Effects of automation transparency on trust and diagnostic decision making with an automated decision aid during medical emergencies

Mohamed Syahid Hassan

2015

Mohamed Syahid Hassan. (2015). Effects of automation transparency on trust and diagnostic decision making with an automated decision aid during medical emergencies. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/62933>

<https://doi.org/10.32657/10356/62933>

**EFFECTS OF AUTOMATION TRANSPARENCY ON
TRUST AND DIAGNOSTIC DECISION MAKING
WITH AN AUTOMATED DECISION AID DURING
MEDICAL EMERGENCIES**

MOHAMED SYAHID BIN HASSAN

MOHAMED SYAHID BIN HASSAN

School of Mechanical and Aerospace Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Doctor of Philosophy

2015

[INTENTIONALLY LEFT BLANK]

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisors (past and present), Asst Prof Park Taezoon, Prof Martin Helander and Assoc Prof Leong Kah Fai for their constant and invaluable guidance and support throughout my PhD journey, without which, none of it would have been possible. I would also like to thank Emeritus Prof Christopher Wickens for being a source of wisdom and inspiration in providing me with the insights that I needed to make this study come true.

I would also like to express my greatest appreciation for Assoc Prof Siau Chiang and Assoc Prof Chen Fun-Gee from National University of Singapore for supporting my research project and providing me with access to the medical simulation facilities at the Centre for Healthcare Simulation, assisting me with their expertise in the subject matter and facilitating recruitment of their students as participants for the study. I would also like to thank the staff at the Centre for Healthcare Simulation, especially Kok Ann and Johanis, for their hard work and dedication in helping me execute the simulations for the experiment.

I would also like to thank all my lab-mates, especially SQ, Rain, Jessie, Benedict, Lisa, and Fitri, not only for being wonderful friends but also for being there when I needed somebody to discuss my ideas with. Our friendships have provided the sweetest moments during this arduous journey.

Last but not least, I would like to thank my wife, my baby daughter and the rest of my family for their undying support and encouragement which had provided me the strength that I needed to keep on going.

[INTENTIONALLY LEFT BLANK]

ABSTRACT

Automated decision aids have the potential to improve decision making performance and reduce errors in complex decision making, such as during a medical emergency. However, since the reliability of such automated systems is imperfect, it gives rise to the problem of automation disuse and misuse, which is a result of miscalibrated trust in the system. Furthermore, when the system eventually commits an error, trust and acceptance is reduced and can only recover slowly. This study investigates if diagnostic performance can be increased by improving trust calibration, trust drop after failure experience and trust recovery through the introduction of automation transparency features. A transparent automation is one that provides the user with insights into the internal processes that produce its outcomes. In the main experiment, novice doctors in a simulated emergency diagnosis task used decision aids with different transparency configurations, manipulated by the presence of two transparency features, a list of Key Diagnostic Cues that explain its recommendation and a Likelihood Rating that displays how likely the aid thinks its recommendation is correct. No significant evidence that appropriate trust in the aid's recommendation was improved by the transparency configuration was found. Instead, Likelihood Ratings had reduced diagnostic confidence during appropriate reliance and the transparency features had intensified the trust drop and corresponding confidence decrement from the experience of error. Nonetheless, Key Diagnostic Cues were found to decrease confidence during distrust while Likelihood Ratings was found to improve trust recovery to the point where it was better than the initial trust. Although transparency features could intensify mistrust immediately after the automation has committed an error, such automated aids should still possess features that provide transparency as it can improve overall trust in the system, especially over long term usage.

[INTENTIONALLY LEFT BLANK]

TABLE OF CONTENTS

Acknowledgement

Abstract

List of Figures

List of Tables

Chapter 1: Introduction	1
1.1 Background	1
1.2 Research Objective	6
1.3 Scope of Research	6
1.4 Overview of Report	7
 Chapter 2: Literature Review	 9
2.1 Misdiagnoses in Medical Emergencies	9
2.2 Diagnostic Decision Making during Medical Emergencies	10
2.3 Challenges of Emergency Decision Making	15
2.4 Automated Diagnostic Decision Aids in Medicine	17
2.5 Acceptance of and Reliance on Automated Decision Aids	22
2.6 Decision Making with Automation	23
2.7 Trust in Automation	25
2.8 Effect of Automation Errors on Trust	27
2.9 Appropriate Reliance on Automation	30
2.10 Factors that Influence Trust in Automation	34
2.11 Automation Transparency	38
2.12 Key Diagnostic Cues and Likelihood Ratings for Transparency	43
2.13 Research Gaps	45
 Chapter 3: Conceptual Model	 47
3.1 Model for Transparency and Reliance on Imperfect Automated Aid	47
3.2 Transparency and Performance during Non-Automation Failure and during Automation Failure	50
3.3 Transparency and the Evolution of Trust and Confidence Resulting from Error Experience and Return to Non-Failure Experience	55
3.4 Summary of Hypotheses	57
3.5 Flow of Research Study	59
 Chapter 4: Exploratory Study	 61
4.1 Background	61
4.2 Objectives	62
4.3 Method	62
4.4 Results	65
4.5 Discussion	71

Chapter 5: Method	75
5.1 Participants	75
5.2 Independent Variables	75
5.3 Dependent Variables	77
5.4 Procedure	78
Chapter 6: Results	89
6.1 Overview of Results	89
6.2 Transparency and Performance during Non-Automation Failure	93
6.3 Transparency and Performance during Automation Failure	104
6.4 Transparency and the Experience of Automation Failure	111
6.5 Transparency and the Experience of Return to Correct Automation Performance	117
6.6 Effect of Scenario on Trust and Diagnostic Confidence	119
6.7 Overall Subjective Measures of Automated Decision Aids	121
6.8 Summary of Results	122
Chapter 7: Discussion	125
7.1 Transparency, Trust and Reliance	125
7.2 Transparency and Confidence	129
7.3 Transparency and the Experience of Automation Failure	130
7.4 Transparency and the Experience of Return to Correct Automation Performance	132
7.5 Overall Assessment of Transparency Features	133
7.6 Summary of Analysis	135
Chapter 8: Conclusion	
8.1 Summary of Findings	137
8.2 Practical Implications of Findings	139
8.3 Limitations of Study	140
8.4 Future Directions	142
References	145
Appendices	
Appendix I	Exploratory Study: Selected Participant Utterances
Appendix II	Detailed Views of Transparency Configurations
Appendix III	Details of Scenario Designs
Appendix IV	Diagnosis Training Notes for Participants
Appendix V	Post-Condition Questionnaire Items
Appendix VI	Experimental Results
Appendix VII	Institutional Review Board Approval Certificate

LIST OF FIGURES

- Figure 2-1 Hygene model for hypothesis generation (Thomas et al, 2008)
- Figure 2-2 Representation of information integration process to form belief between two competing hypotheses (Wickens & Hollands, 2000)
- Figure 2-3 Hybrid Lens Model of operator's judgment with automated decision aids by (Seong & Bisantz, 2002)
- Figure 2-4 Framework to predict automation use (Dzindolet et al., 2002)
- Figure 2-5 Model of Complacency and Bias in Automation (Parasuraman & Manzey, 2010)
- Figure 2-6 Calibrated trust, over-trust and distrust from the operator and the trustworthiness of the automation (Lee & See, 2004)
- Figure 2-7 Resolution of trust in automation
- Figure 2-8 Dynamic model of trust and reliance on automation (Lee & See, 2004)
- Figure 3-1 Model for Transparency and Reliance on Imperfect Automated Aid for Diagnostic Decision Making
- Figure 3-2 Conceptual Model for effect of transparency configuration on trust and decision confidence
- Figure 3-3 Conceptual Model for effect of transparency configuration on the change in trust and confidence due to failure and non-failure
- Figure 3-4 Flow for research study
- Figure 5-1 Transparency configurations for the automated decision aid
- Figure 5-2 Outline of Experimental Procedures
- Figure 5-3 Multiple choice screen for diagnosis on Participant Response Terminal
- Figure 5-4 Diagnostic confidence rating on Participant Response Terminal
- Figure 5-5 Trust in decision aid rating on Participant Response Terminal
- Figure 6-1 Percentage of subjects who relied on the aid appropriately over each trial of the session for different combinations of transparency features
- Figure 6-2 Trust in aid's recommendation over all trials

- Figure 6-3 Mean Confidence in diagnosis over all trials, when appropriately relying or not relying on the decision aid
- Figure 6-4 Mean Confidence in diagnosis over all trials, when distrusting or over-trusting the decision aid
- Figure 6-5 Trust in aid's recommendation when appropriately relying or distrusting a decision aid when there was no automation failure
- Figure 6-6 Diagnostic Confidence when appropriately relying or distrusting a decision aid when there was no automation failure
- Figure 6-7 Trust in aid's recommendation when over-trusting or appropriately not relying on the decision aid when there was automation failure
- Figure 6-8 Diagnostic Confidence Ratings when over-trusting or appropriately not relying on the decision aid when there was automation failure
- Figure 6-9 Trust Ratings from Pre-Failure trial to First Post-Failure Trial
- Figure 6-10 Diagnostic Confidence Ratings when agreeing with decision aid that was correct before and immediately after error
- Figure 6-11 Trust Ratings from Pre-Failure trial to Second Post-Failure Trial
- Figure 6-12 Trust in aid's recommendation for each non-failure scenario
- Figure 7-1. Illustration of effects of transparency features on trust dynamics for an imperfect automation

LIST OF TABLES

Table 2-A	Levels of Automation (Sheridan & Wickens, 2000)
Table 3-A	Possible outcomes during reliance on an imperfect automated diagnostic decision aid
Table 3-B	Summary of Hypotheses
Table 5-A	List of medical diagnosis scenarios
Table 5-B	Assignment of experimental conditions to participants
Table 6-A	Possible outcomes for each simulation trial
Table 6-B	Mean number out of the three non-failure trials where there was appropriate reliance
Table 6-C	Number of subjects who appropriately relied on or distrusted the aid when using one aid without the Likelihood Ratings and another with the Likelihood Ratings for each non-failure trial
Table 6-D	Number of subjects who appropriately relied on or distrusted the aid when the Key Diagnostic Cues were absent or present for each Non-failure trial
Table 6-E	Mean of subjects' mean Trust in Recommendations from decision aid across all trials that the aid was correct
Table 6-F	Mean of subjects' mean Diagnostic Confidence during trials when appropriately relying on decision aid when there was no automation failure
Table 6-G	Mean Diagnostic Confidence when distrusting the decision aid when there was no automation failure
Table 6-H	Number of subjects who appropriately did not rely on or over-trusted the aid when using one aid without the Likelihood Ratings and another with the Likelihood Ratings for the Failure trial
Table 6-I	Number of subjects who appropriately did not rely on or over-trusted the aid when the Key Diagnostic Cues were absent or present for the Failure trial.
Table 6-J	Mean of subjects' Trust in Recommendations from decision aid during the trial that the aid was incorrect

Table 6-K	Mean Diagnostic Confidence when disagreeing with decision aid when there is automation failure (Appropriate Non-Reliance)
Table 6-L	Mean Diagnostic Confidence when agreeing with decision aid when there is automation failure (Over-Trust)
Table 6-M	Mean change in trust rating from Pre-Failure trial to First Post-Failure trial
Table 6-N	Mean Change in Diagnostic Confidence during appropriate reliance from before to after error experience
Table 6-O	Mean change in trust rating from Pre-Failure trial to Second Post-Failure trial
Table 6-P	Questionnaire items and responses
Table 6-Q	Summary of Results

CHAPTER 1: INTRODUCTION

1.1 Background

With advancements in computer technology and artificial intelligence algorithms, automation has become common in many domains to support or enhance human performance. In various decision making environments, human decision makers have access to automated decision support tools to help them make decisions. Some of these decision support systems provide assistance in the acquisition of information and some facilitate decision making by making the information easier to analyze (Parasuraman, Sheridan & Wickens, 2000). More sophisticated systems are able to automate the decision making process itself and produce the decision itself from the available data. Such systems can be seen in common everyday applications such as automated route planning in GPS navigation or automated recommendation of items to online shoppers based on their previous purchase behaviours.

Automated decision aids are especially useful in complex decision making environments and have been utilized in military, process control, aviation and business analytics. In these domains, there may be large amounts of relevant data to be used to make decisions which may too overwhelming for a human decision maker to process on his own. Algorithms in these automated decision aids can be capable of processing the data to assist the human to make the final decision. In business analytics for example, decision support tools process past performances and trends from multiple variables to predict future outcomes.

Practitioners in medicine have also sought to leverage on automated decision aids to address human errors in decision making. In 2000, the Institute of Medicine in the U.S.

published the report, “To Err is Human: Building a Safer Health System”, which highlighted human errors by doctors and healthcare workers as a major cause of morbidity and mortality in the medical field (Kohn, Corrigan, Donaldson, 2000). Among these medical errors, misdiagnoses are one of the most common causes of harm to patients. It is estimated to cause 40 000 to 80 000 deaths in the U.S. every year (Newman-Toker & Provonost, 2009). A misdiagnosis can be committed by any doctor in any healthcare environment; from the general practitioner handling routine cases in a family clinic up to the emergency room doctor managing multiple patients on a busy night. However, forming diagnoses during emergencies are in greater need of intervention because such situations present a greater challenge to the doctor’s diagnosis process. Decision making in the emergency department, where such situations are common, becomes more complex due to aggravating factors such as high levels of diagnostic uncertainty, high decision density, high cognitive load, poor feedback and narrow time windows (Croskerry & Sinclair, 2001).

Given the high potential for errors in medical diagnosis and the potential harm, much effort has gone into developing solutions to minimize them, including the development of clinical decision support systems. A clinical decision aid or a decision support system can be understood as an active knowledge system that use two or more items of patient data to generate case-specific advice (Wyatt & Spiegelhalter, 1991). Commercially available diagnostic decision aids, such as DXplain and DiagnosisPro, are able to generate recommended diagnoses based on available patient information.

As computing technology progresses, diagnostic decision aid systems are expected to be more accurate and more robust. However, as automation for a complex task like medical decision making can be imperfect, which means that there is always a possibility of the automation making a mistake, the human should remain as an

integral component of the decision making system. The automation's decision making capability is only as effective as its developer's design and it may not be able to integrate certain information the way that humans can. It needs to be supported by the flexibility of human reasoning to decide whether to fully or partially integrate the input from the automation into the final decision. Such a decision aid system needs to be designed based on the understanding of the dynamics in collaborative human-automation decision making.

The key problems in the use of imperfect automated systems that are well established in the literature are that of automation disuse and misuse (Parasuraman & Riley, 1997). Disuse refers to when the user chooses not to use the automation when it would actually be more beneficial to use it while misuse refers to when the user uses the automation when it is inappropriate to use it, such as when it fails.

A factor that plays a major role in determining the user behaviour that leads to disuse and misuse is the level of trust in the automation. Trust in automation determines the acceptance of assistance from an automated system (Lee & Moray, 1992). For a doctor to work with an automated diagnostic decision aid, the system must feel trustworthy. However, merely designing for maximal trust is not sufficient nor is it prudent because an imperfect automation can under certain circumstances, fail unexpectedly. Therefore, designing for appropriate trust and reliance is the more desirable approach (Lee & See, 2004).

Within the collaborative human automation decision making paradigm, there should be affordances for the human to calibrate his trust towards the automation. The design of the automation should possess properties that allow the human to perceive how much the automation can be trusted to perform its actions properly. This would allow him to

make decisions with higher confidence. Furthermore, when the user experiences a failure for the first time when using an imperfect automation, it can cause a detriment to trust that then affects acceptance in subsequent uses of the automation (Muir & Moray, 1996). It may take a long time for the trust to return back to its initial levels.

Imperfect automated decision aids need to be carefully designed such that users are able to assess when it is appropriate to rely on it and when it is not. Being imperfect machines, they are bound to eventually commit errors. Effort should be made to design the system such that it is able to recover trust and acceptance as soon as possible when it returns to proper performance.

Some automation designers create their automated systems as “black boxes” that do not reveal the processes involved in generating the outputs from the inputs. The degree of understandability of automation can have a major influence on trust and “black box” designs do not afford any understandability to the user (Sheridan, 1988). Some studies have shown that systems that are more transparent are more positively received by users and are deemed to be more trustworthy (Herlocker et al, 2000; Sinha & Swearingen, 2002).

Manipulating the transparency of an automated decision aid is a potential design parameter to improve the calibration of trust in automated decision aids. An automated system is transparent in design when it allows the user to have an insight into the inner workings of the system and thus allows the user to have a better understanding of the rationale behind its actions (Cramer et al, 2008). The insight could be provided from different levels, ranging from providing a supporting reference from an established decision guideline to details about the causal logic, such as the Bayesian network or decision tree, that was used. Users who have a better understanding of how the

automation works can make a better decision of how much to trust its assistance and thus rely on it more appropriately (Lee & See, 2004). Better trust calibration can improve decision confidence when the user knows that he is relying on the aid appropriately or when he knows that he is appropriately not relying on the aid because of a potential failure.

Previous literature has explored the role transparency in automation trust and acceptance for decision aids in domains that are non-critical in nature, such as recommender systems for movies (Herlocker et al, 2000) and artwork (Cramer et al, 2008). There has been a lack of studies that validated these findings in the context of high-risk and time-pressured decision making environments such as emergency diagnosis. In these domains, there is greater risk involved in the decision to be made collaboratively with the automation and this provides a different dynamics of interaction as compared to the domains that the literature has addressed. These previous studies have also looked at how transparency affects trust in the overall system but not how it interacts with the experience of automation failure. Transparent systems afford the user the ability to understand why a system has made an error and could potentially reduce the trust detriment caused by the failure. A deeper understanding of the effects of transparency of an imperfect automated decision aid on user trust and decision making during time pressure situations could shed light on how decision aids for high-risk and time-pressured decision environments can be better designed to increase its acceptance and for more appropriate reliance.

This research addresses the gaps in the literature regarding the relationship between user trust and the transparency design of imperfect decision support automation in the context of high risk emergency decision making. It will provide empirical evidence of the interaction between human and automation, of different transparency

configurations, in emergency diagnostic decision making within a near-naturalistic setting.

1.2 Research Objective

The objective of this research is to investigate how the presence of transparency design features on an imperfect automated decision aid affects trust, reliance and diagnostic performance during high risk, high pressure and time-critical decision making when the automation is performing properly, when it is failing and immediately after experiencing an automation failure.

1.3 Scope of Research

The study is applied within the domain of medical emergency diagnostic decision making and for novice decision makers. The automated decision aid will serve to provide diagnostic recommendations to a novice doctor within a simulated patient emergency setting. The aid is imperfect and may fail at providing a correct recommendation. The novice doctor is aware of the possibility of failure and will need to make a decision to rely or not rely on the aid.

The variable of interest is the transparency configuration of the automated decision aid, which is manipulated specifically through the display of key diagnostic cues and diagnostic likelihood rating to accompany the provided recommendation.

The study will investigate the effects of automation transparency on the user's trust of the decision aid, their decision to rely on the aid and their diagnostic performance

which will be reflected through their diagnostic confidence when agreeing or disagreeing with the aid. It will also study how transparency affects the evolution of trust over multiple uses of the system which includes the experience of an automation failure.

1.4 Overview of Report

- **Chapter 1** has provided an overview of the background and objectives of this research.
- **Chapter 2** will provide an in-depth review of relevant existing literature, explain the key terms and concepts of this research and identify the gaps in the literature.
- **Chapter 3** will synthesize the knowledge from the available literature into the research conceptual model and form the hypotheses for this research.
- **Chapter 4** will provide details of the initial exploratory study to understand the dynamics of diagnostic decision making in a high-fidelity simulated medical emergency scenario.
- **Chapter 5** will outline the methods for the study that was conducted to address the research hypotheses.
- **Chapter 6** will provide the results of the study and its analyses.
- **Chapter 7** will provide a discussion of the findings of the study.

- **Chapter 8** will provide an overview of the significance of the findings of this research, discuss its limitations and identify future directions for this research.

