networks lack the flexibility to make probabilistic inferences based on object level. Many efforts have been made to empower either the logic-based methods or the Bayesian networks with ideas from each other (Nilsson, 1986; Poole, 1993; Jaeger, 1997; Pfeffer, 2000; Domingos & Richardson, 2004; Raedt & Kersting, 2004; Braz, Amir, & Roth, 2005; Laskey, 2006; Kersting & Raedt, 2007).

The Multi-Layer Bayesian Network developed in this study is another attempt to integrate first-order logic representation with Bayesian networks. A node in the Bayesian network can now represent a relation or a function. Moreover, the conditional probability tables associated with each node can be multiple layers, where each layer is defined by a special constraint or requirement of variable bindings. Therefore, the MLBN is able to make inferences with specific probabilities based on personalized or specialized information. This new feature makes the Bayesian inference more powerful in making probabilistic relational inferences with personalized or specialized requirements.

*It enabled abductive reasoning within Bayesian structure.*

In classical Bayesian networks, a node can represent any variable and a link between two nodes can represent any conditional relationship, but not limited to causal relationships. There is no distinction between the causal relationships and other influencing factors from how a Bayesian network is constructed. Therefore, it is difficult to make abductive reasoning, i.e. to find out the causes of observed information, with the classical Bayesian structure.
The Multi-Layer Bayesian Network further distinguishes the causal relationship and other influencing factors, and uses links between nodes to only represent their causal relationships. Other influencing factors collapse into the internal structure of a node. Therefore, an MLBN is able to represent the causal relationships for complicated nodes, such as function and predicates which have their internal structures defined by variable bindings. Therefore, abductive reasoning for nodes with complicated structures was enabled in MLBN.

*It proposed value-aware information seeking strategy.*

In traditional clinical decision support systems, recommendations on information seeking activities such as lab tests were not usually given. For those systems that have such recommendations, only the names but not their rankings were given. In this study, a value-aware information seeking strategy was proposed, trying to rank the possible lab tests in their usefulness to narrow down the range of differential diagnoses. The value of a test was defined by the expected entropy loss that the test would bring to the suspected differential diagnosis. Furthermore, the effectiveness of a test was defined by the average value per cost unit, if the cost of the test is taken into consideration. In this way, a decision-maker will be aware of the value and the cost of each lab test and choose the ones that have significant influence on the differential diagnosis based on a quantitative rationale.

*It enriched a computational RPD model with a story building component.*

RPD was originally proposed as a cognitive model that mainly describes how human experts make decisions with time stress. The core component of experience matching in RPD has been computationally implemented from different perspectives
(Warwick, McIlwaine, Hutton, & McDermott, 2001; Fan & Yen, 2007). However, the story building component is missing in these implementations. Therefore, it enriched a computational RPD model with a story building component.

9.2.2 Contributions to the Medical Informatics Community

Medical informatics is the area that studies the applications of computer technologies in medical care. This study contributed to the medical informatics from the following perspectives:

*It developed a new clinical diagnostic decision support framework.*

In medical informatics, understanding misdiagnosis and designing and implementing computer systems to reduce misdiagnosis and improve patient safety has always been a central goal. The typical features and process of clinical diagnostic decision-making were captured in the HDSB framework which was also used for designing clinical diagnostic decision support prototype. This prototype highlighted three major challenges of a clinical diagnostic decision-making: (1) information is only partially known; (2) information has cost; and (3) information can change over time. These challenges were addressed in the HDSB framework through four major steps: (1) hypothesis generation, (2) hypothesis evaluation, (3) hypothesis-driven information seeking, and (4) hypothesis revision. An implementation of a new decision support prototype supported the medical diagnostic making in a new way.

*It demonstrated its potential to improve diagnosis accuracy and reduce resource cost at the same time.*
Unlike earlier decision support systems which mostly focused on one aspect of diagnosis accuracy and resource efficiency, SRCAST-Diagnosis has shown its potential to improve diagnosis accuracy and reduce resource cost at the same time through the experiment. HDSB modeled doctors’ diagnostic and lab ordering decisions in a general framework. SRCAST-Diagnosis is then able to provide diagnostic and lab ordering recommendations based on an optimized method, where differential diagnoses are listed in the order of their likelihoods, and lab tests are listed in the order of their usefulness to further differentiating the differential diagnoses. Therefore, a doctor supported by the decision support tool may figure out the correct diagnosis using fewer resources for lab tests.

*It investigated the potential relationship between diagnosis quality and cognitive biases.*

This study not only demonstrated its potential usefulness to improve diagnosis accuracy and resource efficiency, but also investigated the relationship between the improvements and counteraction of cognitive biases. By offering differential diagnoses with certain ranking criterion and recommending relevant lab tests based on their usefulness to the suspected differential diagnosis, SRCAST-Diagnosis increased the awareness of those important issues that may have been ignored by decision-makers before. This is the basic reason why a decision-maker’s cognitive biases may be counteracted. The experiment has shown that: (1) decision-makers were more probable to find out the correct final diagnosis even though their initial tentative diagnosis was incorrect so that their anchoring heuristics were counteracted, and (2) decision-makers
were more probable to spend less resource to find out the correct final diagnosis so that their confirmation biases were counteracted.