CHAPTER 2: LITERATURE REVIEW

2.1 Misdiagnoses in Medical Emergencies

A misdiagnosis or diagnostic error occurs when the identification of the nature or cause of a disease or injury is missed, wrong or delayed (Graber, 2005). It has the potential to cause harm to the patient as a result of a delay or failure to treat a condition that is actually present or a provision of the wrong treatment for a condition that is actually not present (Newman-Toker & Provonost, 2009).

In a review of 723 paid malpractice claims involving negligent adverse events, Holohan et al. (2005) found that the most frequent were for diagnostic related negligence (45%). In comparison, 16% of those total claims were related to medication. Out of the claims for diagnostic related negligence, 41% of them had resulted in death. A review of medical error studies by Schiff et al. (2005) revealed that 10% to 30% of the errors were errors in diagnosis while their review of selected disease-specific studies showed substantial percentages of patients who experienced missed or delayed diagnoses. According to an anonymous survey of 726 paediatricians by Singh et al (2010), 54% of the respondents reported making at least one or two diagnostic errors per month and 45% reported making at least one or two diagnostic errors that caused patient harm per year. An Australian study of adverse events in hospital admissions found that 81% of diagnosis related adverse events were rated as “highly preventable” (Wilson et al., 1995).

2.2 Diagnostic Decision Making during Medical Emergencies

A physician has to make large number of decisions when handling a patient who is experiencing abnormalities. One of the most crucial decision tasks is to reason out and decide the patient's diagnosis based on the available information. Sandhu and Carpenter (2006) had provided an overview of various general models of clinical reasoning that physicians use, usually in a combination, to construct their final diagnoses:-

- *Hypothetico deductive Method.* The physician gathers enough data to assess the seriousness of the condition and prescribe a therapeutic action and generates a working diagnosis, based on the findings from these initial presenting data. As more data becomes available from subsequent investigations and responses to treatments, the hypotheses are refined and competing hypotheses are excluded until the physician is able to confirm a diagnosis.
- *Algorithmic.* The physician performs the diagnosis and carries out treatment following a set of steps in a protocol or flowchart that was previously established based on existing knowledge.
- *Pattern Recognition.* The physician quickly forms a diagnosis when the combination of diagnostic evidence seems obvious to him as matching a particular diagnosis from his previous experience.
- *Rule out worst-case scenario.* The physician forms and maintains a list of critical life-threatening diagnoses in memory that he feels he cannot afford to miss. He will then search for evidence to exclude these diagnoses.

- *Exhaustive.* The physician gathers as much data from the patient as possible before assessing and filtering through all of them to identify the most likely diagnosis.
- *Event-driven.* The physician's main focus is to stabilize the patient by responding to the presenting symptoms with the appropriate therapeutic actions. Depending on the response to the therapy, he will evaluate diagnostic options.

In the initial stages of the scenario, the full picture of the problem is never available. Bits of information have to be detected or obtained over time and the overall picture has to be constructed. From the data, possible explanations of what is happening to the patient has to be inferred and hypothesized by the physician based on his knowledge of the human physiology and relevant diseases (Lee, Vicki, & Geoffrey, 2000; Wears, 2009).

Initial Generation of Diagnostic Hypotheses

A physician typically generates about three to five possible hypotheses over the duration of the encounter before forming his final diagnosis (Elstein, 2009). This can be done rapidly by expert physicians (Barrows & Feltovich, 1987). The success of identifying the correct diagnosis is strongly predicted by the ability to generate the correct diagnosis among the hypotheses generated during the early part of the encounter (Coderre et al., 2003). Diagnostic errors usually occur if these early hypotheses do not contain the correct diagnosis (Barrows & Feltovich, 1987).

Figure 2-1. Hygiene model for hypothesis generation (Thomas et al, 2008)

The Hygiene model proposed by Thomas et al (2008), shown in Figure 2-1, provides a process from which hypotheses are generated based on the observed environmental data. In the first step, an initial symptom or set of symptoms is sampled by the decision maker from the environment and this activates traces from his episodic memory, which represents the collection of the decision maker's past experience with previous patients with similar symptoms. An unspecified probe is then extracted from the episodic memory which resembles the hypotheses that are strongly related to the observed data, based on his past experience. This unspecified probe is then compared against the known prototypical representations of diseases or "illness scripts" in the decision maker's semantic memory. This is similar to the concept of the activation of "illness

scripts” from Charlin, Tardif, & Boshuizen (2000) which are pre-stored schemas for illnesses including default, acceptable or abnormal values for relevant illness attributes. The traces in semantic memory which are sufficiently activated by the degree to which the trace matches the unspecified probe are added to the set of competing hypotheses in working memory. Hypotheses are continually added to the set of competing hypotheses until the unspecified probe consecutively fails to activate a trace a certain number of times. The generated set of leading hypotheses is then evaluated and the decision maker may implement a hypotheses-guided search for more data to refine the set of hypotheses.

The Hygene model highlights the importance of the decision maker’s previous experience with similar scenarios and their knowledge of “illness scripts” in the process of generating possible explanations of what is happening to the patient. It implies that the failure to generate a relevant diagnostic hypothesis is due to the inability to recognize the current scenario as something familiar or the lack of a useable prototypical representation of a disease in the decision maker’s knowledge base to match against what they perceive.

Cue Integration during Hypothesis Evaluation

The cues that the decision maker perceives from the decision making environment are important inputs which he uses to evaluate the likelihood of the different hypotheses that have been generated. In Figure 2-2, Wickens and Hollands (2000) provides an outline of the cue integration process during hypothesis evaluation.

Figure 2-2. Representation of information integration process to form belief between two competing hypotheses (Wickens & Hollands, 2000)

The true state of the world is represented by the stimuli that it generates. Some or all of these stimuli are perceived and paid attention to as cues by the decision maker. The belief towards either of the hypotheses being considered is influenced by the cues that the decision maker pays attention to. In the context of medical diagnosis, these alternative hypotheses may represent a choice between possible disease conditions or it could represent whether a specific disease is present or absent. For each cue that the decision maker attends to, their influence on the respective hypothesis beliefs is determined by the information value of the cue, which is the product of the cue's diagnosticity and reliability and reflects how valuable the cue is for the decision maker to use it to decide on a hypothesis. The cue's diagnosticity refers to how strongly the cue serves as a supportive or contradictory piece of evidence towards either of the alternative hypotheses. The cue's reliability refers to how much the decision maker

feels that the cue is actually what he believes it is. A cue that has a high information value (i.e. high reliability and high diagnosticity) can allow the decision maker to easily choose a hypothesis with confidence but the same cannot be said for highly diagnostic cue that has low reliability. The degree of attention given to a cue is moderated by the physical features of the cue, which can make it more salient or easier to notice than others. Especially under time pressure, decision makers tend to selectively process the more salient cues when forming opinions even if they may be less diagnostic (Wallsten & Barton, 1982)

During the integration of multiple cues, the decision maker will attempt to consider these properties of the available perceived cues to arrive at a decision between the alternative hypotheses but their ability to do so optimally is limited by available attentional resource. As a result, not all perceived cues may be attended to and it is also possible that the cues may not be correctly weighted according to actual their information values (Wickens & Hollands, 2000).

2.3 Challenges of Emergency Decision Making

In a naturalistic setting, the decision making environment is typically characterized by a number of features, as described by Orasanu & Connolly (1993). The problems posed by the situation are ill-structured. The environment is dynamic and there is a lot of uncertainty. Goals are ill-defined, constantly changing and at times can compete with one another. There are action and feedback loops, some of which may be hidden and there are usually delays before the effects of any action can be observed. Time stress and high stakes are usually present. Multiple players are usually involved in decision making and actions. And there is pressure to operate within organizational

goals and norms. In the case of a patient emergency, for example when the state of the patient is critical and deteriorating rapidly, these characteristics are even more pronounced.

Diagnosing and treating a patient while under emergency conditions also resembles a “wicked problem” (Rittel & Webber, 1973; Wears, 2009). Possible diagnostic options may not be evaluated as true or false, but only as better or worse than one another. Tests for possible diagnoses usually cannot be done immediately, and they may evolve over time. Treatment options are usually partial, not resolving the patient’s problem completely, but only moving the patient’s condition in a favourable direction. Trial and error may be very difficult as any action the physician takes can have significant consequences and is often irreversible. Finally, goals in “wicked problem” situations are usually vague, incomplete, constantly changing and may contradict one another.

“Wicked problems” in a naturalistic emergency decision making scenario are likely to stretch the cognitive resources of a physician trying to form an understanding of what is happening to the patient. It would be nearly impossible for a physician to gather every possible relevant piece of information about the patient before making a decision. Therefore, it would be extremely challenging to expect a human decision maker to consider all available information and make optimal decisions in a rational and objective manner. More appropriately, this form of decision making operates within bounded rationality (Simon, 1955), meaning that the decision making capability of the human is limited by what he knows and what information he has gathered, his cognitive limitations and the finite amount of time for decision making.

Operating on bounded rationality, human decision making tends to resort to cognitive biases and heuristics, which are mental shortcuts that the human unconsciously makes

in decision making in order to produce a decision in a timely manner (Tversky & Kahnemann, 1974). Examples of these heuristics and biases that are prevalent in clinical judgment include the availability heuristic, where the decision maker judges the probability of an event based on the ease at which it can be brought to mind (Tversky & Kahnemann, 1973), representativeness, where the decision maker judges the probability of a sample based on the degree to which its salient features are similar to that of the parent population (Kahnemann & Tversky, 1972), and confirmation bias, where the decision maker tends to seek and prioritize evidence that support the currently held belief while failing to seek or utilize evidence that is inconsistent with the belief or supports an alternative hypothesis (Mynatt, Doherty, & Tweney, 1977). While these biases and heuristics are known to make decision making more manageable and efficient, it is also known to increase the likelihood of sub-optimal decisions in clinical decision makers which can then lead to medical errors (Bornstein & Emler, 2001).

2.4 Automated Diagnostic Decision Aids in Medicine

Automation can be defined as the execution of a function previously performed by a human by a machine agent (Parasuraman & Riley, 1997). It can be applied to the different stages of human information processing from information acquisition to information analysis to decision and action selection and finally action implementation (Parasuraman, Sheridan & Wickens, 2000).

Automation at these different stages can be described along a continuum from full manual control over the function with no machine involvement up to complete automation where the machine controls all aspects of the function. Parasuraman,

Sheridan & Wickens (2000) illustrates this continuum through the ten levels of automation in Table 2-A.

Table 2-A. Levels of Automation (Adapted from Sheridan & Wickens, 2000)

Level	Machine Involvement	Human Involvement
1	Machine offers no assistance	Human performs all the decisions or actions himself
2	Machine provides a complete set of possible decisions or actions	Human chooses from all the possible options
3	Machine narrows down the possible decisions or actions to a few alternatives	Human may choose from machine's suggestions or make his own
4	Machine provides a single decision or action suggestion	Human may choose machine's suggestion or make his own
5	Machine provides a single suggestion and executes it if human approves	Human considers machine's suggestion and decides whether to approve its execution
6	Machine makes its choice and executes it automatically if human does not veto it within a set amount of time	Human has a set amount of time to consider the machine's choice and veto it
7	Machine executes its choice automatically	Human is informed of the choice executed.
8	Machine executes its choice automatically	Human is informed of the choice executed if he asks for it
9	Machine executes its choice automatically	Human is informed of the choice if machine decides to do so
10	Machine performs all decisions and actions autonomously	Human is ignored

Automated decision aids are a form of automation that executes the analysis of information for the human operator. They are computer-based systems that utilize facts, rules and inference techniques in order to generate a hypothesis or inference of what is going on based on (Rauch, 1984). They can be classified between levels 2 to 5 on the continuum described above as they are commonly used to support information processing but leave the control of the actual execution of the response action to the

user. Some are able to acquire information automatically from the decision making environment and some are also able to automate the selection of decisions and actions. The decision making algorithms used for information analysis are typically derived from the knowledge of domain experts, leading to these aids sometimes being referred to as “expert systems”.

Automated decision aids have the potential to reduce user workload and increase decision efficiency by reducing information sampling and increasing decision speed (Röttger, Bali & Manzey, 2009). It can also increase the quality of the user’s decisions (McGuirl & Sarter, 2006). These are especially useful benefits for complex decision making environments and various high-risk domains such as aviation, military and process control have used automated decision aids to support the human decision maker in various ways.

Doctors and engineers in the medical field have also recognized the potential for automated decision aids to assist the medical diagnosis process and have developed a number of commercial expert systems, with sophisticated algorithms that are able to generate possible diagnostic suggestions from a given set of patient information. These recommendations can be based on rules of inferences from expert physicians as well as epidemiological data, or the frequency of occurrence of particular diseases.

In automated decision aids for medical diagnosis, automation is typically applied at varying levels to the stages of information acquisition, information analysis and decision selection. Information about the patient can be obtained automatically from the various monitoring equipment or electronic medical records system or it can be manually inputted by the user. Analysis of the information to form an understanding of what is happening to the patient can be done using algorithms to process the available

data at a high level of automation. At a lower level, the automation can provide appropriate visualization of the data to be processed by the human. Based on the analysis of the patient information, the aid can then provide a single or multiple suggestions for the probable diagnoses as well as similar options for appropriate patient management actions that can be taken. The action resulting from the decision such as drug administration or the execution of a medical procedure is usually implemented by the doctor himself.

Interest in computer-assisted diagnostic decision making began in the late 1960s during which one of the earliest computerized decision aid systems, the Leeds Abdominal Pain System, was developed by the University of Leeds. The system used Bayesian reasoning based on an extensive database of patient records to calculate the probability of seven possible diagnoses for acute abdominal pain. In a controlled experiment, the system was able to diagnose accurately for 91.8% of the experimental cases, compared to the 79.6% success rate of the most senior member of the clinical team who assessed the same experimental cases (de Dombal et al, 1972).

Since then, advances in computing technology have allowed these diagnostic decision aids to mature significantly with features such as more complex algorithms, the ability to leverage on the Internet and the ability to directly access electronic health records (Bond et al, 2012). Examples of commercially available clinical diagnostic decision aids include DXplain, QMR, Isabel, and DiagnosisPro.

While some of these systems focus on specific disease categories, some, like DXplain, are also able to generate a list of differential diagnoses from a more general range of medicine. DXPlain is a computer-based reference and decision support system that is able to accept a list of clinical manifestations and then generate a ranked list of

possible hypotheses for the user to consider (Barnett et al, 1987). Leveraging on the Internet, the DXplain is able to constantly update itself with the input of its many users and is apparently one of the most well-used diagnostic decision support tools that is available today (Barnett et al, 1998; Musen, Shahar, Shortliffe, 2006). In a study to evaluate the level of satisfaction of using the DXplain system among internal medicine residents in an actual hospital setting found that 83.3% of its users found the system useful, 11.1% of the users were unsure and 5.6% did not think it was useful (Bauer et al, 2002). Furthermore, 72.2% of them would like to have the system regularly available, 22.2% were unsure and 5.6% did not want it available.

A review of studies on the effects of various clinical decision support systems on performance and patient outcomes by Garg et al (2005) found 10 studies which evaluated clinical diagnostic systems. Among these studies on diagnostic systems, the review found that practitioner performance improved in 40% of them. Two of the studies found that the diagnostic decision aid reduced the rate of unnecessary hospital admissions for cardiac ischaemia by 15% while the aid in another study reduced the time to diagnosis for acute bowel obstruction from 16 hours to 1 hour (Pozen et al, 1984; Selker et al, 1998; Boguševičius et al, 2003).

As advancements in this field of diagnostic decision aids progresses, we can expect future systems to be more accurate and more robust. However, it is highly unlikely that the human decision maker will ever be completely taken out from the diagnosis process. Apart from legality and liability issues for expecting a human doctor to have the ultimate responsibility over the final diagnosis, there will always be unexpected circumstances in certain diagnostic scenarios where subtleties in the situation or the recognition of unusual combination of factors are often the key determinants to a proper decision. These deviations from the norm may not have been predicted by the

machine designers may be overlooked or misinterpreted by the machine. The human's adaptability and creativity still plays a crucial role in the interpretation of the data. Therefore, the designers of clinical decision aids cannot afford to completely exclude the human decision maker from the system and need to recognize that their design must be able to support a cooperative relationship between the automation and the human physician (Musen, Shahar, Shortliffe, 2006).

2.5 Acceptance of and Reliance on Automated Decision Aids

A lot of research effort since the early days of computerised clinical decision aids have gone into the development of more and more sophisticated algorithms that are able to process large amounts of patient data as well as exercise complex expert knowledge rules. However, the improvement of accuracy and reliability of these decision aids are not the only factors that will determine its acceptance. Despite the relative success of various automated diagnostic decision aids as shown in different studies, this has not translated into widespread clinical use (Garg et al, 2005). Other factors that act as barriers to the implementation of automated decision aids in clinical practice include the failure of the clinicians to use the systems, which may be caused by poor usability or poor integration into the clinician's workflow, and also the general non-acceptance of recommendations by the user (Reisman 1996).

Usability and workflow integration are issues that can be tackled with better understanding of user interface designs and better systems designs that minimize the manual effort required to use the system. This could make access to the system more seamless and more efficient, with minimal disruption to workflow. However, even with ease of access, tackling the general non-acceptance of the aid's assistance is

something that requires researchers to carefully look into how the clinical decision maker interacts with the decision aid and the information it provides.

2.6 Decision Making with Automation

Considering the frailties of human decision making while under the stresses of a naturalistic emergency setting, it is common strategy in various work contexts to turn to technology and automation to mitigate human weakness and reduce decision errors. As described earlier in the continuum between full manual control and full automation control, automated decision aids lie between Levels 2 to 5 where the automation provides a single or a set of suggestions and the human still plays a role in the system to act upon the suggestions provided by the aid (Parasuraman, Sheridan & Wickens, 2000).

Figure 2-3. Hybrid Lens Model of operator's judgment with automated decision aids by (Seong & Bisantz, 2002)

In the clinical diagnostic decision aids like DXplain and DiagnosisPro, the automation carries out its own process of cue integration, hypothesis generation and hypothesis evaluation and provides diagnostic suggestions as additional informational cues for the user to consider in his own diagnostic decision making process. This can be illustrated by the Hybrid Lens Model of operator's judgment with automated decision aids proposed by Seong & Bisantz (2002) as shown in Figure 2-3.

In this model, the true state of the environment (ENV) generates cues (X_n), which are processed into a judgment independently by the automated decision aid (ADA) and the human operator (HO). HO_{ADA} represents the judgment that is made when the human operator combines his own independent judgment with that of the automated decision aid.

Within this model, the effect of the automated decision aid on the human operator depends on the interaction between the decision aid and the operator. In reality, the decision by the human operator to actually make use of the information from the automated decision aid is not a given. There are several factors that can influence whether or not the human chooses to rely on the automation to achieve the desired task objective. These include task complexity, mental workload, the effort required to operate the automation, expertise, self-confidence in carrying out the task manually, the perceived risk of using or not using the automation, attitudes towards automation and trust (Parasuraman & Riley, 1997). In a framework developed by Dzindolet et al (2002) as shown in Figure 2-4, the usage of an automation is predicted by cognitive processes such as biases, motivational processes such as the value of the outcome and the effort, and social processes such as trust, among other factors such as task difficulty and the environmental context.

Figure 2-4. Framework to predict automation use (Dzindolet et al., 2002)

Among these factors, there is a wide range of literature that points to user's trust in the automation as one of the major factors that often influences whether the user accepts and uses the automation (Dzindolet et al., 2003; Lee & Moray, 1992; Lee & See, 2004; Muir, 1987; Wickens & Hollands, 2000).

2.7 Trust in Automation

Many researchers have pointed out that the trust between human and automation is analogous to the trust between human and human and that many of the factors that influence trust are applicable in both cases (Lee & See, 2004; Muir, 1987; Muir & Moray, 1996; Parasuraman, Molloy & Singh, 1993; Sheridan, 1975). In the same way that humans typically do not rely on the aid of other humans who they do not trust, humans also tend to avoid having to much reliance on machines that they perceive as

untrustworthy. The more trustworthy the actions of the automation are, the more likely the user would allow the automation's action to play a role in completing the objectives of the human's task. It is therefore important for the designers of automated diagnostic decision aids to understand how they can increase the trust between their designs and the user to promote its usage.

In social interaction, Mayer, Davis and Schoorman (1995) defines trust between two parties as the willingness of the trustor to be vulnerable to the actions of the trustee based on the expectation that the trustee will perform a particular action important to the trustor, irrespective of the ability to monitor or control the trustee. In the same way, when a user places trust in an automation to perform a task, the user expects the automation to achieve the task objective without necessarily being able to have full influence over the automation's actions and the user is willing to be vulnerable to the consequences of the outcome of the automation's actions.

Parasuraman & Miller (2004) refer to trust in automation as the willingness of the user to believe the information provided by the automation or make use of its capabilities. Lee & See (2004) refers to trust in automation as the user's attitude that the automation will help to achieve the user's goals in a situation where uncertainty and vulnerability is present.

In one of the earliest experiments on the effects of trust on the interaction between human and automation by Lee & Moray (1992), participants were asked to maintain the performance of a semi-automated pasteurization plant by adjusting different parameters of the plant's system using either manual or automatic controls. To simulate real-life complexities in process control, system faults were introduced during the experimental trials. They found that the allocation of control between the manual

and automatic controls was dependent on the trust in the automatic controls as well as the confidence of the participants in their ability to control the plant manually. In an extension of the experiment, Lee & Moray (1994), found that operators would rely on the automatic controls when trust in the automation exceeded their confidence to control it manually and would rely on to manual controls when it is vice versa. Muir & Moray (1996) later found that there was a positive correlation between the use of an automated controller and the operator's trust in the automation. Automation that were trusted were utilized by the operator but automation that could not be trusted were rejected.

In another experiment by Dzindolet et al (2003), participants were asked to detect the presence of a camouflaged soldier in a still image of terrain. A pseudo-computer program provided the participants with information on the likelihood that the soldier was present. Even with limited interaction with the aid prior to the actual experimental trials, participants believed that the automated decision aid was trustworthy and expected it to perform better than humans. However, after experiencing errors by the decision aid during the experiment, there was a decline in trust and a corresponding decline in automation reliance in favour of self-reliance.

2.8 Effect of Automation Errors on Trust

Trust can be a cause that determines behaviour and also be an effect as a result of an observation. It influences reliance on the automation and in turn, is affected by the automation's performance (Manzey, Reichenbach & Onnasch, 2012). Trust in the automation builds up with repeated usage and with the increased trust, there is greater reliance on the automation.

As the user interacts with a highly reliable automation over a period of time and experiences constant proper functioning, this can lead to an increased complacency in the automation capability and an increased over-reliance on the system (Parasuraman et al, 1993). It leads to a loss of situation awareness as the over-reliance reduces the attention of the user on the environmental cues. The loss of situation awareness involves a diminished perception of relevant information in the operating environment which then leads to a diminished comprehension of the state of the environment and a diminished ability to predict the future state of the environment (Endsley, 1995).

If after long periods of proper functioning, the automation, being imperfect, unexpectedly fails and makes an error, the increased complacency increases the likelihood that the user continues to rely inappropriately on the automation, which can lead to diminished user performance. This is illustrated in the model of complacency and bias in automation in Figure 2-5 (Parasuraman & Manzey, 2010). After experiencing the automation failure, there is a decreased in automation complacency and increased attention on the task capabilities of the automation.

The decreased complacency potential can be associated with a decrease in trust in the automation. Any sign of incompetence, like an error in automation performance, can reduce the trust and reliance in the system even though there may be no actual change in the system's overall performance (Muir & Moray, 1996). Dzindolet et al (2003) found that even reliable automated decision aids were distrusted and ignored by their users after they had observed the aid commit an error. Manzey et al (2012) found that while positive experiences impacted trust in automated aids positively and negative experiences impacted it negatively, there was a difference in the magnitude between the two feedback loops. Automation failures caused a greater decrease in trust compared to the increase in trust caused by the experience of a reliably working aid.

Figure 2-5. Model of Complacency and Bias in Automation (Parasuraman & Manzey, 2010)

The severity of the error can affect how much impact the error could have on trust. The greater the negative consequence of the error, the greater the trust would drop. Khasawneh et al (2004) found that trust towards an automated printed-circuit-board (PCB) inspection aid was negatively correlated to the severity difference between the aid's decision and the user's own decision. If the aid decides that the severity of a board is low and accepts the board but the user thinks the severity is higher, or if the user feels that the aid over-valued the severity of the board and rejects the board instead of accepting it, the negative trust impact increases according to how different the user's perceived severity is compared to the aid's assessment. Accepting a highly problematic PCB or rejecting a perfectly fine PCB reduced trust more compared to accepting or rejecting a slightly problematic board.

The complexity of the error also affects the impacts of the error on trust. Users tended to be more forgiving to the automation if the error was during a difficult task that they

could have committed themselves. Madhavan, Wiegmann & Lacson (2006) found that when users experienced errors from an automated decision aid in easy tasks that the users themselves could have performed, trust was severely degraded and they were more likely to ignore the aid during difficult tasks, compared to users who only experienced automation errors in difficult tasks.

Although trust declines rapidly after the experience of failure, it can be expected to recover again if the system thereafter performs continuously without making errors (Muir, 1987). Experiencing continuous reliable performance can be expected to build trust. However, with a lower complacency due to the error, the user will consider the aid's assistance more cautiously. As such, it will require several error-free uses before it will recover to previous levels (Lee & Moray, 1992; Rovira, McGarry & Parasuraman, 2007). Kantowitz, Hanowski & Kantowitz (1997) found that users of an advanced traveller information system lost trust in the system after experiencing inaccurate information but recovered it with subsequent accurate information, albeit not completely to initial levels. Manzey et al (2012) found that after experiencing an early error in the use of a highly automated decision aid in a supervisory process control task, it took twenty error free uses of the aid before trust recovered to levels that were similar to users who experienced no early errors. However, if the automation makes the same error consistently, the user may be able to predict the occurrence of the error and calibrate his reliance accordingly (Muir & Moray, 1996)

2.9 Appropriate Reliance on Automation

Forming a medical diagnosis is a complex task that many doctors may believe that only a human can perform more effectively. Given the stakes involved in identifying

the correct diagnosis, to what extent would a doctor be able to trust the diagnostic judgment of a machine and allow it to influence his decision? Newman-Toker & Provonost (2009) suggested that because there is an expectation for the decision aid to have perfect accuracy, their apparent performance always appears to fall short to the user. This perception of the decision aid limits their general trust in the system and their acceptance of it.

In reality, perfect accuracy is unlikely to be an unachievable target for decision aids in complex decision making environments. Users will have to accept a certain margin of error to be existent in these types of decision automation. Even robust automation designs can perform below expectations at some point during its operation. Some sources of sub-optimal automation performance could be due to shortcomings in the design of the automation's internal processes such as the inability to operate within unexpected circumstances, or even some form of damage to the automation's sensors that compromises its ability to gather the environmental cues accurately. Human operators realize this imperfection in automation. In its usage, they need to decide when it is appropriate to rely on the automation to accomplish the task and when it is not. If a physician chooses wrongly when deciding whether to accept or reject a suggestion from an automated diagnostic aid, he may initiate the wrong treatment action and put the patient at risk of further harm.

Inappropriate usage of automation by the operator can be described as the disuse and misuse of automation (Parasuraman & Riley, 1997). Disuse of automation refers to situations where the automation performs properly but it is not used or disregarded by the operator. For example, a nurse may ignore a genuine system alarm in an intensive care unit, dismissing it as a false alarm. Misuse of automation refers to situations where the operator relies too much on the automation when the reliance on it is not

warranted because the automation is functioning improperly. An example could be when a pilot delegates the navigation of the plane to the auto-pilot while not realizing that the GPS tracking system is faulty.

This vulnerability to the consequences of disuse or misuse in the dependence on automation forms a critical element in defining trust in automation (Lee & See, 2004; Mayer et al., 1995). As a reflection of the operator's willingness to place himself in a vulnerable position by assigning some responsibility for the success of the task to the automation, trust in the automation is one of the key factors that will guide whether the operator should rely on the automation. Inappropriate reliance in terms of disuse and misuse is associated to how well the operator's trust in the automation calibrated against the capability of the automation which reflects its trustworthiness (Lee & Moray, 1994; Lee & See, 2004; Muir, 1987). This is illustrated in Figure 2-6. When trust matches the capability of the automation, the trust is well calibrated and the use of the machine is appropriate. Disuse of automation occurs when there is distrust of the system, where the trust is lower than the system's capabilities. Misuse occurs when there is over-trust where the trust exceeds the capabilities.

The resolution of the trust refers to how accurately the trust of the operator varies with respect to the changes in the automation's capabilities (Lee & See, 2004). As shown in Figure 2-7, when resolution is good, changes in the trustworthiness of the automation maps onto the same changes in trust but when resolution is poor, large changes in the trustworthiness of the automation maps onto a smaller change in operator trust leading to deviations from calibrated trust.

Figure 2-6. Calibrated trust, over-trust and distrust from the operator and the trustworthiness of the automation (Lee & See, 2004)

Figure 2-7. Resolution of trust in automation (Lee & See, 2004)

Trust also has functional specificity and temporal specificity. Functional specificity refers to how much trust is associated to specific components of the automation. Trust at higher functional specificity denotes the operator's trust in specific sub-functions or modes of the automation while trust at lower specificity reflects the operator's trust in

the overall system. Temporal specificity refers to changes in trust over time or over different contexts. At higher temporal specificity, trust reflects the dynamic moment to moment changes in the perceived capability and trustworthiness of the system due to specific fluctuations in the operating environment. At lower temporal specificity, trust reflects the changes in the automation's capability over the long term.

While promoting trust in automation is important to increase acceptance and usage, system designers need to be aware that too much trust can also be detrimental especially when the system eventually commits a mistake. Therefore, it is not enough for the designer to simply design to increase operator trust in the system. Instead, it is far more crucial for designers to ensure that the system is able to afford the operator a good resolution of trust by providing the operator a way to perceive its trustworthiness and calibrate his trust appropriately and the necessary levels of specificity (Lee & See, 2004; Muir, 1987). If there exists the potential benefit to outperform unaided decision making when the aid is performing reliably, the challenge to the ADA's designer is then to design for the user to know when the system can be relied upon and when to use his own diagnostic skill.

2.10 Factors that Influence Trust in Automation

Muir (1987) had extrapolated based on the literature on interpersonal trust in social interaction that human trust in machines depended on the machine's predictability of its actions, dependability of its processes and faith in its intentions. These dimensions were applied to the expectations of the persistence of physical laws and the moral social orders that govern that man and its machine are good, the expectations of the

technical competence of the machine in performing its roles and the expectation that the machine will carry out its fiduciary responsibilities.

Similarly, Lee and Moray (1992) had summarized that trust between human and machine similarly depended on three basic dimensions which are performance, process and purpose. Performance refers to the expectation that the automation will operate consistently, stably and desirably in carrying out its intended task. Process refers to the understanding of the underlying algorithms, methods or rules that determine how the automation operates. Purpose refers to the intention of the designer for the automation, describing the reasons for the development of the automation.

The framework by Lee & See (2004), as shown in Figure 2-8, emphasizes that trust in an automation is not a static concept but constantly evolves with further interaction with the automation. This dynamic nature is captured in the closed loop between trust and reliance with the automation. Trust influences the intention to rely on the automation. As the operator uses the automation and observes the consequences of the reliance, this in turn influences belief formation and evolves the trust which then again influences whether the reliance should continue. This interaction is moderated by the characteristics of the individual, the organization, the culture and the environment. One important factor to be noted by the designer of the automation is that how trust evolves also depends on the characteristics of the automation and the information that is displayed to the operator on its interface.

Figure 2-8. Dynamic model of trust and reliance on automation (Adapted from Lee & See, 2004)

According to Lee & See (2004), the decision to rely on an automated system is influenced by the interplay between the individual's analytic, analogical and affective processes that underlie trust. The analytic processes involve the individual accumulating knowledge about the system allowing him to estimate the probabilities of various outcomes from the decision aid. However, such analytic processes may be cognitively demanding to be the sole basis for the reliance decision. Analogical processes are less cognitively demanding processes that relate the level of trust to the characteristics of the system and its environmental context, as developed through direct observations of the system, information about the system as conveyed by other observers and presumptions about the system. As it can be too cognitively complex to build trust based only on analytic and analogical processes, the individual also depends on affective processes or their emotions to guide their trust behaviour when cognitive resources to make a purely rational decision are not available. Therefore to guide appropriate trust and reliance, the automation's display which acts as mediator

between the user and the internal mechanisms of the automation should convey automation-related information in a way that supports assimilation of the information through these analytic, analogical and affective processes.

Sheridan (1988) suggested several characteristics of automation that affect how much trust an operator would place on the automation, which are reliability, robustness, transparency, familiarity, explication of intention, dependability, usefulness, validity and understandability.

Seong and Bisantz (2002) summarized these different attributes into three major characteristics for automated decision aids which are reliability, validity and understandability. Reliability refers to the consistency of the automation to produce the same outcome given the same set of informational cues. Previous studies have shown that the typically, the minimum perceived reliability before users would rely on an imperfect automated system lies around 60% to 70% (Kantowitz et al, 1997; Wickens & Dixon, 2007; Wiegmann, Rich & Zhang, 2001; Yeh & Wickens, 2001) Validity refers to the proximity between the aid's estimate and the truthful state of the environment. Understandability refers to the degree to which the user understands how and why the system behaves the way it does.

Other aspects of decision aids that have been found to influence trust include the level of automation (Parasuraman et al, 2000; Manzey et al, 2012), the etiquette in which the system delivers its assistance (Parasuraman & Miller, 2004) and even the degree of image realism in the aid (Yeh & Wickens, 2001).

2.11 Automation Transparency

In high risk scenarios, the user of the automated decision support, the designer of the automated decision support and the automated decision support itself need to function as “partners” in a cooperative decision making team, with avenues for appropriate feedback and communication between them (Rasmussen & Goldstein, 1987). To allow the automation to be an extension of the user’s decision making capabilities, designers need to implement facilities that provide these avenues so that operators can have accurate mental model of what the machine is doing (Rasmussen & Goldstein, 1987). The operator’s mental representation of how the machine works or his thought processes about the relationships between the machine’s output and input should be a good representation of how the machine actually works. This allows him to make his own predictions about the behaviour of the machine.

A lot of automation operates as a “black box” where the user is not privy to the internal processes of the automation. The user is usually only able to perceive the input and the output, without knowing what happens in between. However, the understandability of the system, or the extent to which the operator is able to make sense of why and how the automation is doing what it is doing, is one of the system characteristics that is expected to influence the trust that an operator has in the automation (Seong & Bisantz, 2002; Sheridan, 1988).

One way of increasing the understandability of a system is to train the operator on how it works before its actual use (Lee & See, 2004). This allows the operator to form a mental model of how the machine processes the inputs into the output. While this gives the operator a general understanding of the overall system, a design dimension that can be manipulated by the designer of the system is by making the “black box”

more transparent, that is by building some information related to the internal workings of the system into the display interface so that the operator has access to some information on why the machine is performing certain actions and how well it is doing so.

According to Koenderink, Broekstra & Top (2010), a “black box” approach to design may be acceptable for automation where the designers are able to anticipate all expected circumstances and the automation can perform its task well at all times. In those situations, it does not matter to the user how the machine does it as long as it does it right all the time. However, they reasoned that for knowledge intensive applications, like medical diagnostic decision aids, which require extensive expert knowledge, it is not possible for the designer to master all the required domain expertise in order to design the machine to function optimally for all possible scenarios. Hence, a completely “black box” design is not appropriate as the automation needs to be able to explain its complex reasoning patterns to the user who will act as the moderator for the final action. The user needs to know this so that he can ensure that the automation is performing reliably and safely in a decision making environment that can have many deviations from the normal because of factors like information uncertainties, technical faults and inappropriate actions from other humans (Rasmussen & Goldstein, 1987). Herlocker, Konstan & Riedl (2000) have pointed out that the inability to question the “black box” to determine whether to trust or doubt its advice is one of the reasons why such designs have limited acceptance in all but low-risk content domains.

Between human decision makers and the humans who are affected by those decisions, open communication or communication transparency between superiors and subordinates is considered to be an important factor for an effective organization

(Rogers, 1987; Myers et al, 1999). Transparent communication is characterized by how individuals in the organization exchange information with one another and how receptive and responsive they are to the information provided. Leaders who are transparent about the decisions they make are found by their subordinates to be more trustworthy and more effective (Norman, Avolio & Luthans, 2010).

Drawing from the literature on advice giving and taking, van Dongen & van Maanen (2013) suggested that people tend to be more biased to their own opinions more than to that of others because their own reasons that support those opinions are cognitively more available and are brought to mind more effortlessly (Tversky & Kahneman, 1973). The advice from the advisor tends to be under-used by people because of the concept of differential information, where people have no direct access to the advisor's reasoning behind the advice while having direct access to their own reasoning process behind their own judgment (Yaniv & Kleinberger, 2000). Therefore, conversely, when the reasoning of the advisor is made more transparent and more cognitively available, it should make it more likely for the recipient of the advice to accept it.

The issue of the effects of transparency, when applied to automation, has been explored by a number of previous researchers in its use in user-adaptive recommender systems (Cramer et al, 2008; Herlocker et al., 2000; Sinha & Swearingen, 2002; Tintarev & Masthoff, 2007). These recommender systems are designed to help the user find items that they are likely to find interesting by learning the preferences of the user, processing the abundance of parameters from the relevant items and then providing item suggestions to the user based on their profile (Burke, 2002). Examples include the product recommendation systems that are used by online commerce websites like Amazon to help customers make decisions on which products to choose from a wide range of their products. These studies have argued that by providing features that

increase the transparency of the system by providing explanations for its recommendation, user satisfaction and trust in the system can be improved.

Herlocker et al (2000) found that users had a positive reaction to inclusion of explanation facilities in a web-based movie recommender system. Sinha & Swearingen (2002) conducted a user study on existing music recommender systems and found that the systems that make the user feel that they understood why particular recommendations were presented scored higher on user liking and user confidence than those that did not. They suggested that users were not satisfied with just blind recommendations but expected the system to justify the choices that it made for them. Cramer et al (2008) studied the effects of introducing transparency features on a content-based artwork recommender system. One feature was confidence ratings that showed how confident the system was that the artwork would be interesting to the user and the other feature was an explanation feature that listed the properties of the artwork that were similar to those of artworks the user had rated positively. They found that users were more likely to accept recommendations from a system that explained its recommendations. However, they found that transparency did not affect user trust of the overall system. They pointed out that it was possible that the increase in trust in the system was cancelled out by the lowered perceived competence of the transparent system, which also has a major influence on trust. Users could see the reasons for the recommendation but did not find the criteria for recommendation suitable and that the aesthetic and affective dimension that would influence one's interest in an artwork could not be addressed by the system.

Herlocker et al (2000) had identified four main benefits of transparency in recommender systems. Firstly, users are able to justify how much confidence they are able to place in a recommendation. Secondly, users are involved in the

recommendation process allowing them to factor in their own inputs in the decision process. Thirdly, users are educated to the processes that the system uses to generate recommendations and this would let them gain a better understanding of the system's strengths and weaknesses. Fourthly, there would be greater acceptance of the system when its limits and strengths are fully visible and its suggestions are justified.