9.3 Future Work

Although HDSB has shown its contribution to the AI community as well as the medical informatics community, there are other related important topics that have not yet been addressed. In this section, I outline several topics that may lead to future work.

9.3.1 Large-scale Experiment

Although the experiment has provided useful and important insights, the sample size of the experiment is still relatively small. A large-scale experiment is expected. This not only requires to extend the overall sample size, but also requires to extend the roles of participants. If possible, experiments within other department and other hospitals are expected.

9.3.2 Extension to Collaborative Settings

The current implementation of SRCAST-Diagnosis assumes a single decision-making involved in the entire decision-making process. This is not the case in real hospital settings. A clinical diagnostic decision-making process is usually a collaborative activity that involves physicians, residents, nurses, and technicians. Therefore, it is expect
to extend SRCAST-Diagnosis to the collaborative settings and allow collaborative teamwork.

**9.3.3 Implementation on Mobile Devices**

Another important observation on clinical decision-makers is that they are not always sitting in front of a computer. Therefore, there is a need to implement SRCAST-Diagnosis on mobile devices, so that a decision-maker will be able to input information/requests and receive information/recommendations in a timely manner. The web-based mechanism applied in developing SRCAST-Diagnosis makes it easy to be implemented on mobile devices.

**9.3.4 HDSB with Data Mining**

The knowledge base currently used by the existing HDSB decision support tool was built based primarily on a literature search and several expert estimates. The knowledge engineering method may work, but it has limitations. First, the method will only work for small scale problems. Once the problem space becomes large and involves thousands of nodes, it will become very time-consuming to extract knowledge from domain experts, and practically infeasible. Second, the knowledge may already have been biased by the experts, since their views are also limited.

Therefore, there exists a need to construct a more reliable knowledge base in an automatic way. Fortunately, because of the application of EMR (Electronic Medical
Records) systems, years of patient data have been recorded and are ready for research use.
There may be an opportunity to run data mining methods on the EMR system and obtain more reliable patterns of knowledge in a systematic and automatic way.

9.3.5 HDSB with Natural Language Processing

In medical decision-making, much information is saved in free text with no structure constraint. This is because free text has the flexibility to express ideas and opinions, and to communicate information and decisions. For example, a nurse’s note may record her impression about a patient’s appearance and her conversation with the patient or his family; or the radiology image may be interpreted with natural language sentences. This information is not able to be processed by the HDSB tool at present, since no natural language processing component has been incorporated into it. Therefore, the tool must have a natural language processing capability in the future so that it can be used for real patient scenarios.

9.4 Closing Remarks

One goal of the healthcare reform is to bring health care coverage to a broader population, which requires cost savings without compromising patient safety. This research work was aimed at addressing the dilemma of misdiagnosis and cost. HDSB was proposed as a framework to model physicians’ activities in making diagnostic and testing decisions. It was found that certain human cognitive biases which are the reasons for
misdiagnoses and ineffective use of resources can be counteracted by the decision support tool SRCAST-Diagnosis that was built based on the proposed HDSB framework, and therefore diagnosis accuracy and resource efficiency can be improved. It will be potentially useful to contribute to the health care reform and benefit every individual.
Bibliography


Appendix A

Diseases and Related Content

A.1 Viral Pneumonia (ICD-9: 480)

Description:

An inflammatory illness of the lung. Frequently, it is described as lung parenchyma/alveolar inflammation and abnormal alveolar filling with fluid (consolidation and exudation).

Risk factors:

- Virus infection

Possible Symptoms:

- Cough
- Sputum (yellow or green)
- Shaking chills
- Shortness of breath
- Chest pain: Sharp or stabbing pain
- Sweaty and clammy skin
- Cyanosis
- Nausea

Physical Examinations:

Vital signs

- Temperature;
  - > 100F (37.8C)
- Blood pressure
  - low
- Pulse:
  - Increased (>100/min)
- Respiratory rate:
  - Increased: tachypnea
- Saturation of oxygen:
Physical exams

- Rales or other abnormal breathing sounds may be heard
- Decreased breath sound
- Percussion dullness
- Rhonchi
- Egophony (Increased vocal resonance)

Lab Tests:

- Pulse oximetry
  - Reveals hypoxemia
- ABG
  - pO2:
- Blood tests
  - CBC
    - High white blood cell count: presence of an infection or inflammation.
    - Low blood sodium: extra anti-diuretic hormone produce when the lungs are diseased
- Urine test
- Liver function tests
- X-ray
  - Opacity represents consolidation
  - May see dense, lobar, fluffy, patchy, or diffuse infiltrates
  - There maybe an associated pleural effusion
  - misleading
- CT scan

A.2 Pulmonary Embolism (ICD-9: 415.1)

Description:

- Blockage of the pulmonary artery or one of its branches, usually occurring when a deep vein thrombus (blood clot from a vein) becomes dislodged from its site of formation and travels, or embolizes, to the arterial blood supply of one of the lungs.