Sinha & Swearingen (2002) had stated that the recipient of a recommendation in a human-human recommendation-making interaction decides whether to trust the recommendation by scrutinizing how similar the recommendation is to the recipient's own tastes, by assessing how successful the recommender was in previous recommendations and by asking the recommender for the reasons behind the recommendation. Drawing an analogue of this for the automated diagnostic decision aid, the user of the aid would assign trust based on the similarity of the aid's recommended diagnosis is to his own diagnosis, the success of the aid in previous recommendations and the ability to assess the aid's reasons for providing that recommendation.

For an automated decision aid software, the complexity of the machine language may be one of the reasons which deters designers from exposing the inner workings of the system. For machines with more complex algorithms, users may be less willing or less able to learn about the machines internal processes (Parasuraman & Miller, 2004). Herlocker et al (2000) had found that poorly designed explanations limits the acceptance of recommendations. However, given the potential benefits of providing transparency, the challenge of converting complex machine reasoning into a form that can be understandable to the user is important challenge that designers of such decision aid systems need to address (Herlocker et al, 2000). For example, developers of the MYCIN system, an early expert system used to identify bacteria that causes

severe infections and recommend antibiotics, had included a design feature that explained its recommendation to its user by translating its rule-based reasoning in the LISP programming language into plain English for the user to understand (Buchanan & Shortliffe, 1984).

2.12 Key Diagnostic Cues and Likelihood Ratings for Transparency

To increase the degree to which the user is able to understand the automated decision aid and its recommendations and thus be able to calibrate his trust appropriately, designers of the automation should look at transparency features for the display interface that afford the understandability of the three basic dimensions of trust as identified by Lee & Moray (1992) which are purpose, process and performance.

For transparency of purpose, information on the system's intent or what it is designed to do has to be apparent. For a medical diagnostic decision aid, its purpose is to provide a suggestion for a possible diagnosis. This could be afforded through the use of clear labelling for displayed information or through instructions to the user prior to use.

For transparency of process, the diagnostic decision aid needs to provide information to the user that relates to the rationale behind its recommendation, such as the key diagnostic cues that were used by the aid to generate the recommendation. If the automation's reliability is not perfect, the automation should provide information that allows for the user to inspect and analyse the raw data that is relevant to the task (Rovira et al, 2007). Lim, Dey & Avrahami (2009) had found that an intelligent system that automatically provided explanations as to why it provided a particular output

improved understanding of the system, system trust and task performance. Providing such information lets the user inspect the system's reasoning, that is to evaluate the appropriateness of the process used by the system to come up with recommendation from the data, and allows him or her to detect errors in the diagnostic process or errors in the data (Herlocker et al, 2000).

Finally for transparency of performance, the diagnostic decision aid should provide information on how sure the system is for its given recommendation or a rating of the likelihood that the recommendation would be correct. Given that the aid is imperfect and could possibly be making an error, the human decision maker who needs to decide on whether to rely on the aid needs to be aware of the degree of uncertainty in the machine. Sorkin, Kantowitz & Kantowitz (1988) first explored the use of incorporating likelihood information into alarm displays that conveyed information about the systems's confidence in the certainty of the alarm. It was found that users better allocated attention between multiple tasks including the monitoring of the alarm. McGuirl & Sarter (2006) use a line graph to indicate dynamic system confidence of a decision aid to assist pilots in detecting handling in-flight icing encounters and found that the presence of the uncertainty information improved trust calibration and led to better performance. Beller, Heesen & Vollrath (2013) showed that presenting uncertainty information on an automated driving aid increased trust and acceptance of the system compared to when it was not present. It also increased the user's knowledge of fallibility, that it was possible for the automation to make an error. This led them to account for it in their understanding of the system and thus be more prepared for automation failure.

2.13 Research Gaps

Previous literature have suggested that a more transparent decision aid is more understandable than a decision aid that functions as a “black box” because it lets the user better calibrate his trust towards it (Seong & Bisantz, 2002; Sheridan, 1998; Herlocker, Konstan & Riedl, 2000). Previous studies have investigated how a more transparent automated decision aid can provide more information about itself and influence the beliefs and trust of the human towards it. They have shown that transparency is beneficial to the interaction between human and machine by increasing trust, user acceptance of the machine’s recommendations and overall user satisfaction (Cramer et al, 2008; Herlocker et al., 2000; Sinha & Swearingen, 2002; Tintarev & Masthoff, 2007).

However, none of these studies have quantitatively investigated if these benefits will occur in a decision making environment that is more high-risk in nature. Automated decision aids can be incredibly helpful to human decision makers in such complex decision making domains. However, since the automation cannot be expected to be perfect, correct calibration of trust towards the aid’s recommendation and knowing when to rely on the aid and when not to rely on it is more critical because wrong decisions may have more serious consequences.

The study addresses this gap by providing evidence on the effectiveness of transparency on improving trust in automation within the context of medical emergency decision making. Compared to the previous studies, this decision making domain involves time pressure and higher stakes and could be a better representation of how transparency affects automation trust in high risk decision making.

Another gap that has not been sufficiently covered in the literature on automation transparency and trust is the effectiveness of transparency in addressing trust recovery after an imperfect decision aid has committed an error. Previous literature have shown how detrimental automation errors can be towards user trust (Muir & Moray, 1996; Dzindolet et al, 2003, Madhavanm Wiegmann & Lacson, 2006; Manzey et al, 2012). Even when the automation is functioning correctly, a previous experience of automation error makes the user more suspicious of the automation and it takes several error-free uses before the user would trust it as much as he has before the error. If the system is more transparent, it allows better calibration of trust because the user can understand the automation's rationale and justify his level of trust towards it and it also allows him to be more involved in the final decision (Herlocker et al, 2000). Compared to a "black box" automation, transparency features could potentially help mitigate the trust detriment resulting from a previous error experience and improve trust recovery. However, there has been a lack of studies to investigate if transparency can benefit trust in this aspect and this study fills the gap by investigating how transparency affects the drop in trust that results from a previous experience of automation error.

This study addresses the need for empirical evidence to better understand how transparency affects the dynamics of human-machine collaborative decision making in these critical environments. The knowledge can lead to a better understanding of how to design decision aids that allow the user to better calibrate their trust towards it.

CHAPTER 3: CONCEPTUAL MODEL

3.1 Model for Transparency and Reliance on Imperfect Automated Aid

The diagram in Figure 3-1 represents how trust in automation is involved during the interaction with an imperfect automated diagnostic decision aid. It adapts Seong & Bisantz (2002) Hybrid Lens Model and expands it to include trust in automation as a mechanism that influences the reliance on the automation's aid during its performance.

Figure 3-1. Model for Transparency and Reliance on Imperfect Automated Aid for Diagnostic Decision Making

The true state of the environment for the context of the medical emergency setting in the current study comprises the patient, the accompanying medical history of the patient and the condition that the patient is going through. The objective of the decision maker (the physician) is to identify the most likely state of the environment, which in this case is the diagnosis for the condition that the patient is experiencing.

The environment will contain informational cues from which a subset is sensed and processed by the automated diagnostic decision aid's sensors and another subset is perceived by the human decision maker. The information processed by both the automation and the human is incomplete. For the automation, it is limited by its sensing and processing capabilities. For the human, it depends on the information value of the cues and their salience (Wickens & Hollands, 2000).

The automated decision aid processes its subset of informational cues and produces a recommended diagnosis for the true state of the environment. If the automation fails, it produces a wrong recommendation and if it performs correctly, it produces the correct recommendation.

The human decision maker also processes the subset of cues that had been perceived from the environment in order to come up with the final diagnosis. Aware that the automation is imperfect and may make errors, the human needs to decide whether to rely on the recommendation or to ignore it and rely completely on his own decision making process to produce the final diagnosis.

This decision occurs while the human is subjected to the conditions of a naturalistic emergency environment (Orasanu & Connolly, 1993; Rittel & Webber, 1973; Wears, 2009). Given the uncertainty, time pressure and high risk of this critical decision making environment, the human needs to make the most efficient use of the available

cues and come up with the most likely diagnosis as quickly as possible. With the automation aiding immediately providing a possible diagnostic option to consider, it may save the human some time and effort for the decision making process (Röttger, Bali & Manzey, 2009; McGuirl & Sarter, 2006).

The reliance decision is also influenced by how much trust the human feels is appropriate for the automation at that point in time (Lee & See, 2004; Muir, 1987; Muir & Moray, 1996). The human decision maker assimilates information about the automated system and forms his beliefs towards it from its interface features and previous experiences with it (Lee & See, 2004; Parasuraman & Manzey, 2010). One property of the automation that can influence his beliefs and affect the human's calibration of trust is the transparency of the system (Seong & Bisantz, 2002; Sheridan, 1988).

Depending on whether the automation fails or performs correctly and whether the human decides to rely on not rely on the automation's recommendation, there are four possible reliance outcomes as shown in Table 3-A below (Lee & See, 2004; Parasuraman & Riley, 1997).

Table 3-A. Possible outcomes during reliance on an imperfect automated diagnostic decision aid

		Automated Diagnostic Decision Aid	
		<i>Automation Failure</i>	<i>No Automation Failure</i>
Human Decision Maker	<i>Rely on Automation</i>	Over-Trust	Appropriate Reliance
	<i>Don't Rely on Automation</i>	Appropriate Non-Reliance	Distrust

The consequence of the final diagnostic decision can evolve the trust towards the automation in future usage, creating a positive or negative feedback loop (Lee & See, 2004; Parasuraman & Manzey, 2010). If the automation does not fail, the human will either experience a successful outcome from appropriately relying on it or a negative outcome from ignoring the recommendation. In either case, the human has experienced the ability of the automation and may trust it more in future. On the other hand, if the automation were to fail, it creates a negative impression of the automation, which may lead to the human trusting it less in future.

3.2 Transparency and Performance during Automation Non-Failure and during Automation Failure

Figure 3-2 describes the first conceptual model for this study which relates the effect of transparency configuration on trust and decision confidence.

Since the study is concerned with appropriate trust and reliance, it was necessary to differentiate when the automation is performing properly without failure and when the automation fails. In either case, the level of trust that the user has in the system influences the decision to rely or to not rely on the automation (Lee & Moray, 1992; Muir & Moray, 1996; Parasuraman & Miller, 2004)

Figure 3-2. Conceptual Model for effect of transparency configuration on trust and decision confidence

The human decision maker's trust towards the automation is influenced by the contents of the interface between him and the automation (Lee & See, 2004). Trust in the automation evolves based on the human's perceptions of the automation's purpose, process and performance (Lee & Moray, 1992). In this model, we assume that the human decision maker understands the purpose of the machine, that it is designed to provide him with a recommendation for the interpretation of the true state of the environment. The human's perceptions of the aid's process and performance, on the other hand, can be supported by the presence of transparency features.

A transparent interface affords the user a greater understandability of the system which is expected to influence trust and appropriate reliance (Herlocker et al, 2000; Seong &

Bisantz, 2008; Sheridan, 1988). The transparency configurations used in this study is made up of two features, Key Diagnostic Cues and Likelihood Rating.

Key Diagnostic Cues support the perception of process by providing the human with informational cues on the reasons underlying the recommendation by the decision aid. The Key Diagnostic Cues allows the user to scrutinize the machine's rationale against their own reasoning and lets him justify whether to place his confidence in the recommendation (Sinha & Swearingen, 2002; Herlocker et al, 2000).

Likelihood Rating supports the perception of performance by providing the human with informational cues that allow the human to determine how reliable the automation is at the time that it is providing the recommendation. The additional likelihood information to accompany the recommendation provides the user with a tangible reference to calibrate his trust accordingly (McGuirl & Sarter, 2006).

Trust & Reliance during Automation Non-Failure

When the aid performs properly and provides a correct recommendation, the presence of these features will increase the tendency that the human will perceive that the automation's process and performance is acceptable and improve the trust and acceptance in the correct recommendation (Lee & See, 2004).

H1A: *The presence of the Key Diagnostic Cues and/or Likelihood Ratings will increase trust in the decision aid's recommendation during non-failure*

If the human agrees with the aid and appropriately relies on its recommendation, the additional cues from the transparency features will provide additional justification for

the chosen diagnosis. This is expected to increase the confidence of the decision maker in the final diagnosis if he agrees with the automation.

H1B: *The presence of the Key Diagnostic Cues and/or Likelihood Ratings will increase confidence in diagnosis during appropriate reliance*

However, if the human does not agree with the aid even though it is performing correctly and chooses to distrust the aid, the justification for the recommendation from the transparency features is expected to raise doubts about the human's self-generated diagnosis by allowing him to compare his reasoning process and performance against the automation's own. Where there may be differences, the transparency features would increase the salience of those differences and make the user reconsider how correct he feels about his own reasoning. Hence, even if he proceeds to distrust the aid, it is expected that the confidence will be lower if the information from the transparency features are available.

H1C: *The presence of the Key Diagnostic Cues and/or Likelihood Ratings will decrease confidence in diagnosis during distrust*

Trust & Reliance during Automation Failure

In this model, we assume that the human decision maker knows that the automation is imperfect and can fail. And it can fail because of limitations in its sensing and processing capabilities. Ideally, if the machine is failing to provide the correct recommendation, the human should calibrate his trust accordingly by lowering it and relying more on his own self-generated diagnosis (Lee & See, 2004).

Transparency features are intended to assist the human decision maker to calibrate his trust. The justifications from the Key Diagnostic Cues will provide the opportunity for the human to scrutinize and notice weaknesses in the underlying rationale such as doubtful evidence or erroneous information (Herlocker et al, 2000; Sinha & Swearingen, 2002).

The presence of Likelihood Ratings is expected to increase the tendency for the human will consider that the machine is making an error by providing the user with a salient cue that it is imperfect and that there is a possibility for error. This higher awareness of the vulnerability of the automation is expected to increase the decision maker's attention on the environmental cues to notice inconsistencies between the recommended diagnosis and the actual diagnosis.

Therefore, with the presence of the transparency features, the automation failure will be more noticeable and it is expected that it will reduce the trust in the aid's recommendation when the aid is making a wrong recommendation.

H2A: *The presence of the Key Diagnostic Cues and/or Likelihood Ratings will decrease trust in the decision aid's recommendation during automation failure*

If the automated decision aid is providing a wrong recommendation and the human appropriately chooses to disagree with its recommendation, the additional cues from the transparency features is expected to provide additional justification that the recommendation is likely to be wrong. This is expected to increase the confidence of the decision maker in choosing to disagree with the aid and rely completely on his own diagnosis process.

H2B: *The presence of the Key Diagnostic Cues and/or Likelihood Ratings will increase confidence in diagnosis during appropriate non-reliance*

Conversely, if the human decision maker chooses to agree with the erroneous recommendation from the automated decision aid, the additional justification afforded by the presence of the transparency features is expected to increase the perceived likelihood that an error is occurring and thus decrease the confidence when relying on the erroneous recommendation.

H2C: *The presence of the Key Diagnostic Cues and/or Likelihood Ratings will decrease confidence in diagnosis during over-trust*

3.3 Transparency and the Evolution of Trust and Confidence

Resulting from Error Experience and Return to Non-Failure

Experience

Figure 3-3 describes the second conceptual models for this study which relates the effect of transparency configuration on the evolution of trust and confidence as the user experiences the first automation failure followed by a return to non-failure.

Figure 3-3. Conceptual Model for effect of transparency configuration on the change in trust and confidence due to failure and non-failure

Since the automation is imperfect, the user will eventually experience a failed automation performance and experience the negative consequences of it. The complacency potential is expected to decrease and the user is less likely to rely on the aid in subsequent usage without a more careful consideration of the environmental cues (Parasuraman & Manzey, 2010).

Even if the aid is performing without failure in subsequent usage, trust in the aid recommendation can be expected to decline sharply immediately after the error experience (Muir & Moray, 1996; Dzindolet et al, 2003; Manzey et al, 2012). With the decline in trust, it is also expected that the decision maker's diagnostic confidence when appropriately relying on the aid will be lower than it was before the experience of the failure.

Experience of Automation Failure

With the transparency features present, it affords the user the ability to learn about the potential weaknesses of the automation and how the error may have happened. It allows the user to access and scrutinize the information on the diagnostic process and the expected performance of the automation, and assess the potential for error in the current recommendation. Knowing how the automation may have erred or learning of the circumstances that will induce weakness in the automation's diagnostic process can improve trust (Dzindolet et al, 2003).

During the non-failure performance immediately after the experience of failure, the transparency features are thus expected to mitigate the decline of trust due to the experience of failure. With better trust in the aid's recommendation, confidence when agreeing with the aid is also expected to improve and thus decrease less.

H3A: *The presence of the Key Diagnostic Cues and/or Likelihood Ratings will reduce the drop in trust in the recommendation from a decision aid during non-failure immediately after the experience of an automation failure*

H3B: *The presence of the Key Diagnostic Cues and/or Likelihood Ratings will reduce the drop in diagnostic confidence during appropriate reliance immediately after the experience of an automation failure*

Experience of Return to Non-Failure Automation Performance

After the automation failure experience and the accompanying fall in trust, trust will recover as the automation continues to perform without any further failures albeit slower than the rate of decrease due to failure (Muir, 1987; Manzey et al, 2012). Since the transparency features is expected to improve trust, it is thus predicted that the recovery of trust as the user experiences non-failure performance will be greater when the transparency features are present.

H4: *The presence of the Key Diagnostic Cues and/or Likelihood Ratings will improve the recovery of trust in the recommendation from a decision aid during non-failure after the experience of a return to correct automation performance*

3.4 Summary of Hypotheses

Table 3-B summarizes the hypotheses that are investigated in the main study.

Table 3-B. Summary of Hypotheses

	Trust in Aid's Recommendation	Diagnostic Confidence	
		if making the correct reliance decision	if making the wrong reliance decision
During Automation Non-Failure	H1A <i>The presence of the Key Diagnostic Cues and/or Likelihood Ratings will increase trust in the decision aid's recommendation during non-failure</i>	H1B <i>The presence of the Key Diagnostic Cues and/or Likelihood Ratings will increase confidence in diagnosis during appropriate reliance</i>	H1C <i>The presence of the Key Diagnostic Cues and/or Likelihood Ratings will decrease confidence in diagnosis during distrust</i>
During Automation Failure	H2A <i>The presence of the Key Diagnostic Cues and/or Likelihood Ratings will decrease trust in the decision aid's recommendation during automation failure</i>	H2B <i>The presence of the Key Diagnostic Cues and/or Likelihood Ratings will increase confidence in diagnosis during appropriate non-reliance</i>	H2C <i>The presence of the Key Diagnostic Cues and/or Likelihood Ratings will decrease confidence in diagnosis during over-trust</i>
Effect of Experience of Failure	H3A <i>The presence of the Key Diagnostic Cues and/or Likelihood Ratings will reduce the drop in trust in the recommendation from a decision aid during non-failure immediately after the experience of an automation failure</i>	H3B <i>The presence of the Key Diagnostic Cues and/or Likelihood Ratings will reduce the drop in diagnostic confidence during appropriate reliance immediately after the experience of an automation failure</i>	
Effect of Experience of Return to Correct Performance	H4 <i>The presence of the Key Diagnostic Cues and/or Likelihood Ratings will improve the recovery of trust in the recommendation from a decision aid during non-failure after the experience of a return to correct automation performance</i>		

3.5 Flow of Research Study

Figure 3-4 describes the flow for the research study.

Figure 3-4. Flow for research study

Before starting the main study, an exploratory study was conducted to better understand how novice doctors make their own diagnostic decisions without any automation assistance within a simulated emergency setting. Through the exploratory study, potential factors that may lead to suboptimal diagnostic decision making will be investigated. It also serves to build up the researcher's experience and knowledge in medical diagnostic decision making and patient emergency simulation in order to more effectively carry out the main study.

The main study will then investigate the effects of automation transparency on the collaborative diagnostic decision-making between the novice doctor and the automated decision aid.

[INTENTIONALLY LEFT BLANK]

CHAPTER 4: EXPLORATORY STUDY

4.1 Background

Diagnostic decision making in medicine can be a complex task for novice doctors as compared to more experienced doctors. Trainee doctors were found to have been involved in 56% of the misdiagnosis cases studied by Kachalia et al (2007). While under the challenging conditions of emergency decision making, Experts are more likely to be able to adapt themselves to the complexity of the situation so that they can make better decisions (Sandhu & Carpenter, 2006). Novices on the other hand tend to experience degradation in their diagnostic skills while under pressure, leading to problems of information gathering and integration (Graber, Franklin & Gordon, 2005; Kachalia et al, 2007).

Initial observations of medical students participating in medical simulation training involving a rapidly deteriorating patient revealed difficulties among some students in their ability to diagnose the situation. There were delays in identifying what was wrong with the simulated patient and instances of wrong diagnosis. Some students were not even able to form any working diagnosis and either resorted to ordering a wide range of investigative tests in the hopes that one of them would produce a positive result or resorted to waiting for the patient to deteriorate in some recognizable manner that would allow them to identify what was wrong. In an actual emergency situation, where delays in decision-making and action can lead to detrimental consequences, such behaviour may not be desirable.

To gain a better awareness of the problem faced by these novice doctors as they are making decisions during these time-critical scenarios, the initial observations led to the

conducting of detailed interviews with the students in order to elicit their thought processes as they form their diagnostic hypotheses as test them over the course of the scenario.

4.2 Objectives

The objective of this exploratory study is to investigate the possible factors that may lead to breakdowns of diagnosis performance among novice doctors during a simulated patient emergency event.

4.3 Method

Crisis Management Training

The study involved the observations and detailed individual interviews of fourth-year undergraduate medical students participating in a Crisis Management Training course. Each Crisis Management Training session contained a set of six different crisis scenarios involving the METI high-fidelity human patient simulator (HPS). The six conditions suffered by the simulated patient were Severe Asthma Attack, Acute Myocardial Infarction, Hypovolemic Shock, Anaphylaxis, Tension Pneumothorax and Pulmonary Embolism.

The patient simulator was programmed to produce the perceptual cues that are relevant to the above conditions, which includes the appropriate lung and heart sounds, pulses, chest movements and vital signs on the patient monitor. Certain cues that could not be simulated on the patient simulator could be obtained verbally from the patient or the

trainer controlling the simulation. Investigative tests, such as a 12-lead electrocardiogram (ECG) or a chest x-ray, could be requested but there was typically a time delay in the results or the results may not even arrive. The perceptual cues from the patient simulator would also respond accordingly to the treatment actions carried out, such as medication or resuscitative procedures.

The students went through the scenarios in teams of 3 to 5 students. During the scenarios, each lasting 10 to 15 minutes, the students acted as a team of physicians assessing the patient prior to the commencement of surgery. Their objective was to diagnose what was happening to the patient and manage the situation accordingly. Each team identified a team leader who would act as the main decision maker for the team and coordinate the actions of the team members during the scenario.

Before the start of the scenario, the students verbally received some brief patient information and history. While the specific details of the patient history, such as age, race and existing medical conditions, varied between scenarios, a common narrative element in all of them was that the patient had suffered a fracture of the right femur and that he was complaining of breathlessness. The actions of the team during the scenario were video recorded.

Immediately after the scenario and prior to the scenario debriefing, the team of students were required to fill up individual questionnaires on what they thought was the most likely diagnosis, what other diagnoses they had considered during the scenario and what cues they used to derive those diagnoses. The purpose of this questionnaire was to obtain details about the diagnostic thought processes of the students before their views were biased by the information provided during the

scenario debriefing. The questionnaire responses would be used as a reference during the interviews.

Participants

After the entire training session was completed, all the team leaders were invited to participate in an interview that would delve into more detail about their decision making process during the scenario.

A total of 12 students (age range: 21 to 25 years), who acted as the team leaders, were interviewed. The participants were given a token sum of money as reimbursement for their participation in the interview.

Interview Structure

The interview structure was adapted from the interview questions and methods from the Critical Decision Method in incident-based cognitive task analysis (Crandall, Klein, & Hoffman, 2006).

Before the interview began, the participant was asked to review the video recording of his or her scenario and create a time-line of key events that occurred during it. Four types of key event were defined for the participant to identify. The first type was any points in the scenario where the participant had to make any form of decision. Secondly, any points where the participant had made any assessment or diagnosis of what is going on to the patient. Thirdly, any points where the participant had noticed

something important that had changed about the patient. Lastly, any points where the participant had carried out or intended to carry out any plan of actions.

For each key event on the time-line, the interviewer would then probe the participant to elaborate as much as possible about what happened during the key event. Whenever it was relevant, the participant was asked to describe the assessment of the patient or any working diagnosis, any plan of action in response to the assessment, their main goals and concerns at that moment and any development that was surprising or went against expectations. The participant was also asked to describe any key pieces of information that was used to form the assessment or plan of action and where the information was obtained from. If there were any instances where the participant had made or had almost made an error or if another team member had had a different assessment of the situation, the participant was also asked to describe them and how he or she thought it happened.

4.4 Results

Out of the 12 interviews conducted, one interview was not used because the mannequin had malfunctioned during the scenario and affected the diagnosis process. Two other interviews were also dropped as there was insufficient elaboration of the responses from the interviewee limiting the usability of the data. The remaining nine interviews were transcribed and reviewed.

Scope of Analysis

To be able to manage the patient effectively, the students needed to seek diagnostic cues from the environment to generate hypotheses of what was happening to the patient before they formulated an action plan to evaluate their hypotheses and refine their working diagnosis. If they believed they had identified the correct diagnosis, they then had to formulate an action plan to treat the patient accordingly.

The analysis of the interview focused on the ability of the students to generate relevant hypotheses that would then guide their choice of action plan to manage the patient. Sub-optimal patient management actions would include the commission of a treatment plan that was irrelevant to the actual diagnosis of the patient, the unnecessary ordering of irrelevant investigative tests or an inability to form a proactive action plan, that is to be uncertain of the necessary action plan which leads to inaction or unnecessarily waiting for an event to occur that would give them more information.

These sub-optimal patient management actions were typically a result of a misperception or misinterpretation of the environmental cues that led to an inability to generate any working diagnosis or at least any awareness of the possible failures in a specific physiological process or system that led to the observable signs and symptoms (i.e. a partial diagnosis). It could also lead to generation of an incorrect diagnosis or partial diagnosis altogether.

Notable instances of these sub-optimal diagnostic performance from the different participants and the factors that may have led to them are discussed here (See Appendix I for relevant participant utterances).

Participant 01 (Scenario: Severe Asthma Attack)

Early in the scenario, the participant had perceived a combination of cues (falling blood pressure, increased heart rate, decreased central venous pressure and the fractured femur) and had wrongly recognised it as a pattern for hypovolemia (a state of decrease in blood volume) and diagnosed it as such. He then proceeded to treat the condition by administering fluids to the patient, which was unnecessary.

Verification of the vital signs from the video recording revealed that the blood pressure at that point was 108/62mmHg, after falling slightly from 116/62mmHg. This could still be considered as within the acceptable range for blood pressure. During hypovolemia, the expected drop in blood pressure should have been greater. The misinterpretation of this vital sign was perceived early in the encounter and had triggered the erroneous initial working diagnosis of hypovolemia. The other cues were then interpreted as supporting this diagnosis (confirmation bias). Even though he noticed that there were some cues that were not in line with the hypovolemia diagnosis, there was still a reluctance to shift away from it as an existing condition(anchoring bias).

Participant 03 (Scenario: Hypovolemia due to bleeding from fracture site)

The participant had perceived two critically relevant cues, low blood pressure and low central venous pressure. However, this did not trigger any hypothesis generation. Instead they only reacted by administering fluids with the aim of correcting the blood pressure and central venous pressure. Although the treatment was appropriate for the actual condition, it was not because they had correctly diagnosed hypovolemia. Participant 03 seemed to have followed an automatic rule-response to the perceived

decreased blood pressure. If he had recognized the pattern of cues as a possible hypovolemia, it may have led to better patient management actions, such as being more aggressive with the fluid resuscitation. Instead, they were not sure if they could give more fluids than they already had.

When the patient complained of giddiness, which is typical in hypovolemia, the team related this to the patient's history of diabetes (i.e. sub-optimal blood sugar control) that was told to them before the scenario (availability heuristic). Thus, they ordered a hypocount test to differentiate between hypoglycaemia (complication due to insufficient blood sugar) and diabetic ketoacidosis (complication due to excessive blood sugar) so that they could carry out the appropriate treatment plan but they were not able to proceed as they did not receive the results of the test. When they did not observe the expected response from the patient, they could not generate any other alternatives and decided to stick to their erroneous working diagnosis (anchoring bias).

Although the team had noted the low blood pressure and central venous pressure and treated it by administering fluids, there was no mention of how these cues could have factored into the working diagnosis even though they had high diagnosticity. They seem to have paid too much attention to the giddiness and the diabetes history and had ignored the more potent cues.

Participant 05 (Scenario: Anaphylaxis)

Early in this scenario, the participant had generated a hypothesis of pulmonary embolism (i.e. blockage of critical blood vessel in lungs due to a substance from somewhere else that has travelled through the bloodstream) based on the fractured leg. It was consistent with the breathlessness and chest pain.

However, the participant then noticed that it was not consistent with the fact that the patient was not as bad as he should be in the case of pulmonary embolism. Although it was right for him to doubt the pulmonary embolism, the subject was unable to generate an alternative hypothesis. He had overvalued the fractured femur as a supporting cue when it was actually irrelevant to the actual condition of anaphylaxis (i.e. an allergic reaction to a substance such as a specific medication which could lead to swelling of the throat). The participant was attempting to generate a diagnosis that could take into account all the cues perceived, even though only a subset of it was actually relevant, and was unable to do so.

Participant 07 (Scenario: Severe Asthma Attack)

The participant managed to rule out a heart-related cause early in the scenario and correctly identified a lung-related cause by detecting an abnormal lung sound during the chest auscultation. The subject was able to generate two different hypotheses, chronic obstructive pulmonary disorder (COPD), which may be similar to the actual condition and pneumothorax (i.e. an abnormal presence of air in the chest cavity that can lead to breathing difficulties)

However, the subject was not able to ascertain exactly what was wrong with the lungs and was not able to formulate a targeted search plan to evaluate their hypotheses. Their main goal instead was to correct the perceived desaturation by providing oxygen support. There was insufficient attention paid to formulating an action plan to look for specific cues to confirm or dispute the two hypotheses.

Later on in the scenario, the participant claimed to have detected unequal breath sounds which would support the diagnosis of pneumothorax. However, the cue was

misperceived as there was actually no unequal breath sounds in the lungs. Fortunately, being undecided, the team did not act on the suspected pneumothorax by performing a needle decompression. It would have been an inappropriate treatment that could cause even more harm to the patient.

Participant 08 (Scenario: Hypovolemia due to bleeding from fracture site)

The participant had correctly noticed the red-stained bandage at the fracture site but had not assigned enough weightage to this cue because it was misinterpreted as being normal. Nonetheless, the subject was still able to correctly diagnose that the patient was suffering from hypovolemia based only on the perceived vital signs and correctly initiated an aggressive fluid resuscitation action plan.

However, unknown to the participant, the patient's condition continued to deteriorate because he was losing blood from the wound site which the team did not address. When the participant perceived that the patient's condition was not improving despite their aggressive treatment plan, they started searching for other alternative explanations by ordering different investigative tests with the goal of ruling out acute myocardial infarction (by using a 12-lead ECG), hypoglycaemia (by using a hypocount) and other respiratory causes (by using a chest x-ray). When the participant reflected on the scenario, he admitted that too much attention was paid to finding other causes instead of figuring out why their treatment action plan was not working.

4.5 Discussion

From the nine participant interviews, five of them had revealed instances of diagnostic failures in their attempts to diagnose the underlying condition of their patient. While the specific reasons for the breakdown in diagnostic performance leading to the failures in the five cases were different from one another, there were some similarities between some of them.

Subject 03 and Subject 05 experienced problems in pattern recognition. Even though they had reported detecting the critical cues for their respective scenarios, they were unable to match them to a relevant diagnosis. It is unlikely that the reason for this is because they did not possess the declarative knowledge of the conditions because the conditions in the training scenarios were chosen based on them being already covered in their undergraduate training. It was probably that the perceived cues did not trigger the recall of this existing relevant knowledge during the hypothesis generation process but had triggered other pieces of knowledge that were more cognitively available.

Misperception and misinterpretation of cues occurred in the scenarios of Subject 01, Subject 07 and Subject 08. In these cases, certain cues were interpreted as supporting something else and thus affected how the hypothesis generation. These instances could be attributed to confirmation bias, where the subject tended to seek confirmatory evidence and certain cues were perceived incorrectly because they could easily support the belief that was being entertained at the time.

Another common theme that threads through the five identified scenarios was the misevaluation of the relevance of the perceived cues. Certain irrelevant cues were dwelled on as if they were relevant, which led to the generation of irrelevant hypotheses while others were undervalued, paid less attention to and were not used to

trigger the relevant hypotheses. Again, this may have been due to confirmation bias which caused the subjects to focus on cues that were relevant to the existing hypothesis. It could also have been due to the order effects of primacy and recency, where cues that were presented earliest and latest are more likely to be recalled and therefore processed.

Based on the sub-optimal performance of these participants, breakdowns in the diagnostic process could occur in the initial hypothesis generation and cascade into the improper evaluation of cues. The generated hypotheses provide the lens through which the cues are then considered. Hence when the initial hypotheses is not correct, cues became misperceived, misinterpreted and misevaluated, leading to a degradation in the overall diagnostic performance and treatment actions.

While experience in diagnostic decision making can be built up over time and sufficient practice, an appropriate technological intervention could improve the diagnostic performance and shorten the performance gap due to experience by mitigating these issues in hypothesis generation and evaluation. Automated diagnostic decision aids have the potential to complement the decision strategy of the novice physician and provide an alternative source for hypothesis generation. This would be critical during instances of the inability to generate a hypothesis or the generation of an irrelevant hypothesis because such an intervention could provide a timely alternative hypothesis that can then be evaluated with the appropriate investigative tests.

However, as observed in some of the participants during the exploratory study, an erroneous initial hypothesis can be detrimental to performance. An imperfect automation can unexpectedly fail and provide a wrong recommended diagnosis and still result in faulty evaluation of cues. Providing transparency features gives the

physician access to additional informational cues that can then be used to evaluate the validity of the recommendation as well as a basis for comparison for the physician's own diagnostic thought processes. Providing the key pieces of evidence used by the machine can help the novices overcome their tendency to tunnel their attention onto cues that may be irrelevant while providing them with additional cues that they may have overlooked. Providing the machine's certainty rating in its recommendation may reduce their tendency to focus too much on the aid's recommendation by highlighting that it may be incorrect. It increases their awareness that there may be alternative diagnoses to consider and could lead to a more careful assessment of the environmental cues.

However, the effects of automation transparency on trust and performance during human-automation collaborative decision making under emergency conditions needs to be better understood before the potential for automated decision aids to improve diagnostic performance can be maximized.

This exploratory study has provided the author with the opportunity to learn important domain knowledge in the field of medical diagnostic decision making as well as provided valuable insights into thought processes and potential pitfalls involved when forming and evaluating diagnoses. It has also provided the author with experience relating to working with human patient simulators and medical simulation scenarios which would be relevant and useful for the subsequent study.

[INTENTIONALLY LEFT BLANK]

CHAPTER 5: METHOD

5.1 Participants

To simulate a naturalistic medical diagnostic decision making task, 20 male and 20 female third-year undergraduate medical students (Age, $M = 22.05$, $SD = 0.78$), were recruited as participants for the experiment for their level of expertise. Medical students at this level of training were chosen as participants because they possessed the necessary medical knowledge and field experience required to act as physicians making basic diagnostic decisions. Although using experienced physicians might have provided a simulation that was closer to real-life diagnostic decision making, the recruitment of such participants was not feasible for the experiment because their relative availability was much less compared to that of medical students. Furthermore, previous studies have shown that advanced level students within a particular discipline can make suitable surrogates for professionals of the same discipline (Höst, Regnell & Wohlin, 2000; Mortensen, Fisher & Wines, 2012).

No pre-test was conducted to assess their medical knowledge as it was reasonable to assume that by the third year of training, they will have the prerequisite skills and knowledge to recognize the diagnoses that will be used for the experiment and process the accompanying signs and symptoms from the decision environment.

5.2 Independent Variables

The design of the experiment for the study is a 2 x 2 mixed factorial design. The independent variable of the study was automation transparency, manipulated by the availability of two different informational display elements.

Figure 5-1. Transparency configurations for the automated decision aid

The first was **Key Diagnostic Cues**, where the aid would display in bulleted text, the main cues that it had used to generate its suggestion. This was a between-subject factor with two levels which were when the cues were absent (**Cues-OFF**) or when they were present (**Cues-ON**). When this information was available, it will provide transparency of process, as described in the literature review.

The second was **Likelihood Ratings**, where the aid would display a number which represented the probability that it has calculated for its recommendation to be accurate. This was a within-subject factor with two levels which were when the ratings were absent (**Like-OFF**) or when they were present (**Like-ON**). When this information was available, it will provide transparency of performance, as described in the literature review.

By manipulating the presence of these two informational display elements, there were thus 4 possible transparency configurations for the decision aid, as shown in Figures 5-1(a) to (d). (See Appendix II for a more detailed view of the screens)

5.3 Dependent Variables

The measured variables of the study were how much the participant trusted the aid's recommendation during each encounter, whether they decided to rely or not rely on the aid's recommendation during the encounter, how confident the participant was with this decision, how much the above trust and confidence changed after having experienced an automation failure in a previous usage and after experiencing a return to correct performance and lastly, how much they trusted the system as a whole.

Trust in Aid's Recommendation was measured using a Likert-type questionnaire item "To what extent did you trust the recommendation made by the diagnostic aid during the scenario?" anchored at 1: "Did not trust at all" to 7: "Completely trusted".

Reliance Decision was indicated by whether the participant chooses to agree or disagree with the recommendation given by the automated decision aid for the participant's final diagnosis.

Diagnostic Confidence was measured using a Likert-type questionnaire item “How confident are you that the chosen diagnosis will be the correct diagnosis?” anchored 1: “Not at all confident” to 7: “Extremely Confident”.

Change in Trust and Confidence due to the effect of automation failure experience and the effect of the return to correct performance was measured by comparing the ratings for trust and confidence during the encounters immediately after the failure encounter and immediately after the encounter with the return to correct performance respectively against the ratings in the initial encounter with the aid.

Overall subjective trust in the system was measured using a 6-item questionnaire administered at the end of the session of usage with a given transparency configuration. The items were adapted from the questionnaire developed by Jian, Bisantz & Drury (2000).

5.4 Procedure

Medical Simulation Laboratory Setup

For the experiment, a medical simulation laboratory at the National University Hospital, Singapore, was set up to a simulated intensive care unit (ICU) equipped with a high-fidelity computerized human patient simulator, a Laerdal SimMan 3G. The patient simulator simulated a number of pre-programmed medical crises with abnormal patient parameters. It was able to simulate physiological features, such as breath sounds, chest movements, heartbeats and pulses, which could be observed by the participant. The patient simulator also showed the real-time vital signs such as heart

rate, blood pressure, blood oxygen saturation and temperature on a display monitor akin to a typical vital signs monitor found in regular ICUs.

The automated decision aid, presented on a different display monitor which was positioned next to the vital signs monitor, would provide a diagnostic suggestion to the participant during the patient crisis event. The decision aid was not an actual functioning decision aid but instead, the diagnostic suggestions and other informational content that it displayed had been pre-programmed for each diagnostic scenario. Although the decision aid was not real, participants were not informed of this fact but instead attempts were made to lead them to believe that it was a fully functional machine.

The administrator from behind a one-way mirror controlled the patient simulator and the automated decision aid while observing the actions of the participants. The participants' would provide their responses for the simulation scenario on a specially designed computer application on a laptop placed near the participant using a mouse.