Risk factors:

- Trauma
- Obesity
Possible Symptoms:

- Difficulty breathing (dyspnea, shortness of breath, SOB) (73%)
- Chest pain (66%)
- Palpitations
- Diaphoresis (36%)
- Cough (37%)
- Cough up blood (hemoptysis, haemoptysis); (13%)
- Cyanosis (19%)
- Collapse
- Syncope
- Wheezing
- Leg swelling
- Circulatory instability (shock)
- 32% have lower extremity edema
- Pallor

Physical Examinations:

Vital signs

- Temperature;
  - Less than 39C may be present in 14% of patients;
  - Higher than 39.5C is not from PE.
  - 43% have fever (temperature > 37.8C)
- Blood pressure
  - hypotension
- Pulse:
  - Rapid heart rate (tachycardia, >100/min) (30-44%)
- Respiratory rate:
  - Rapid breathing (tachypnea >16/min) (70-98%)
- Saturation of oxygen:
  - Hypoxia (Low blood oxygen saturation)

Physical exams

- Diminished breath sounds
- Crackles (51-58%)
- Pleural rub
- JVD
- Fourth heart sound (24%)
- Accentuated pulmonic component of the second heart sound (23-53%)
- Percussion dullness
34% have S3 and S4 gallop
23% have a cardiac murmur
Ankle edema

Lab Tests / medical imaging

- PT
- aPTT
- TT
- blood test
  - Electrolytes (sodium, potassium)
  - Markers of renal function (creatinine, urea)
  - Erythrocyte sedimentation rate
  - Liver enzymes
  - CBC (Complete blood count)
  - D-dimer
    - Non-specific
  - ABG:
    - Increased A-a gradient (maybe normal in up to 15%)
- EKG
  - Limited diagnostic value of PE
  - Sinus tachycardia (8-69% of people with PE were found sinus tachycardia)
  - Right axis deviation
  - Right bundle branch block
- Echocardiogram
  - In massive and submassive PE, dysfunction of the right side of the heart can be seen on Echocardiogram
  - The appearance of right ventricle is referred to as the McConnell sign.
    - Sensitivity on PE: 77%
    - Specificity on PE: 94%
- Pulmonary angiography
  - Gold standard (to find the clot)
- X-ray
  - Are often done on patients with shortness of breath to help rule-out other causes, such as congestive heart failure and rib fracture.
  - CXR in PE are rarely normal, but usually lack signs that suggest the diagnosis of PE (e.g. Westermark sign, Hampton’s hump).

- MRI
- CT scan
  - CT pulmonary angiogram (non-invasive)
    - Clinical equivalence,
    - Non-invasive
    - Greater availability to patients
    - The possibility of identifying other lung disorders from the differential diagnosis in case there is no pulmonary embolism
Ventilation/perfusion scan (V/Q scan or lung scintigraphy)
- This is used less often because of the more widespread availability of CT.
- It may be useful in patients who have an allergy to iodinated contrast or in pregnancy due to lower radiation exposure than CT.
  - Normal: rules out PE
  - High: diagnostic for PE

Ultrasound
- Leg Doppler in search of DVT. The presence of DVT is in itself enough to warrant anticoagulation, without requiring the V/Q or spiral CT scan
- The negative scan does not rule out PE

Plethysmography of the legs

Venography of the legs
- Positive: diagnostic
- Negative: does not rule out PE

A.3 Cardiogenic Pulmonary Edema (ICD-9: 514)

Description:
- fluid accumulation in the lungs.

Risk factors
- Congestive heart failure;
- Severe heart attack with left ventricular failure;
- Severe arrhythmias (tachycardia/fast heartbeat or bradycardia/slow heartbeat)
- Hypertensive crisis
- Pericardial effusion with tamponade
- Fluid overload, e.g. from Kidney failure or intravenous therapy

Possible Symptoms
- Difficulty breathing (dyspnea, shortness of breath, SOB)
- Accessory muscle use
- Cough (early clue)
- Hoarseness
- Chest pain
- Cough up blood (hemoptysis, haemoptysis);
- Excessive sweating (diaphoresis)
- Anxiety
- Pale skin (pallor);
- Mottling
- Pink frothy sputum;
- Nocturia
- Ankle edema (swelling of the legs, the skin is slow to return to normal when pressed upon)
- Orthopnea (in ability to lie down flat due to breathlessness)
- Paroxysmal nocturnal dyspnea
- Shock

Physical Examinations:

Vital signs

- Temperature;
- Blood pressure
  - hypertension
  - Increased pulmonary blood pressure from normal 15 mmHg to above 25 mmHg.
- Pulse:
  - tachycardia
- Respiratory rate
  - tachypnea
- Saturation of oxygen:
  - Low

Physical exams

- Elevated JVP
- End-inspiratory crackles (sounds heard at the end of deep breath)
- Rhonchi
- Wheezes
- Third heart sound (s3) for cardiogenic pulmonary edema
- Cardiac murmur or rub

Lab Tests

- Pulse oximetry
  - Reveal hypoxemia
- PT
- aPTT
- Electrolytes (sodium, potassium)
- Markers of renal function (creatinine, urea)
- Liver enzymes
- Inflammatory markers (C-reactive protein)
- CBC (Complete blood count)
- BNP (B-type natriuretic peptide)
  - Low level of BNP (<100 pg/ml) make a cardiac cause very unlikely.
- ABG (arterial blood gas)
  - Low pCO2: hypoxemia and respiratory alkalosis (due to tachypnea)
- Fick Principle (invasive, time-consuming, inaccurate)
  - To test cardiac output
- Dilution method (invasive)
  - To test cardiac output
- EKG:
  - Signs of ventricular hypertrophy
  - Atrial enlargement
  - Conduction abnormalities
  - Ischemia/infarction
- X-ray
  - Increased fluid in the alveolar walls
  - Mild congestion:
    - Cephalization of pulmonary vessels
    - Pleural effusion
    - Azygous vein enlargement
  - Worsen congestion:
    - Interstitial edema
    - Loss of distinct vascular margins
    - Alveolar infiltrates
    - Kerley B lines
  - Cardiogenic pulmonary edema:
    - Kerley B lines
    - Increased vascular filling
    - Pleural effusions
    - Upper lobe diversion (increased blood flow to the higher parts of the lung)
  - Noncardiogenic pulmonary edema:
    - Patchy alveolar infiltrates with air bronchograms
- Doppler Ultrasound Method (non-invasive, effective)
  - Echocardiogram
    - Cardiac output (low, normal, high)
  - Transcutaneous Doppler (USCOM)
    - Cardiac output
  - Transoesophageal Doppler (TOD) (requires patient sedation and is accepted for use only in adults and large children)
- Swan-Ganz Catheter (rarely used)
  - Pulmonary artery thermodilution (invasive, time-consuming, inaccurate, complex, dangerous, and expensive)
• Cardiac output (low, normal)

A.4 Myocardial Infarction (ICD-9: acute 410)

Description:

- Commonly known as a heart attack, occurs when the blood supply to part of the heart is interrupted causing some heart cells to die.

Risk factors:

- Older age
- Tobacco smoking
- Obesity

Possible Symptoms:

- Sudden chest pain (typically radiating to the left arm or left side of the neck)
- Shortness of breath
- Nausea
- Vomiting
- Palpitations
- Sweating (diaphoresis: excessive form of sweating)
- Anxiety
- Fatigue
- Feeling of indigestion
- Weakness
- Light-headedness / dizziness
- Loss of consciousness
- Cool and pale skin

Physical Examinations:

Vital signs

- Temperature;
  • Low grade fever (38-39)
- Blood pressure
  • low
- Pulse:
• irregular

◆ Respiratory rate:
  ○ increased

◆ Saturation of oxygen:
  ○ low

### Physical exams

◆ pulse rhythm: irregular
◆ Elevated JVP (jugular venous pressure)
◆ Elevated hepatojugular reflux
◆ Precordial examination: a cardiac bulge with a pace difference from the pulse rhythm can be felt.
◆ Rales or other abnormal breathing sounds may be heard
◆ Decreased breath sound
◆ Third and fourth heart sound
◆ Systolic murmurs
◆ Paradoxical splitting of the second heart sound,
◆ A pericardial friction rub
◆ Percussion dullness
◆ Increased vocal resonance

### Lab Tests

- autopsy
- Blood tests
  - Detect elevations in cardiac markers
    - Creatine kinase-MB (CK-MB) fraction
    - Troponin I (Tnl) or troponin T (TnT) levels
- EKG
- Coronary angiogram
- Echocardiogram
- X-ray
- CT scan

### A.5 Status asthmaticus (ICD-9: 493.01, 493.02)

#### Description:

- Asthma is a common chronic disorder of the airways that is complex and characterized by variable and recurring symptoms, airflow obstruction, bronchial hyperresponsiveness (bronchospasm), and an underlying inflammation. Status asthmaticus is an acute
exacerbation of asthma that does not respond to standard treatments of bronchodilators and corticosteroids. Symptoms include chest tightness, rapidly progressive dyspnea (shortness of breath), dry cough, use of accessory muscles, labored breathing and extreme wheezing. It is a life-threatening episode of airway obstruction considered a medical emergency. Complications include cardiac and/or respiratory arrest.