Design of Medical Diagnosis Scenarios

A total of 10 medical diagnosis scenarios were designed by a subject-matter expert, who was an experienced educator in the field of medical simulation training, as shown in Table 5-A.

Each scenario was designed as a patient encounter during an emergency event in the ICU. During each scenario, the patient, who was suffering from an unknown medical condition, had become unconscious and slipped into a deteriorated state. Each scenario has a unique patient with its own patient identity, medical history, medical

investigation results and observable patient simulator vital signs data and physical cues. Most of these details were designed to altogether form the picture of the unknown diagnosis while some of the details were included as extraneous information which was meant to create some uncertainty in the scenario. The specific details of each scenario can be found in Appendix III.

Table 5-A. *List of medical diagnosis scenarios*

Sce.	Type	Actual Diagnosis	Decision Aid's Recommended Diagnosis
A	Non-Error	Haemorrhagic Shock	Haemorrhagic Shock
B	Non-Error	Congestive Cardiac Failure	Congestive Cardiac Failure
C	Non-Error	Acute Myocardial Infarction	Acute Myocardial Infarction
D	Non-Error	Exacerbation of COPD	Exacerbation of COPD
E	Non-Error	Hypoglycemic Attack	Hypoglycemic Attack
F	Non-Error	Diabetic Ketoacidosis	Diabetic Ketoacidosis
G	Error	Tension Pneumothorax	Cardiac Tamponade
H	Error	Hyperkalemia	Hypoglycemic Attack
I	Practice	Septic Shock	Septic Shock
J	Practice	Cardiovascular Accident	Cardio Vascular Accident

Out of the 10 simulation scenarios created by the subject-matter expert, 8 were non-failure scenarios, where the automated decision aid would generate the correct diagnostic recommendation. 2 of these non-failure scenarios were designated as practice scenarios, to be used by the participant to familiarize themselves with the experimental setup and the decision aid. The other 6 non-failure scenarios were to be used for the actual experimental trials.

The last 2 simulation scenarios were designed as failure scenarios, where the automated decision would commit an error and provide an erroneous suggestion. The

failure scenarios were designed such that the automation error was a plausible one for the operating environment where one would expect to find such an aid. For Scenario G (Tension Pneumothorax), the aid would make an a wrong recommendation of Cardiac Tamponade, which has similar signs and symptoms, because the aid had missed a critical cue that was the unequal air entry between the two lungs. For Scenario H, the aid missed the diagnosis of Hyperkalemia and instead provided the recommendation of Hypoglycemic Attack based on the signs and symptoms made available to it. The automation failed because unbeknownst to the aid and the physician, the actual cause of the deterioration was because the patient was administered with the wrong drug for his pain, which led to the symptom of further pain.

Experimental Procedures

The 40 participants were randomly assigned to the two equally-sized between-subject groups, the Cues-OFF group and the Cues-ON group, as shown in Table 5-B. The first group used automated decision aids without the display of Key Diagnostic Cues (Cues-OFF group) while the second group used decisions aids that displayed the Key Diagnostic Cues (Cues-ON group).

Table 5-B. Assignment of experimental conditions to participants

Transparency Configuration	Cues-OFF group	Cues-ON group
Cues-OFF / Like-OFF	●	
Cues-OFF / Like-ON	●	
Cues-ON / Like-OFF		●
Cues-ON / Like-ON		●

The participants within each group underwent two experimental sessions with 4 scenario trials per session. Each session would involve a different decision aid, one with the Likelihood Ratings absent (Like-OFF) and the other one with the Likelihood Ratings present (Like-ON). The order of presentation of the two sessions to the participants was counter-balanced within the two groups.

Figure 5-2. Outline of Experimental Procedures

Figure 5-2 shows an outline of the experimental procedures for each participant. The second trial of each session (Trials 2 and 6 in Figure 5-2) was assigned with the two pre-designed failure scenarios, where the decision aid provided the participant with an erroneous suggestion. The order of presentation of the two failure scenarios between Trial 2 and Trial 6 was also counter-balanced across all participants within the group.

The other six trials for both sessions were then assigned the non-failure scenarios, where the decision aid will perform correctly and provide the right suggested diagnosis. The order of presentation of the six non-error scenarios was determined randomly for each participant.

Briefing

During the briefing, written consent and demographic information was first obtained from the participant before the experiment commenced. The participant was then briefed on the procedures of the experiment. The participant was informed that there were two different types of automatic decision aid systems to be tested over two sessions of four scenario trials. In the first session, they will use one decision aid (first Likelihood Rating condition) and in the second set, they will use a completely different decision aid (second Likelihood Rating Condition). The participant was asked to imagine as much as possible that they were handling the scenario in real life and behave as they would if it was a real patient, even though the scenarios were only simulations.

The participant would be instructed to consider the suggestion from the decision aid and whatever additional information it provided but it was up to the participant to decide how to use the suggestion, as the final decision as to what the diagnosis was would be entirely up to the participant. As an incentive to perform to the best of their ability, an additional monetary reward was given if the participant made correct diagnoses for 5 or more out of 8 experimental trials. As vulnerability is an important element of trust, the incentive ensures that the participant have something at stake as they are forming their diagnostic decisions.

Training

After the briefing session, the participant reviewed a set of training notes for up to 20 minutes to familiarize with a total of 12 different possible diagnoses. (See Appendix IV) The 12 diagnoses used for the experiment and their corresponding signs and

symptoms were common medical conditions, which the participants were expected to know at their current level of training as medical students. The participant was informed that the diagnoses that would be used in the different simulation scenarios would be chosen randomly from the list of 12 diagnoses, and that it might be possible for diagnoses to be repeated in more than one scenario.

Familiarization Trial

After the training was completed, the participant went through an orientation of the simulation lab. The administrator explained the possible interactions with the patient simulator, which included the different types of observable cues that can be gathered from an examination of the patient's body.

At the beginning of both sessions, the administrator introduced the decision aid for the relevant experimental condition (Cues-OFF or Cues-ON, Like-OFF or Like-ON) to the participant. The different informational content available on the decision aid screen was explained and the participant was then told that although the decision aid might not 100% correct all the time, it could be considered as highly reliable.

This was followed by a practice simulation trial for the participant to familiarize with the patient simulator and the decision aid. After the practice trial, the participant was given an opportunity to ask for clarifications regarding the experimental setup.

Experimental Sessions

The participant then went through the two sessions of four consecutive scenario trials, one for each within-subject condition (Like-OFF and Like-ON). There was a five-minute break between the two sessions.

Before the start of each trial in a session, the participant was given a pre-simulation briefing. With the exception of the first trial of the session, the experimenter first briefed the correct diagnosis for the previous trial as well as the suggestion of the decision aid (“In the previous scenario, your diagnosis was ____, the aid’s recommendation was ____, the correct diagnosis was ____”). The participant then received written details of the patient information and available history for the current trial. The participant was then reminded of the main task and that the decision aid was reliable but not 100% correct all the time. When the participant was ready, the administrator then initiated the trial.

The participant in the scenario would be required to play the role of the physician attending to the patient. The participants’ main task was making the most probable diagnosis for the condition that the patient was experiencing from the information they could gather from the patient information, the real-time vital signs and a physical examination of the patient simulator. To emulate a realistic emergency scenario, the participant was required to formulate the diagnosis as quickly as possible while also carrying out some basic medical examination actions that would be typically carried out in an initial emergency encounter, while formulating the diagnosis.

As vulnerability is an important element of trust, an incentive to perform well is necessary so that the participant have something at stake as they are forming their diagnostic decisions. To provide this incentive, the subject was told that if the

participant could correctly identify at least 5 diagnoses out of the total of 8 trials over the two sessions, an additional cash incentive would be awarded in addition to the specified cash reimbursement for participating in the study.

Within a few seconds from the start of the patient encounter in the scenario, the automated decision aid would display a diagnostic suggestion to the participant which was determined beforehand. Similarly, the pre-determined key diagnostic cues and likelihood ratings were also displayed, depending on the required transparency configuration. The likelihood ratings assigned to each scenario was a randomly determined probability between 60 and 70 percent.

When the participant was ready to finalize the diagnostic decision, he or she would input the chosen diagnosis through the response terminal (see Figure 5-3, 5-4, 5-5) along with their confidence in their decision and trust in the recommendation made by the automated decision aid (7-point Likert-type scale). The simulation scenario terminated immediately when the participant made the decision. The participant was then given a short break after each trial.

Which of the following represents the most likely diagnosis for what the patient is experiencing?
Click on your selected diagnosis

Haemorrhagic Shock	Cardiac Tamponade
Septic Shock	Tension Pneumothorax
Congestive Cardiac Failure	Exacerbation of Chronic Obstructive Pulmonary Disease
Acute Myocardial Infarction	Hypoglycemic Attack
Pulmonary Embolism	Diabetic Ketoacidosis
Cerebrovascular Accident	Severe Hyperkalemia

Figure 5-3. Multiple choice screen for diagnosis on Participant Response Terminal

You had chosen the following as the most likely diagnosis:
 Haemorrhagic Shock

How confident are you that the above diagnosis will be the correct diagnosis?

Drag the slider along the scale to indicate your level of confidence, then click OK

Not at all Confident 1 2 3 4 5 6 7 Extremely Confident

Neutral

OK

Figure 5-4. Diagnostic confidence rating on Participant Response Terminal

To what extent did you trust the recommendation made by the diagnostic aid during this scenario?
Drag the slider along the scale to indicate your level of trust, then click OK

Did not trust at all Completely trusted

1 2 3 4 5 6 7

Neutral

OK

Figure 5-5. Trust in decision aid rating on Participant Response Terminal

Post Condition Questionnaire

At the end of each session, the participant was presented with the post-condition questionnaire, where they were asked to rate their level of agreement with a series of items regarding their overall trust of the system and the extent to which they understood how the aid arrived at its recommendations on a Likert scale (1=Strongly Disagree, 4=Neutral, 7=Strongly Agree) (See Appendix V for questionnaire items). Items in the questionnaire have been adapted from the items used by Jian et al (2000) and Cramer et al (2008) to measure trust and perceived understanding.

Debriefing

At the end of experiment, the experimenter briefed the overall performance of the participants and collected qualitative responses to the two decision aids systems.

CHAPTER 6: RESULTS

6.1 Overview of Results

Overview of proper reliance on aid over all trials

Table 6-A below shows the possible outcomes that may arise during each of the simulation trials.

Table 6-A. Possible outcomes for each simulation trial

		Decision Aid	
		<i>Automation Failure</i>	<i>No Automation Failure</i>
Subject	<i>Agree with Automation</i>	Over-Trust	Appropriate Reliance
	<i>Disagree with Automation</i>	Appropriate Non-Reliance	Distrust

During the first (Pre-Failure), third (Post-Failure 1) and fourth (Post-Failure 2) trials of the session, there were no automation failures. The aid would perform properly and provided the subject with the correct recommended diagnosis. During these non-failure trials, if the subject chose to agree with the aid, the reliance on the aid would be appropriate and the subject's diagnosis would be accurate. If the subject chose to disagree with the aid, it would be an occurrence of distrust and it would be misdiagnosis.

During the second (Failure) trial of the session, the automated decision aid would not perform properly and provided the subject with a wrong recommended diagnosis. If the subject chose to agree with the aid in spite of its error, it would be an occurrence of over-trust. The subject would have made a misdiagnosis. The proper action during the Failure trial required the subject to disagree with the aid and come up with his own

diagnosis. Even though the subject's own diagnosis may or may not be the correct diagnosis, the study is more interested in the proper reliance decision towards the automated diagnostic aid than the actual accuracy of the subject's diagnosis. Therefore, the analysis did not differentiate between disagreeing subjects who identified the correct diagnosis and disagreeing subjects who made the wrong diagnosis.

Figure 6-1 below shows the percentage of subjects out of 20 subjects within each combination of transparency features who made the appropriate reliance decision with the decision aid during the four trials of the session. A general trend that was observed was that a lower number of subjects did not make the correct reliance decision during the Failure trial. A more detailed analysis of the results, including the differences between the transparency configurations is presented in a later section.

Figure 6-1. Percentage of subjects who relied on the aid appropriately over each trial of the session for different combinations of transparency features

Overview of trust ratings over all trials

Figure 6-2 below illustrates how the mean of the subject's trust rating in the aid's recommendation changed across the four trials for the different configurations of the transparency features. A trend observed is the lower trust ratings during the Failure trial as subjects may have perceived that something was amiss between the aid's recommendation and their own interpretation of the scenario. As expected, trust immediately after the failure in the Post Failure 1 trial was generally observed to be lower than the Pre-Failure Trial. A more detailed analysis of the results of the trust and its evolution according to the transparency configurations is presented later in the chapter.

Figure 6-2. Trust in aid's recommendation over all trials

Overview of Diagnostic Confidence

Figure 6-3 below illustrates how the mean of the subjects' confidence ratings in their chosen diagnosis changed between the four trials for the different configurations of the transparency features for when the subjects made the correct reliance decision (i.e. Appropriate Reliance and Appropriate Non-Reliance). Similarly to trust, it is observed that confidence for most of the transparency configurations was lower in the Failure trial as subjects noticed a greater possibility of error but recovered with an upward trend over the two Post-Failure trials. A more detailed analysis of the diagnostic confidence over the different transparency configuration is presented later.

Figure 6-3. Mean Confidence in diagnosis over all trials, when appropriately relying or not relying on the decision aid

Figure 6-4 below shows how the mean confidence ratings changes for when the subjects made an incorrect reliance decision (i.e. Distrust and Over-Trust). A missing

point is noted in the Pre-Failure trial as there were no subjects who disagreed with the aid for the Cues-ON/Like-ON configuration. A more detailed analysis of the diagnostic confidence is presented later.

NOTE: There were no subjects who disagreed with the decision aid during the Pre-Failure Trial of the Cues-ON/Like-ON condition.

Figure 6-4. Mean Confidence in diagnosis over all trials, when distrusting or over-trusting the decision aid

6.2 Transparency and Performance during Non-Automation Failure

Reliance Decision during Non-Failure Trials

Table 6-B below shows the mean number of trials out of the 3 non-failure trials (Pre-Failure, Post-Failure 1 and Post-Failure 2) in the session that a subject relied on the aid appropriately (i.e. agreed with the aid).

Table 6-B. Mean number out of the three non-failure trials where there was appropriate reliance

	Like-OFF	Like-ON	Total
Cues-OFF	2.55 (85.1%) (SE 0.14, n=20)	2.60 (86.8%) (SE 0.11, n=20)	2.58 (86.0%) (SE 0.09, n=40)
Cues-ON	2.70 (90.1%) (SE 0.11, n=20)	2.55 (85.1%) (SE 0.14, n=20)	2.63 (87.6%) (SE 0.09, n=40)
Total	2.63 (87.6%) (SE 0.09, n=40))	2.58 (86.0%) (SE 0.09, n=40)	2.60 (86.8%) (SE 0.07, n=80)

2 (between-subject, Cues: OFF vs ON) x 2 (within-subject, Like: OFF vs ON) repeated measures ANOVA, with order of presentation of the within-subject condition as a blocking factor, was performed on the results. There were no significant main effects found on the mean percentage for both the presence of the Key Diagnostic Cues, $F(1,36) = 0.140$, $p = .710$, and the presence of the Likelihood Ratings, $F(1,36) = 0.184$, $p = .670$, nor was there a significant interaction between them, $F(1,36) = 0.760$, $p = .389$.

Further analysis on the effects of the presence of the transparency features was done for each individual non-failure trial. McNemar's chi-square test was performed on the within-subject factor, Presence of Likelihood Ratings, at the different levels of the between-subject factor, Presence of Key Diagnostic Cues, and at the different non-failure trials of the session. Then, for the between-subject factor, Presence of Key Diagnostic Cues, Pearson's chi-square test was used to evaluate if there was a significant change in the proportion who agreed with the aid when the Key Diagnostic Cues were absent or present, at different levels of within-subject factor, Presence of Likelihood Ratings, during the individual non-failure trials of the session. Tables 6-C

and 6-D show the contingency tables for the respective chi-square tests and their results.

The presence of the Likelihood Ratings did not significantly change the proportion of subjects who agreed with a correctly performing aid during most of the conditions. The only exception where the presence of the Likelihood Ratings changed the proportion significantly, $\chi^2(1) = 6, p = .014$, was when the Key Diagnostic Cues were present and during the third (Post-Failure 1) trial of the session, which was immediately after the subject had learnt that the decision aid had made an error in the previous trial. The proportion of subjects who agreed with the aid during this condition decreased from 95% ($n = 19$) when the Likelihood Ratings were absent to 65% ($n = 13$) when they were present.

The proportion of subjects who agreed with an aid that was performing properly did not change significantly with the presence of the Key Diagnostic Cues in all the trials at the 5% level of significance. However, the change in proportion was marginally significant when the Likelihood Ratings were present during the first (Pre-Failure) trial, $\chi^2(1) = 3.24, p = .072$, and the third (Post-Failure 1) trial, $\chi^2(1) = 3.58, p = .058$, of the session. When the Likelihood Ratings were presented, the presence of the Key Diagnostic Cues increased the proportion of subjects who agreed from 85% ($n = 17$) to 100% ($n = 20$) in the Pre-Failure trial. On the other hand, immediately after learning of the automation error during the second trial, the proportion of subjects who agreed with the aid when it returned to correct performance in the Post Error 1 trial reduced from 90% ($n = 18$) when only the Likelihood Ratings were present to 65% ($n = 13$) when both Likelihood Ratings and Key Diagnostic Cues were present.

Table 6-C. Number of subjects who appropriately relied on or distrusted the aid when using one aid without the Likelihood Ratings and another with the Likelihood Ratings for each non-failure trial

PRE-FAILURE TRIALS						
Cues-OFF				Cues-ON		
Like-OFF	Like-ON			Like-OFF	Like-ON	
	Appropriate				Appropriate	
	Reliance				Reliance	
	Distrust				Distrust	
	Appropriate Reliance	15	3		Appropriate Reliance	18
Distrust	2	0	Distrust	2	0	
$\chi^2(1) = 0.2, p = .654$ (non-significant)				$\chi^2(1) = 2, p = .157$ (non-significant)		

POST-FAILURE 1 TRIALS							
Cues-OFF				Cues-ON			
Like-OFF	Like-ON			Like-OFF	Like-ON		
	Appropriate				Appropriate		
	Reliance				Reliance		
	Distrust				Distrust		
	Appropriate Reliance	16	2		Appropriate Reliance	13	6
	Distrust	2	0		Distrust	0	1
$\chi^2(1) = 0, p = 1.000$ (non-significant)				$\chi^2(1) = 6, p = .014$ (significant)			

POST-FAILURE 2 TRIALS							
Cues-OFF				Cues-ON			
		Like-ON				Like-ON	
		Appropriate				Appropriate	
		Reliance	Distrust			Reliance	Distrust
Like-OFF	Appropriate	13	2	Like-OFF	Appropriate	15	2
	Reliance				Reliance		
Like-OFF	Distrust	4	1	Like-OFF	Distrust	3	0
$\chi^2(1) = 0.667, p = .414$ (non-significant)				$\chi^2(1) = 0.2, p = .654$ (non-significant)			

Table 6-D. Number of subjects who appropriately relied on or distrusted the aid when the Key Diagnostic Cues were absent or present for each Non-failure trial

PRE-FAILURE TRIALS					
Like-OFF			Like-ON		
	<i>Appropriate Reliance</i>	<i>Distrust</i>		<i>Appropriate Reliance</i>	<i>Distrust</i>
Cues-OFF	18	2	Cues-OFF	17	3
Cues-ON	18	2	Cues-ON	20	0
$\chi^2(1) = 0, p = 1.000$ (non-significant)			$\chi^2(1) = 3.24, p = .072$ (marginally significant)		

POST-FAILURE 1 TRIALS					
Like-OFF			Like-ON		
	<i>Appropriate Reliance</i>	<i>Distrust</i>		<i>Appropriate Reliance</i>	<i>Distrust</i>
Cues-OFF	18	2	Cues-OFF	18	2
Cues-ON	19	1	Cues-ON	13	7
$\chi^2(1) = 0.36, p = .548$ (non-significant)			$\chi^2(1) = 3.58, p = .058$ (marginally significant)		

POST-FAILURE 2 TRIALS					
Like-OFF			Like-ON		
	<i>Appropriate Reliance</i>	<i>Distrust</i>		<i>Appropriate Reliance</i>	<i>Distrust</i>
Cues-OFF	15	5	Cues-OFF	17	3
Cues-ON	17	3	Cues-ON	18	2
$\chi^2(1) = 0.625, p = .429$ (non-significant)			$\chi^2(1) = 0.229, p = .633$ (non-significant)		

Trust in Aid's Recommendation during Non-Failure Trials

Table 6-E below show the mean and standard error of the subjects' mean trust ratings towards the aid's recommendations over the three non-failure trials when the aid performed properly.