Risk factors:

- Tobacco smoke
- Poor air quality (traffic pollution, high ozone levels)
- Viral respiratory infections

Possible Symptoms:

- Nighttime coughing
- Shortness of breath with exertion but no dyspnea at rest
- A chronic ‘throat-clearing’ type of cough
- Wheeze
- Diaphoresis
- Clear sputum
- Prolonged expiration
- Over-inflation of the chest / accessory muscle use
- Cyanosis
- Chest pain / tightness
- Loss of consciousness
- Feel numbness in the limbs
- Diminished breath sounds
- Vocal cord dysfunction
- Vocal cord paralysis
- Thyroid enlargement

Physical Examinations:

Vital signs

- Temperature;
- Blood pressure
- Pulse:
  - Increased: tachycardia
- Respiratory rate:
  - Increased: tachypnea
- Saturation of oxygen:
  - Decreased: hypoxia
Physical exams

- Auscultation: Rhonchous lung sound
- Pulse rhythm: paradoxical

Lab Tests

- Pulmonary function tests:
  - Peak flow meter (measures airway obstruction)
    - 400-600: normal
    - 100-300: moderate exacerbation
    - <100: severe exacerbation
- ABG
  - Decreased pCO2: hypoxia and respiratory alkalosis
  - Normal or elevated pCO2: severe obstruction or respiratory fatigue with impending respiratory failure
- Capnography (measures the amount of exhaled carbon dioxide)
- Pulse oximetry (monitor the oxygenation of a patient’s hemoglobin)
  - pO2 monitoring
- Breath test (measure exhaled nitric oxide)
- Bronchial challenge test (measure bronchial hyperresponsiveness)
- Chest X-ray (mostly for exclusion of other pathology)
  - Required in new onset asthma, febrile patients, patients unresponsive to treatment, or those with a suspicion of pneumonia or other lung pathology
  - May show hyperinflation and atelectasis from mucus plugging
- CT scan (mostly for exclusion of other pathology)

A.6 Acute Bronchitis (ICD-9: 466)

Description:

- Bronchitis is inflammation of the mucous membranes of the bronchi, the airways that carry airflow from the trachea into the lungs.

Risk factors:

- Common cold
- Influenza
- Viral respiratory infection
Possible Symptoms:

- Persistent dry or wet cough
- Sputum
- Sore throat
- Runny nose
- Nasal congestion (coryza)
- Malaise
- Chest discomfort
- Fatigue
- Shortness of breath
- Wheezing

Physical Examinations:

Vital signs

- Temperature
  - Low grade fever
- Blood pressure
- Pulse:
  - Increased
- Respiratory rate:
  - Increased
- Saturation of oxygen:

Physical exams

- Decreased intensity of breath sounds
- Rhonchi
- Prolonged expiration

Lab Tests

- Blood test:
  - Raised white blood cell count
  - Elevated C-reactive protein
- Sputum sample
  - For signs of inflammation or bacterial infection
- Chest X-ray (to exclude pneumonia)
A.7 Tension Pneumothorax (ICD-9: 512.0)

Description:

■ A pneumothorax is a potential medical emergency wherein air or gas is present in the pleural cavity. A tension pneumothorax is a medical emergency as air accumulates in the pleural space with each breath.

Risk factors:

■ Trauma
■ Smoking

Possible Symptoms:

■ Sudden shortness of breath
■ Pale, cool, clammy skin
■ Dry coughs
■ Cyanosis
■ Chest pain
■ Backache
■ Arm pain
■ Subcutaneous emphysema
■ Loss of consciousness
■ Coma

Physical Examinations:

Vital signs

◆ Temperature;
◆ Blood pressure
  ○ hypotension
◆ Pulse rate:
  ○ tachycardia
◆ Respiratory rate:
  ○ Increased: tachypnea
◆ Saturation of oxygen:
  ○ hypoxia

Physical exams

◆ Decreased expansion of the chest on the affected side
Sound of air flowing (in penetrating chest wounds)
Absence of audible breath sounds
Coin test
  - Produces a tinkling resonant sound which is audible on auscultation
Hyperresonance (higher pitched sounds than normal)
Elevated JVP
Reduced cardiac preload (how to get this?)
Decreased cardiac output (how to get this?)

Lab Tests

- X-ray
  - Deep sulcus sign: low lateral costophrenic angle on the affected side
  - Presence of a thin radiolucent pleural line
  - Absence of vascular lung marking peripheral to the radiolucent line
  - Upright position and expiration may help visualize the PTX
  - Lateral decubitus chest X-ray with affected side up may also aid in visualization
- Chest CT
  - Very sensitive and may detect very small PTXs not seen on chest X-ray
- AGB:
  - May reveal hypoxemia due to V/Q mismatching
- Hemodynamic
  - Hemodynamic improvement
  - A rush of air following needle decompression or chest tube insertion

A.8 Dissecting Thoracic Aneurysm (ICD-9: 441.0)

Description:

- Aortic dissection is a tear in the wall of the aorta that causes blood to flow between the layers of the wall of the aorta and force the layers apart. Aortic dissection is a medical emergency and can quickly lead to death.

Risk factors:

- Chest trauma
- Age: high in 50-70
- Gender: male:female = 2:1

Possible Symptoms:
Severe sudden chest pain: tearing, stabbing, sharp (96%)
Syncope (9%)

Physical Examinations:

Vital signs

◆ Temperature;
◆ Blood pressure
  ● Hypertension (36%)
  ● Hypotension (25%)
  ● Severe hypotension at presentation is a grave prognostic indicator
◆ Pulse:
◆ Respiratory rate:
◆ Saturation of oxygen:

Physical exams

◆ The murmur of aortic insufficiency is audible in about 32% of proximal dissections

Lab Tests

◆ EKG
  ◆ 33%: signs of left ventricular hypertrophy (due to long-standing hypertension)
  ◆ 33%: normal

◆ Echocardiogram
  ◆ Transesophageal echocardiography (TEE)
    ● Sensitivity: 98%
    ● Specificity: 97%

◆ Aortography / aortogram
  ◆ Sensitivity: 88%
  ◆ Specificity: 94%

◆ X-ray
  ◆ Widening of the mediastinum on an X-ray of the chest:
    ● Sensitivity: 67%
    ● Specificity: low
  ◆ Calcium sign
    ● Separation of the intimal calcification from the outer aortic soft tissue border by 10mm
    ● Pleural effusion may be seen, typically in the left hemithorax
- Obliteration of the aortic knob
- Depression of the left mainstem bronchus
- Loss of the paratracheal stripe
- Tracheal deviation
- 12%-20% of individuals presenting with an aortic dissection have a "normal" chest x-ray

- MRI (current gold standard test)
  - Sensitivity: 98%
  - Specificity: 98%

- CT scan
  - Sensitivity: 96%-100%
  - Specificity: 96-100%
Appendix B

Conditional Probability Tables

B.1 Conditional Probabilities for Symptoms

**VP:** viral pneumonia (480.*): 0.139  
**PE:** pulmonary embolism (415.*): 0.139  
**CPE:** cardiogenic pulmonary edema (514): 0.139  
**Asthma:** extrinsic asthma with acute exacerbation (493.*): 0.208  
**PTX:** pneumothorax (512.*): 0.014  
**MI:** myocardial infarction (410.*): 0.208  
**AB:** acute bronchitis (446.*): 0.139  
**DTA:** dissecting thoracic aneurysm (441.*): 0.014

<table>
<thead>
<tr>
<th></th>
<th>VP</th>
<th>PE</th>
<th>CPE</th>
<th>Asthma</th>
<th>PTX</th>
<th>MI</th>
<th>AB</th>
<th>DTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cough (786.2)</td>
<td>✓0.85</td>
<td>✓0.37</td>
<td>✓0.85</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shaking chills (780.64)</td>
<td>0.15✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dyspnea (786.0)</td>
<td>✓0.55</td>
<td>0.77✓</td>
<td>1✓</td>
<td>0.97✓</td>
<td>0.9✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chest pain (786.5)</td>
<td>0.3✓</td>
<td>0.66✓</td>
<td>0.5✓</td>
<td>0.5✓</td>
<td>0.66✓</td>
<td>0.3✓</td>
<td>0.96✓</td>
<td></td>
</tr>
<tr>
<td>Diaphoresis (780.8)</td>
<td>0.07✓</td>
<td>0.36✓</td>
<td>0.2✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.3✓</td>
</tr>
<tr>
<td>Cyanosis (782.5)</td>
<td>0.07✓</td>
<td>0.19✓</td>
<td></td>
<td>0.2✓</td>
<td>0.2✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nausea (787.0)</td>
<td>0.07✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.1✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hemoptysis (786.3)</td>
<td></td>
<td>0.13✓</td>
<td>0.1✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accessory muscle use</td>
<td></td>
<td></td>
<td>0.5✓</td>
<td>0.7✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety (300)</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Pallor (782.1)</td>
<td></td>
<td></td>
<td></td>
<td>0.1✓</td>
<td>0.1✓</td>
<td></td>
<td>0.1✓</td>
<td></td>
</tr>
<tr>
<td>Ankle edema (leg swelling)</td>
<td>0.07✓</td>
<td>0.49✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>Shock (785)</td>
<td></td>
<td>0.07</td>
<td>0.25✓</td>
<td></td>
<td>0.07 ✓</td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subcutaneous emphysema</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feeling of indigestion (536.8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syncope (780.2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.1</td>
<td></td>
<td>0.09✓</td>
</tr>
<tr>
<td>Sore throat (462)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.3✓</td>
</tr>
<tr>
<td>Runny nose (478.19)</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.5✓</td>
</tr>
<tr>
<td>Malaise (780.7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.15</td>
<td>0.5✓</td>
</tr>
</tbody>
</table>
## B.2 Conditional probabilities on Physical Examinations