Table 6-E. Mean of subjects' mean Trust in Recommendations from decision aid across all trials that the aid was correct
(1=Did not trust at all, 4=Neutral, 7=Completely trusted)

	Like-OFF	Like-ON	Total
Cues-OFF	5.350 (SE 0.180, n=20)	5.383 (SE 0.147, n=20)	5.367 (SE 0.137, n=40)
Cues-ON	5.432 (SE 0.180, n=20)	5.140 (SE 0.147, n=20)	5.286 (SE 0.137, n=40)
Total	5.391 (SE 0.127, n=40)	5.262 (SE 0.104, n=40)	5.326 (SE 0.097, n=80)

To test Hypothesis H1A, a 2 (between-subject, Cues: OFF vs ON) x 2 (within-subject, Like: OFF vs ON) repeated measures ANOVA, with order of presentation of the Like condition as a blocking factor, was performed on the mean trust ratings.

Shapiro-Wilk's Test was used to test assumption of normality for each cell. For Cues-OFF, p -values for Like-OFF and Like-ON were .668 and .117 respectively. For Cues-ON, p -values for Like-OFF and Like-ON were .570 and .780 respectively. The data met the assumption of normality.

Levene's Test was used to test assumption of homogeneity of variance. The p -values for Like-OFF and Like-ON were .486 and .738 respectively. The data met the assumption of homogeneity of variance.

There were no significant main effects found on the mean trust ratings for both the presence of the Key Diagnostic Cues, $F(1,36) = 0.173$, $p = .680$, and the presence of the Likelihood Ratings, $F(1,36) = 1.034$, $p = .316$, nor was there a significant interaction between them, $F(1,36) = 1.626$, $p = .210$.

There was no sufficient statistical evidence to support Hypothesis H1A that the presence of the transparency features on the automated diagnostic aid would increase the trust in the aid's recommendation when there is no automation failure.

Figure 6-5 shows how trust in the recommendation relates to the reliance decision during the non-failure trials for each transparency configuration. As expected, appropriate reliance corresponded with a higher trust rating ($\bar{X} = 5.60$, $SE = 0.07$, across all subjects in all conditions) compared to that of distrust ($\bar{X} = 3.41$, $SE = 0.16$), $t(38.8) = 12.3$, $p < .001$.

Figure 6-5. Trust in aid's recommendation when appropriately relying or distrusting a decision aid when there was no automation failure

Diagnostic Confidence during Non-Failure Trials

The reliance decision was found to have a significant effect on the subjects' confidence in their chosen diagnosis, $t(36.1) = 3.55$, $p = .001$. Subjects relying appropriately on the aid had a higher confidence rating ($\bar{X} = 5.86$, $SE = 0.06$, across all subjects in all conditions) than when they were distrusting it ($\bar{X} = 5.21$, $SE = 0.17$). Figure 6-6 below illustrates the difference in the mean confidence rating across all the non-failure trials for the different transparency configurations.

Furthermore, during appropriate reliance, trust in the aid's recommendation significantly predicted confidence ratings, $\beta = .571$, $t(206) = 10.07$, $p < .001$, and also explained a significant proportion of variance in the confidence ratings, $R^2 = .33$, $F(1, 206) = 101.30$, $p < .001$.

Figure 6-6. Diagnostic Confidence when appropriately relying or distrusting a decision aid when there was no automation failure

Table 6-F and Table 6-G show the values for the mean diagnostic confidence ratings in their chosen diagnosis for when they appropriately relied on and when they distrusted the decision aid respectively.

Table 6-F. Mean of subjects' mean Diagnostic Confidence during trials when appropriately relying on decision aid when there was no automation failure

(1=Not confident at all, 4=Neutral, 7=Extremely Confident)

	Like-OFF	Like-ON	Total
Cues-OFF	5.989 (SE 0.120, n=20)	5.801 (SE 0.116, n=20)	5.895 (SE 0.095, n=40)
Cues-ON	5.902 (SE 0.120, n=20)	5.749 (SE 0.116, n=20)	5.825 (SE 0.095, n=40)
Total	5.945 (SE 0.085, n=40)	5.775 (SE 0.082, n=40)	5.860 (SE 0.067, n=80)

Table 6-G. Mean Diagnostic Confidence when distrusting the decision aid when there was no automation failure

(1=Not confident at all, 4=Neutral, 7=Extremely Confident)

	Like-OFF	Like-ON	Total
Cues-OFF	5.644 (SE 0.332, n=8)	5.338 (SE 0.332, n=8)	5.491 (SE 0.235, n=16)
Cues-ON	4.917 (SE 0.383, n=6)	4.875 (SE 0.332, n=8)	4.896 (SE 0.253, n=14)
Total	5.332 (SE 0.071, n=14)	5.106 (SE 0.071, n=16)	5.212 (SE 0.173, n=30)

To test Hypothesis H1B, a 2 (between-subject, Cues: OFF vs ON) x 2 (within-subject, Like: OFF vs ON) repeated measures ANOVA, with order of presentation of the Like condition as a blocking factor, was performed on the subjects' mean diagnostic confidence ratings during the trials that they appropriately relied on the decision aid.

Shapiro-Wilk's Test was used to test assumption of normality for each cell. For Cues-OFF, p -values for Like-OFF and Like-ON were .114 and .322 respectively. For Cues-ON, p -values for Like-OFF and Like-ON were .011 and .090 respectively. The data met the assumption of normality, except for the Cues-ON/Like-OFF conditions where the p -value is significant. However, as ANOVA is robust to deviations from normal, ANOVA was continued.

Levene's Test was used to test assumption of homogeneity of variance. The p -values for Like-OFF and Like-ON were .553 and .273 respectively. The data met the assumption of homogeneity of variance.

The presence of the Key Diagnostic Cues did not have a significant main effect on the mean confidence ratings, $F(1,36) = 0.269$, $p = .607$. However, the presence of the Likelihood Ratings had a marginally significant main effect on the mean confidence ratings, $F(1,36) = 2.907$, $p = .097$ with the subjects being less confident in their diagnosis when the Likelihood Ratings were present. There was no significant interaction effects, $F(1,36) = 0.032$, $p = .859$, between the two factors.

Then to test Hypothesis H1C by examining the trials where there was distrust in the aid's recommendation, a 2 (Cues: OFF vs ON) x 2 (Like: OFF vs ON) ANOVA was performed on the subjects' mean confidence ratings during these trials. Repeated measures ANOVA was not used as there were too few subjects who had incidences of distrust during both of the within-subject conditions.

Shapiro-Wilk's Test was used to test assumption of normality for each cell. For Cues-OFF, p -values for Like-OFF and Like-ON were .159 and .105 respectively. For Cues-ON, p -values for Like-OFF and Like-ON were .679 and .010 respectively. The data met the assumption of normality, except for the Cues-ON/Like-ON condition where

the p -value is significant. However, as ANOVA is robust to deviations from normal, ANOVA was continued.

Levene's Test was used to test assumption of homogeneity of variance. The p -values for Like-OFF and Like-ON were .674 and .052 respectively. The data met the assumption of homogeneity of variance.

It was found that the presence of the Key Diagnostic Cues had a marginally significant main effect on the mean confidence, $F(1,26) = 2.969$, $p = .097$. The subjects who had the Key Diagnostic Cues present on their decision aids were less confident in their own chosen diagnosis, when disagreeing with the decision aid's recommendation which was actually correct. On the other hand, the presence of the Likelihood Ratings had no significant main effect on the mean confidence, $F(1,26) = 0.254$, $p = .619$, and neither was there any interaction effects, $F(1,26) = 0.147$, $p = .705$.

Therefore, there was no sufficient evidence to support Hypothesis H1B that the presence of the transparency features would increase confidence when appropriately relying on a decision aid that was performing correctly. However, there was marginal statistical evidence that the presence of the Likelihood Ratings might instead decrease diagnostic confidence during appropriate reliance. Hypothesis H1C was marginally supported as there was some statistical evidence that the presence of the Key Diagnostic Cues might decrease confidence when distrusting a decision aid that was correct.

6.3 Transparency and Performance during Automation Failure

Reliance Decision during Failure Trials

The automated diagnostic decision aid failed only during the second trial in the session. To see if the presence of the transparency features had changed the proportion of subjects who appropriately did not rely on or over-trusted the error-committing aid, McNemar's chi-square test was performed on the within-subject factor, Presence of Likelihood Ratings, at the different levels of the between-subject factor, Presence of Key Diagnostic Cues. And for the between-subject factor, Presence of Key Diagnostic Cues, Pearson's chi-square test was used to see if there was a significant change in the proportion who agreed with the aid when the Key Diagnostic Cues were absent or present, at different levels of within-subject factor, Presence of Likelihood Ratings. Tables 6-H and Table 6-I show the contingency tables for the respective chi-square tests and their results.

Table 6-H. Number of subjects who appropriately did not rely on or over-trusted the aid when using one aid without the Likelihood Ratings and another with the Likelihood Ratings for the Failure trial

FAILURE TRIALS						
Cues-OFF				Cues-ON		
Like-OFF	Like-ON			Like-ON		
	Appropriate Non-Reliance		Over-Trust	Appropriate Non-Reliance		Over-Trust
	Appropriate Non-Reliance	6	2	Appropriate Non-Reliance	3	1
	Over-Trust	2	10	Over-Trust	5	11
$\chi^2(1) = 0, p = 1.000$ (non-significant)				$\chi^2(1) = 2.67, p = .102$ (non-significant)		

Table 6-I. Number of subjects who appropriately did not rely on or over-trusted the aid when the Key Diagnostic Cues were absent or present for the Failure trial.

FAILURE TRIALS					
Like-OFF			Like-ON		
	<i>Appropriate Non-Reliance</i>	<i>Over-Trust</i>		<i>Appropriate Non-Reliance</i>	<i>Over-Trust</i>
Cues-OFF	8	12	Cues-OFF	8	12
Cues-ON	4	16	Cues-ON	8	12
$\chi^2(1) = 1.905, p = .168$ (non-significant)			$\chi^2(1) = 0, p = 1.00$ (non-significant)		

The presence of the within-subject factor, Likelihood Ratings, was found to not significantly change the proportion of subjects who appropriately did not rely on an incorrectly performing aid for both conditions where the Key Diagnostic Cues were absent and present. The presence of the of the between subject factor, Key Diagnostic Cues, also did not cause a significant change in the proportion of appropriate non-reliance for both conditions where the Likelihood Ratings were absent and present.

Trust in Aid's Recommendation during Failure Trial

Table 6-J show the means and standard errors of the subjects' trust ratings for the different transparency configurations during the failure trials when the aid performed incorrectly.

Table 6-J. Mean of subjects' Trust in Recommendations from decision aid during the trial that the aid was incorrect

(1=Did not trust at all, 4=Neutral, 7=Completely trusted)

	Like-OFF	Like-ON	Total
Cues-OFF	4.675 (SE 0.293, n=20)	4.540 (SE 0.306, n=20)	4.608 (SE 0.254, n=40)
Cues-ON	5.065 (SE 0.293, n=20)	4.690 (SE 0.306, n=20)	4.878 (SE 0.254, n=40)
Total	4.870 (SE 0.207, n=40)	4.615 (SE 0.216, n=40)	4.743 (SE 0.179, n=80)

To test Hypothesis H2A, a 2 (between-subject, Cues: OFF vs ON) x 2 (within-subject, Like: OFF vs ON) repeated measures ANOVA, with order of presentation of the Like condition as a blocking factor, was performed on the means.

Shapiro-Wilk's Test was used to test assumption of normality for each cell. For Cues-OFF, p -values for Like-OFF and Like-ON were .053 and .505 respectively. For Cues-ON, p -values for Like-OFF and Like-ON were .016 and .465 respectively. The data met the assumption of normality, except for the Cues-ON/Like-OFF condition where the p -value is significant. However, as ANOVA is robust to deviations from normal, ANOVA was continued.

Levene's Test was used to test assumption of homogeneity of variance. The p -values for Like-OFF and Like-ON were .718 and .115 respectively. The data met the assumption of homogeneity of variance.

Similarly to the non-failure trials, there were also no significant main effects found on the mean trust ratings for both the presence of the Key Diagnostic Cues, $F(1,36) = 0.566$, $p = .457$, and the presence of the Likelihood Ratings, $F(1,36) = 1.286$, $p = .264$, nor was there a significant interaction between them, $F(1,36) = 0.285$, $p = .597$.

Hence, there was no significant statistical evidence to support H2A that the presence of the Key Diagnostic Cues or the Likelihood ratings would decrease trust in an automated decision aid that was performing incorrectly.

The reliance decision on the aid's recommendation was found to correspond with differences in the trust in the aid's recommendation, as it did during the non-failure trials. Trust was significantly higher during over-trust ($\bar{X} = 5.48$, $SE = 0.11$, across all subjects in all conditions) compared to during appropriate non-reliance ($\bar{X} = 3.37$, $SE = 0.17$), $t(48.4) = 10.3$, $p < .001$. Figure 6-7 illustrates the difference in trust between over-trust and appropriate non-reliance.

Figure 6-7. Trust in aid's recommendation when over-trusting or appropriately not relying on the decision aid when there was automation failure

Diagnostic Confidence during Failure Trial

Unlike the non-failure trials, the reliance decision did not have a significant effect on the diagnostic confidence during the failure trials. Figure 5-8 shows the diagnostic confidence for the different transparency configurations during over-trust and during appropriate non-reliance. There was no significant difference between the confidence during over-trust ($\bar{X} = 5.39$, $SE = 0.13$, across all subjects in all conditions) and during appropriate non-reliance ($\bar{X} = 5.60$, $SE = 0.13$), $t(69.5) = 1.16$, $p = .251$. Subjects who agreed with the erroneous recommendation were equally confident in their diagnosis as those who disagreed with it and produced their own diagnosis.

Figure 6-8. Diagnostic Confidence Ratings when over-trusting or appropriately not relying on the decision aid when there was automation failure

On the other hand, trust in the aid's recommendation still significantly predicted confidence ratings during over-trust, $\beta = .798$, $t(50) = 7.00$, $p < .001$, as it did during

appropriate reliance. And it also explained a significant proportion of variance in the confidence ratings, $R^2 = .495$, $F(1, 50) = 49.08$, $p < .001$.

Table 6-K and Table 6-L show the mean and standard error for the subjective confidence ratings of the subjects in their chosen diagnosis under different configurations for the diagnostic aid's transparency features during appropriate non-reliance and over-trust respectively.

Table 6-K. Mean Diagnostic Confidence when disagreeing with decision aid when there is automation failure (Appropriate Non-Reliance)

(1=Not confident at all, 4=Neutral, 7=Extremely Confident)

	Like-OFF	Like-ON	Total
Cues-OFF	5.738 (SE 0.248, n=8)	5.563 (SE 0.248, n=8)	5.650 (SE 0.175, n=16)
Cues-ON	5.500 (SE 0.351, n=4)	5.550 (SE 0.248, n=8)	5.525 (SE 0.215, n=12)
Total	5.619 (SE 0.215, n=12)	5.556 (SE 0.175, n=16)	5.588 (SE 0.139, n=28)

Table 6-L. Mean Diagnostic Confidence when agreeing with decision aid when there is automation failure (Over-Trust)

(1=Not confident at all, 4=Neutral, 7=Extremely Confident)

	Like-OFF	Like-ON	Total
Cues-OFF	5.517 (SE 0.260, n=12)	5.433 (SE 0.260, n=12)	5.475 (SE 0.184, n=24)
Cues-ON	5.475 (SE 0.225, n=16)	5.133 (SE 0.260, n=12)	5.304 (SE 0.172, n=28)
Total	5.496 (SE 0.172, n=28)	5.283 (SE 0.184, n=24)	5.390 (SE 0.126, n=52)

To test Hypothesis H2B and Hypothesis H2C, a 2 (Cues: OFF vs ON) x 2 (Like: OFF vs ON) ANOVA was performed on the subjects' mean diagnostic confidence ratings for both cases of over-trust and appropriate non-reliance. The Presence of Likelihood Ratings was not used as within-subject factor in a repeated measures ANOVA like it was in the other analyses because there were too few participants with the same reliance decision in the two within-subject conditions for both situations of over-trust and appropriate non-reliance.

Shapiro-Wilk's Test was used to test assumption of normality. For appropriate non-reliance, Cues-OFF and Cues-ON, p -values were .102 and .199 respectively and for Like-OFF and Like-ON, p -values were .121 and .306 respectively. For over-trust, Cues-OFF and Cues-ON, p -values were .176 and .687 respectively and for Like-OFF and Like-ON, p -values were .213 and .186 respectively. The data met the assumption of normality.

Levene's Test was used to test assumption of homogeneity of variance. For appropriate non-reliance, the p -values for the Cues conditions and the Like conditions were .830 and .861 respectively. For over-trust, the p -values for the Cues conditions and the Like conditions were .561 and .137 respectively. The data met the assumption of homogeneity of variance.

During appropriate non-reliance, both the presence of Key Diagnostic Cues, $F(1,24) = 0.203$, $p = .656$, and the presence of Likelihood Ratings, $F(1,24) = 0.051$, $p = .823$ was found to not have any significant main effects on diagnostic confidence. There were also no significant interaction effects, $F(1,24) = 0.165$, $p = .688$.

Similarly, during over-trust, there no significant main effects found for both the presence of the Key Diagnostic Cues, $F(1,48) = 0.461$, $p = .500$, and the presence of

the Likelihood Ratings, $F(1,48) = 0.713$, $p = .403$. There were also no significant interaction effects between the two factors, $F(1,48) = 0.264$, $p = .610$.

Hence, there was insufficient statistical evidence to support H2B that the transparency configuration would increase confidence when appropriately not relying on a decision aid that is performing incorrectly. There was also not enough statistical evidence to confirm Hypothesis H2C that the transparency configuration will decrease confidence when over-trusting an aid that is performing incorrectly.

6.4 Transparency and the Experience of Automation Failure

Change in Trust from before to after Failure Trial

The aid performs incorrectly in the second trial of the session and the subject would learn of this automation error before starting the third trial. To study if the presence of the transparency features would cause a difference in the effects of learning about this error on the trust in a decision aid that was performing properly, the trust ratings in the Pre-Failure trial (first trial) were compared to those in the First Post-Failure trial (third trial) for each combination of the transparency features.

The experience of the automation failure during the second trial of the session was found to have a marginally significant main effect on the trust ratings, $t(79) = 1.86$, $p = .067$. The subjects across all conditions tended to have lower trust in the aid's recommendation in the First Post-Failure Trial ($\bar{X} = 5.15$, $SE = 0.13$) compared to the Pre-Failure Trial ($\bar{X} = 5.43$, $SE = 0.10$).

Figure 6-9 shows the how the trust ratings varied from the Pre-Failure Trial to the First Post Failure trial for each transparency configuration. Table 6-M below show the mean and standard error of the change in the subjects' trust ratings from the Pre-Failure trial to the First Post-Failure trial.