<table>
<thead>
<tr>
<th></th>
<th>VP</th>
<th>PE</th>
<th>CPE</th>
<th>Asthma</th>
<th>PTX</th>
<th>MI</th>
<th>AB</th>
<th>DTA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Body Temperature</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- lower than normal (&lt;36.1°C)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>- normal (36.1°C-37.5°C)</td>
<td>0.1</td>
<td>0.85</td>
<td></td>
<td></td>
<td>0.85</td>
<td>0.9</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>- low fever (37.5 – 39.5)</td>
<td>0.1</td>
<td></td>
<td></td>
<td></td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>- high fever (&gt;39.5)</td>
<td>√ 0.8</td>
<td>0.01</td>
<td>0.05</td>
<td>0.1</td>
<td>0.05e</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Blood pressure (systolic)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- low (&lt;90)</td>
<td>0.04</td>
<td>0.05</td>
<td>0.1</td>
<td>0</td>
<td>0.85</td>
<td>0.25</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>- normal (90 – 140)</td>
<td>√0.95</td>
<td>√0.8</td>
<td>√0.75</td>
<td>√1</td>
<td>0.15</td>
<td>√0.7</td>
<td>1</td>
<td>√0.39</td>
</tr>
<tr>
<td>- high (&gt; 140)</td>
<td>0.01</td>
<td>0.15</td>
<td>0.15</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
<td>0</td>
<td>0.36</td>
</tr>
<tr>
<td><strong>Pulse rate</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- low (&lt; 50)</td>
<td>0.01</td>
<td>0.05</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>- normal (50-110)</td>
<td>0.8</td>
<td>0.55</td>
<td>√0.5</td>
<td>0.5</td>
<td>0.25</td>
<td>√0.7</td>
<td>√0.5</td>
<td>√0.75</td>
</tr>
<tr>
<td>- high (&gt; 110)</td>
<td>√0.19</td>
<td>√0.4</td>
<td>√0.25</td>
<td>√0.5</td>
<td>√0.75</td>
<td>√0.25</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td><strong>Respiratory rate</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- low (&lt; 6)</td>
<td>0.01</td>
<td>0</td>
<td>0.05</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>- normal (6 – 20)</td>
<td>0.5</td>
<td>0.2</td>
<td>√0.65</td>
<td>0.1</td>
<td>√0.8</td>
<td>√0.5</td>
<td>√0.7</td>
<td></td>
</tr>
<tr>
<td>- high (&gt; 20)</td>
<td>0.49</td>
<td>√0.8</td>
<td>√0.3</td>
<td>√0.9</td>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td><strong>Saturation of oxygen</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- very low (&lt; 92%)</td>
<td>0.05</td>
<td>0.1</td>
<td>0.25</td>
<td>0.4</td>
<td>0.6</td>
<td>0.1</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>- low (92 – 97%)</td>
<td>√0.3</td>
<td>√0.8</td>
<td>√0.7</td>
<td>√0.6</td>
<td>√0.35</td>
<td>0.6</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>- normal (&gt;97%)</td>
<td>0.65</td>
<td>0.1</td>
<td>0.05</td>
<td>0</td>
<td>0.05</td>
<td>√0.3</td>
<td>0.3</td>
<td>√0.2</td>
</tr>
<tr>
<td><strong>Capnography</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- decreased exhaled CO2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Perception</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- JVD</td>
<td></td>
<td></td>
<td>0.2</td>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Auscultation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- rales/crackles (786.7)</td>
<td>0.05</td>
<td>√0.54</td>
<td>√0.95</td>
<td>0.05</td>
<td>0.2</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>- rhonchi</td>
<td>0.5</td>
<td>√0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>- pleural rub</td>
<td>0.02</td>
<td>√0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- wheeze (786.07)</td>
<td></td>
<td>0.2</td>
<td>0.4</td>
<td>0.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- cardiac murmur (785.2)</td>
<td>0.23</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td>0.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- sound of air flow</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8</td>
</tr>
<tr>
<td>- tinkling resonant sound by coin test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- S3 and S4 gallop</td>
<td>0.34</td>
<td>√0.45</td>
<td></td>
<td></td>
<td>0.63</td>
<td>(me)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- 4th heart sound</td>
<td>0.24</td>
<td>√0.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- 2nd heart sound paradoxical</td>
<td>√0.38</td>
<td>0.27(me)</td>
<td></td>
<td></td>
<td>√0.78(me)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Percussion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- dullness</td>
<td></td>
<td></td>
<td>0.2</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
B.3 Conditional probabilities of Lab Tests

The price information (with blue color) is from

http://userpages.bright.net/~ach/Patient%20Price%20Information.htm

http://www.bloodworksusa.com/cs_productpages/33.html

normal range of the tests are according to:

http://www.dalesplace.net/lab_values.php

<table>
<thead>
<tr>
<th></th>
<th>VP</th>
<th>PE</th>
<th>CPE</th>
<th>Asthma</th>
<th>PTX</th>
<th>MI</th>
<th>AB</th>
<th>DT A</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ABG (259$, 15min)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- acidotic (&lt;7.35)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- normal (7.35-7.45)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- alkalemic (&gt;7.45)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pO2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- low (&lt;69.8mmHg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- normal (69.8-99.8 mmHg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- high (&gt;99.8mmHg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pCO2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- low (&lt;36.3mmHg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- normal (36.3-45mmHg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- high (&gt;45mmHg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-a gradient</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- normal (&lt; 15 mmHg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- increased (&gt; 15 mmHg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CBC (30$, 15min)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WBC (10^9/L)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- low (&lt;3.9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- normal (3.9 – 11)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- high (&gt;11)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hemoglobin (g/dl)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- low (&lt;13.8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal (13.8-17.2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High (&gt;17.2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low (&lt;12.1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal (12.1-15.1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High (&gt;15.1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Children
Low (<11)
Normal (11-16)
High (>16)

Pregnant women
Low (<11)
Normal (11-12)
High (>12)

**Hematocrit (%)**

**Men**
Low (<41)
Normal (41-53)
High (>53)

**Women**
Low (<36)
Normal (36-46)
High (>46)

Children
Low (<49)
Normal (49-61)
High (>61)

**Platelet Count (1000/mm³)**
Low (<150)
Normal (150-400)
High (>400)

**Basic Metabolic Panel (25$, 15min)**

- **blood sodium**
  - low (< 135 mmol/L)
  - normal (135 – 147)
  - high (> 147)

- **Potassium**
  - low (< 3.5 mmol/L)
  - normal (3.5 – 5.0)
  - high (> 5.0)

- **Chloride**
  - low (< 98 mmol/L)
  - normal (98-108)
  - high (> 108)

- **Bicarbonate**
  - low (< 22 mmol/L)
  - normal (22 - 30)
  - high (> 30)

- **BUN**
  - low (< 7 mg/dl)
<table>
<thead>
<tr>
<th>Test</th>
<th>Result</th>
<th>P-value</th>
<th>P-value (unadjusted)</th>
<th>P-value (adjusted)</th>
<th>P-value (adjusted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creatinine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- low (&lt;68 umol/L)</td>
<td></td>
<td></td>
<td>0.0001</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>- normal (68 – 118)</td>
<td></td>
<td></td>
<td>0.5</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>- high (&gt; 118)</td>
<td></td>
<td></td>
<td>0.4999</td>
<td>0.2999</td>
<td></td>
</tr>
<tr>
<td>Cardiac Panel</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>CK-MB fraction (57)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>- normal (0-5 ug/L)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>- increased (&gt;5)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>TnI or TnT level (69)</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>- low (&lt; 8 hours)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>- normal (8 – 16 hours)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>- high (&gt; 16 hours)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Other Labs</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>C-reactive protein</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>- normal (&lt; 6 mg/L)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>- increased (&gt; 6 mg/L)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>D-dimer</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>- normal (0-500 ng/mL)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>- increased (&gt; 500)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>BNP (79)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>- Normal (0-100 pg/mL)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>- Mild increase (100 – 500)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>- Severe increase (&gt;500)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>X-Ray (117)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>- positive</td>
<td>0.4</td>
<td>0.1</td>
<td>0.95</td>
<td>0.85</td>
<td>0.05</td>
</tr>
<tr>
<td>- negative</td>
<td>0.6</td>
<td>0.9</td>
<td>0.05</td>
<td>0.15</td>
<td>0.95</td>
</tr>
<tr>
<td>CT Chest (835)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>- positive</td>
<td>0.5</td>
<td>0.05</td>
<td>0.9</td>
<td>0.05</td>
<td>0.95</td>
</tr>
<tr>
<td>- negative</td>
<td>0.5</td>
<td>0.05</td>
<td>0.9</td>
<td>0.05</td>
<td>0.95</td>
</tr>
<tr>
<td>EKG (89)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>- positive</td>
<td>0.3</td>
<td>0.2</td>
<td>0.3</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>- negative</td>
<td>0.7</td>
<td>0.8</td>
<td>0.7</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>Echocardiogram (849)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>- positive</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>- negative</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>V/Q scan (638)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>- positive</td>
<td>0.5</td>
<td>0.1</td>
<td></td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>- negative</td>
<td>0.5</td>
<td>0.1</td>
<td></td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>Ultrasound (342)</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>- positive</td>
<td>0.8</td>
<td>0.2</td>
<td></td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>- negative</td>
<td>0.8</td>
<td>0.2</td>
<td></td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>Pulmonary function test (87)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Peak flow:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- normal (400-600)</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- moderate exacerbation (100 – 300)</td>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- severe exacerbation (&lt;100)</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
VITA

Shizhuo Zhu

Education

Ph. D. in Information Sciences and Technology 2010

Dissertation: Hypothesis-Driven Story Building: Counteracting Human Cognitive Biases to Improve Medical Diagnosis Support

College of Information Sciences and Technology, The Pennsylvania State University, University Park, Pennsylvania, USA

M. E., in Computer Science and Technology 2003

Department of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui, China

B. Phil in Mathematical Logic 2000

Department of Philosophy, Peking University, Beijing, China

Publications