Figure 6-9. Trust Ratings from Pre-Failure trial to First Post-Failure Trial

Table 6-M. Mean change in trust rating from Pre-Failure trial to First Post-Failure trial

	Like-OFF	Like-ON	Total
Cues-OFF	-0.160 (SE 0.245, n=20)	+0.265 (SE 0.318, n=20)	+0.053 (SE 0.202, n=40)
Cues-ON	-0.170 (SE 0.245, n=20)	-1.035 (SE 0.318, n=20)	-0.603 (SE 0.202, n=40)
Total	-0.165 (SE 0.173, n=40)	-0.385 (SE 0.225, n=40)	-0.275 (SE 0.143, n=80)

To test Hypothesis H3A, a 2 (between-subject, Cues: OFF vs ON) x 2 (within-subject, Like: OFF vs ON) repeated measures ANOVA, with order of presentation of the Like condition as a blocking factor, was performed on the mean changes.

Shapiro-Wilk's Test was used to test assumption of normality for each cell. For Cues-OFF, p -values for Like-OFF and Like-ON were .537 and .743 respectively. For Cues-ON, p -values for Like-OFF and Like-ON were .884 and .417 respectively. The data met the assumption of normality.

Levene's Test was used to test assumption of homogeneity of variance. The p -values for Like-OFF and Like-ON were .647 and .963 respectively. The data met the assumption of homogeneity of variance.

It was found that the presence of the Key Diagnostic Cues had a significant effect on the change in trust, $F(1,36) = 5.278$, $p = .028$. With the Key Diagnostic Cues present, subjects experienced a larger drop in trust in the aid after learning of the error in the previous trial compared to when the Cues were absent. There was no significant effect of the presence of the Likelihood Ratings on the change in trust, $F(1,36) = 0.606$, $p = .441$.

There was also a significant interaction effect, $F(1,36) = 5.208$, $p = .028$, between the presence of the Key Diagnostic Cues and the presence of the Likelihood Ratings. Analysing the simple effects of the presence of Key Diagnostic Cues using independent t -tests, the factor did not result in a significant difference in the change in trust when the Likelihood Ratings were absent, $t(38) = 0.029$, $p = .977$. However, when the Likelihood Ratings were present, there was a larger drop in the subjects' trust in the recommendation when the Key Diagnostic Cues were provided, $t(38) = 2.903$, $p = .006$. Paired t -tests were then used to study the simple effects of the within-subject

factor, presence of Likelihood Ratings, for both the presence and absence of the Key Diagnostic Cues. There was no significant simple effect when found in the absence of the Key Diagnostic Cues, $t(19) = 0.890$, $p = .384$, but in the presence of the Key Diagnostic Cues, the presence of the Likelihood Ratings caused a significantly greater drop, $t(19) = 2.835$, $p = .011$, in the trust ratings than when the Likelihood Ratings were absent.

Hence, although there was no evidence to support H3A that the transparency configuration will reduce the trust drop due to the experience of automation failure, there was a still significant effect of transparency which had caused the trust drop to increase instead.

Change in Diagnostic Confidence during Appropriate Reliance from before to after Failure Trial

Before the start of the third trial of the session, the subject would learn that the diagnostic aid had performed incorrectly during the second trial. While the aid returns to correct performance in the third trial, the experience of the error in the second trial was expected to affect the confidence of the subject even when he or she appropriately relied on the aid during the third trial. Subjects who appropriately relied on the decision aid during the first (Pre-Failure) trial and the third (Post-Failure 1) trial were selected for the analysis.

Figure 6-10 shows the mean diagnostic confidence ratings of these subjects for both the first and third trials of the session for each combination of transparency features of the decision aid. Table 6-N shows the mean change in the subjects' confidence rating for each combination of transparency features.

Figure 6-10. Diagnostic Confidence Ratings when agreeing with decision aid that was correct before and immediately after error

Table 6-N. Mean Change in Diagnostic Confidence during appropriate reliance from before to after error experience

	Like-OFF	Like-ON	Total
Cues-OFF	+0.050 (SE 0.210, n=12)	+0.383 (SE 0.294, n=12)	+0.217 (SE 0.181, n=24)
Cues-ON	-0.100 (SE 0.210, n=12)	-0.542 (SE 0.294, n=12)	-0.321 (SE 0.181, n=24)
Total	-0.025 (SE 0.148, n=24)	-0.079 (SE 0.208, n=24)	-0.052 (SE 0.128, n=48)

To test Hypothesis H3B, a 2 (between-subject, Cues: OFF vs ON) x 2 (within-subject, Like: OFF vs ON) repeated measures ANOVA was conducted on the results.

Shapiro-Wilk's Test was used to test assumption of normality for each cell. For Cues-OFF, p -values for Like-OFF and Like-ON were .626 and .590 respectively. For Cues-ON, p -values for Like-OFF and Like-ON were .110 and .740 respectively. The data met the assumption of normality.

Levene's Test was used to test assumption of homogeneity of variance. The p -values for Like-OFF and Like-ON were .957 and .415 respectively. The data met the assumption of homogeneity of variance.

The presence of the Key Diagnostic cues had a significant main effect on the change in the diagnostic confidence, $F(1,22) = 4.412$, $p = .047$. The confidence during appropriate reliance showed a greater drop when the Key Diagnostic Cues were present. However the presence of the Likelihood Ratings was found to not have any significant effect on the change, $F(1,22) = 0.045$, $p = .834$. There was also no significant interaction effects, $F(1,22) = 2.306$, $p = .143$.

Hence, there was not enough statistical evidence to support Hypothesis H3B that the transparency features will reduce the drop in diagnostic confidence. However, there was still a significant effect of the presence of the Key Diagnostic Cues on confidence ratings where it had instead intensified the drop in diagnostic confidence.

6.5 Transparency and the Experience of Return to Correct

Automation Performance

Trust Recovery after experiencing return to correct performance

The aid returns to correct performance in the third trial of the session and the subject would learn that it did not make any errors in the third trial before starting the fourth trial. The trust in the aid's recommendation was expected to drop in the third trial and the learning of the return to correct performance was expected to have a recovery effect on trust. The trust in the fourth trial was compared to the initial trust in the first trial to see how well the trust would return or exceed the initial trust under the different combinations of transparency features.

Figure 6-11 illustrates how the trust in the aid's recommendation had changed from the pre-failure trial to the second post-failure trial for the different transparency configurations.

Table 6-O shows the mean and standard error of the change in the subjects' trust ratings from the Pre-Failure trial to the Second Post-Failure trial.

Figure 6-11. Trust Ratings from Pre-Failure trial to Second Post-Failure Trial

Table 6-O. Mean change in trust rating from Pre-Failure trial to Second Post-Failure trial

	Like-OFF	Like-ON	Total
Cues-OFF	-0.335 (SE 0.313, n=20)	+0.375 (SE 0.288, n=20)	+0.020 (SE 0.181, n=40)
Cues-ON	-0.335 (SE 0.313, n=20)	+0.195 (SE 0.288, n=20)	-0.070 (SE 0.181, n=40)
Total	-0.335 (SE 0.221, n=40)	+0.285 (SE 0.204, n=40)	-0.025 (SE 0.128, n=80)

To test Hypothesis H4, a 2 (between-subject, Cues: OFF vs ON) x 2 (within-subject, Like: OFF vs ON) repeated measures ANOVA, with order of presentation of the Like condition as a blocking factor, was performed on the mean changes.

Shapiro-Wilk's Test was used to test assumption of normality for each cell. For Cues-OFF, p -values for Like-OFF and Like-ON were .471 and .159 respectively. For Cues-ON, p -values for Like-OFF and Like-ON were .107 and .121 respectively. The data met the assumption of normality.

Levene's Test was used to test assumption of homogeneity of variance. The p -values for Like-OFF and Like-ON were .465 and .178 respectively. The data met the assumption of homogeneity of variance.

The difference in the change in the trust ratings was found to be marginally significant between the aid with the Likelihood Ratings and the aid without the Likelihood Ratings, $F(1,36) = 3.337$, $p = .076$. The change was more positive with the Likelihood Ratings present but was negative when it was absent. There was no significant main effect found for the presence of the Key Diagnostic Cues, $F(1,36) = 0.123$, $p = .728$, and there was also no significant interaction effects between the presence of the two transparency features, $F(1,36) = 0.070$, $p = .792$.

Hence, there was marginal statistical evidence to support Hypothesis H4 where the presence of the Likelihood Ratings improved the recovery of trust in the aid's recommendation after experiencing a return to non-failure performance.

6.6 Effect of Scenario on Trust and Diagnostic Confidence

The non-failure scenarios that were designed for the experimental trials were randomly assigned to the six non-failure scenarios underwent by each participant over the two sessions. ANOVA was used to see if the scenario had an effect on the trust in the aid's recommendation and the diagnostic confidence. Figure 6-12 and Figure 6-13 show the

trust and confidence ratings respectively. Scenario was found to have a significant effect on the trust ratings, $F(5,234) = 7.469$, $p < .001$, as well as the confidence ratings, $F(5,202) = 2.279$, $p = .048$, during the non-failure scenarios.

Figure 6-12. Trust in aid's recommendation for each non-failure scenario

Figure 6-13. Diagnostic Confidence during appropriate reliance for each non-failure scenario

Bonferonni pairwise comparison for trust showed that Scenario D (Exacerbation of COPD) resulted in significantly lower trust ratings than all the other scenarios ($p < .004$) except Scenario E. Scenario E (Hypoglycaemic Attack) had significantly lower trust ratings than Scenario A ($p = .044$) and Scenario F ($p = .029$). However, pairwise comparisons with Bonferonni adjustments did not show any significant differences in the confidence ratings between the different scenarios.

The order of assignment of the two failure scenarios, Scenario G (Tension Pneumothorax) and Scenario H (Severe Hyperkalemia), between the two sessions was counterbalanced across the participants in each group. Independent sample t-tests were used to see if the scenario had an effect on the trust in the aid's recommendation and the diagnostic confidence during the trials where the participant agreed with the aid's erroneous recommendation. There was no significant difference found for trust (Scenario G: $M=4.655$, $SE=0.208$; Scenario H: $M=4.830$, $SE=0.206$), $t(78) = 0.597$, $p = .552$. There was also no significant differences found for confidence (Scenario G: $n=24$, $M=5.296$, $SE=0.209$; Scenario H: $n=28$, $M=5.481$, $SE=0.143$), $t(50) = 0.753$, $p = .455$.

6.7 Overall Subjective Measures of Automated Decision Aids

Table 6-P shows the results of the post-condition questionnaire items that were administered at the end of each session of trials. Scales for reverse-scaled items (R) were inverted for analysis. The Cronbach's Alpha values calculated for the items for Perceived Understanding and Overall System Trust are 0.703 and 0.767 respectively. 2 (between-subject, Cues: OFF vs ON) x 2 (within-subject, Like: OFF vs ON) repeated

measures ANOVA was performed on the results. The p -values for the test are presented in the table.

For the marginally significant interaction effects found at Overall System Trust and its components at TrustQ3 and TrustQ5, paired sample t -tests were done to analyze the simple effects of the Likelihood Ratings at the Cues-OFF and Cues-ON conditions respectively. A marginally significant simple effect of Likelihood Ratings was found for TrustQ3 at Cues-OFF, $t(19) = 1.89$, $p = .074$. The other simple effects in the other paired sample t -tests were found to be non-significant.

Independent sample t -tests were then done to analyze the simple effects of the Key Diagnostic Cues at the Like-OFF and Like-ON conditions respectively. Key Diagnostic Cues was found to have a significant simple effect on the Overall Trust at the Like-OFF condition, $t(38) = 2.195$, $p = .034$, and also a significant simple effect on TrustQ5 at the Like-OFF condition, $t(38) = 3.167$, $p = .003$. The other simple effects in the other independent sample t -tests were found to be non-significant.

6.8 Summary of Results

Table 6-Q provides an overview of the results according to the hypotheses presented in this study.

Table 6-P. Questionnaire items and responses

	Cues-OFF		Cues-ON		CUES	LIKE	CUES x LIKE
	Like-OFF	Like-ON	Like-OFF	Like-ON			
PERCEIVED UNDERSTANDING							
UndQ1. I did not know how the automated diagnostic aid arrived at its recommendations. (R)	4.36 <i>SD=1.50</i> Mdn=4.55	3.97 <i>SD=1.75</i> Mdn=4.30	4.38 <i>SD=1.69</i> Mdn=5.00	4.53 <i>SD=1.34</i> Mdn=5.00	.531	.570	.200
UndQ2. I understood why the diagnostic aid gave the recommendations that it did.	5.02 <i>SD=1.01</i> Mdn=5.10	4.96 <i>SD=1.24</i> Mdn=5.20	5.38 <i>SD=0.65</i> Mdn=5.25	5.22 <i>SD=0.63</i> Mdn=5.25	.240	.380	.710
OVERALL	4.69 <i>SD=1.21</i> Mdn=4.78	4.46 <i>SD=1.35</i> Mdn=4.68	4.88 <i>SD=1.04</i> Mdn=5.05	4.87 <i>SD=0.88</i> Mdn=5.05	.372	.397	.417
OVERALL SYSTEM TRUST							
TrustQ1. The diagnostic aid's recommendations provided me with security.	4.84 <i>SD=0.83</i> Mdn=5.00	5.20 <i>SD=0.99</i> Mdn=5.25	5.40 <i>SD=0.81</i> Mdn=5.55	5.33 <i>SD=0.83</i> Mdn=5.50	.124	.374	.203
TrustQ2. I was confident in the diagnostic aid's recommendations.	4.61 <i>SD=0.84</i> Mdn=4.75	5.02 <i>SD=0.84</i> Mdn=5.05	5.36 <i>SD=0.71</i> Mdn=5.30	5.18 <i>SD=0.85</i> Mdn=5.25	.021 **	.521	.100
TrustQ3. I trusted the diagnostic aid to provide me with accurate recommendations.	4.54 <i>SD=1.04</i> Mdn=4.75	5.00 <i>SD=1.09</i> Mdn=5.25	4.92 <i>SD=0.82</i> Mdn=4.95	4.80 <i>SD=0.95</i> Mdn=5.00	.738	.256	.055 *
TrustQ4. Using the recommendations from the diagnostic aid would have a harmful outcome. (R)	4.50 <i>SD=1.22</i> Mdn=4.90	4.60 <i>SD=1.20</i> Mdn=4.75	4.75 <i>SD=0.79</i> Mdn=4.90	4.54 <i>SD=1.25</i> Mdn=4.35	.764	.782	.436
TrustQ5. The recommendations from the diagnostic aid were reliable.	4.48 <i>SD=0.87</i> Mdn=4.65	4.89 <i>SD=1.07</i> Mdn=5.10	5.25 <i>SD=0.65</i> Mdn=5.05	5.07 <i>SD=0.63</i> Mdn=5.00	.033 **	.465	.055 *
TrustQ6. The recommendations from the diagnostic aid were deceptive. (R)	4.88 <i>SD=0.96</i> Mdn=5.00	4.74 <i>SD=1.19</i> Mdn=5.00	4.54 <i>SD=1.17</i> Mdn=4.65	4.38 <i>SD=1.24</i> Mdn=4.90	.282	.375	.940
OVERALL	4.64 <i>SD=0.56</i> Mdn=4.68	4.91 <i>SD=0.81</i> Mdn=4.99	5.03 <i>SD=0.57</i> Mdn=5.03	4.88 <i>SD=0.67</i> Mdn=5.03	.282	.649	.096 *

* p < .10, **p < .05

Table 6-Q. Summary of Results

	Trust in Aid's Recommendation	Diagnostic Confidence	
		if making the correct reliance decision	if making the wrong reliance decision
During Automation Non-Failure	H1A Rejected <i>No significant effect found</i>	H1B Rejected <i>Likelihood Ratings marginally decreased confidence, instead of increasing it.</i>	H1C Marginally Supported <i>Diagnostic Cues marginally decreased confidence.</i>
During Automation Failure	H2A Rejected <i>No significant effect found</i>	H2B Rejected <i>No significant effect found</i>	H2C Rejected <i>No significant effect found</i>
Effect of Experience of Failure	H3A Rejected <i>Diagnostic Cues increased the drop in trust, instead of reducing it. Drop is greater when Diagnostic Cues and Likelihood Ratings are present together.</i>	H3B Rejected <i>Diagnostic Cues increased the drop in confidence, instead of reducing it.</i>	
Effect of Experience of Return to Correct Performance	H4 Marginally Supported <i>Likelihood Ratings marginally increased trust recovery.</i>		

CHAPTER 7: DISCUSSION

7.1 Transparency, Trust and Reliance

As predicted from the literature, the trust that the participant had in the aid's recommendation during the scenario influenced the decision to rely or not rely on the aid's recommendation. Agreement with the aid corresponded with higher trust ratings. However, the study showed that there was not sufficient evidence to confirm that trust when the aid was performing correctly was significantly affected by the presence of the Key Diagnostic Cues or the Likelihood Ratings (Hypothesis H1A). Correspondingly, during the non-automation failure trials, the percentage of appropriate reliance was also not affected by the transparency configuration. This was also observed during the automation failure trials where there was no significant difference in the trust ratings between the transparency configurations when the aid was providing an erroneous recommendation (Hypothesis H2A).

The visibility of the underlying rationale through the Key Diagnostic Cues and the automation's level of certainty through the Likelihood Ratings had not significantly increased the impression of trustworthiness for the imperfect aid. This was observed at the high specificity level of trust (i.e. trust in the recommendation as measured during the experimental trials).

Based on the questionnaire responses, there was no difference in the participants' perceived understanding of how the system worked between the transparency configurations. The study found that perceived understanding positively predicted trust in the automated aid, which was consistent with the literature. However, the presence of the Key Diagnostic Cues and the Likelihood Ratings did not significantly affect the

perceived understanding of the aid compared to the aid that had none of the transparency features.

The participants' responses for their perceived understanding were likely to be due to their expertise in the subject matter as medical students. They had the ability to reasonably analyse the raw environmental data completely by themselves. Even when the system had not revealed the Key Diagnostic Cues that outlined its underlying rationale (Cues-OFF), the medical students felt that they were able to deduce how the system had derived the recommendations over the multiple trials by comparing its recommendations to their own interpretation of the raw data. Hence, despite the rationale not being directly observable, they still had the impression that they could understand how the system arrived at its recommendations. This may have been a contributing factor to why the mean rating for trust during the non-failure and failure trials and the overall perceived trust in the aid did not differ much between the different configurations.

During the experiment, there was no real need to rely on the aid based on their trust of the system because the participants could come up with the diagnosis themselves. If the scenarios had been more complex and the raw data had been more difficult to perceive or to analyse on their own, there may have been a greater need to rely on the aid for a decision. In such a situation, it is possible that a less transparent aid that did not provide any justifications to its diagnostic recommendations would have led to lower levels of perceived understanding and hence less trust.

This view was supported by the participants' comments about how they made use of the recommendation provided by the aid. During the post-experiment debriefing, most of the participants reported forming their own diagnostic decision first before

consulting the decision aid. Even though the aid's recommendation was available to them from the beginning of the scenario, these participants chose not to refer to the aid's recommendation until they had formed their own independent diagnosis from the available environmental cues, or at least narrowed down the diagnosis to a few possible choices. They then used the aid's recommendation as a check against their own diagnosis.

If the aid's recommendation matched their independent diagnosis, they could then proceed with the final decision. The participant would treat the matching recommendation as a piece of supporting evidence for their final decision. Some participants also reported that when the recommendation matched, it boosted their confidence in the chosen diagnosis. In such an interaction, where the aid was used to validate a correct diagnosis made independently, trust might not play a major role as there was no reliance decision to be made. It could also be said that since the aid matched the user's own independent diagnosis, there was little reason to distrust the aid, leading to higher trust ratings regardless of the transparency configuration during those trials. Although trust may not be a factor during this trial, it is likely to increase overall trust which can become a factor for future interactions as the user observes that the machine is capable of matching the user's own diagnostic skills.

However, if there was a mismatch between the aid's recommendation and the participants' diagnosis, trust became a more relevant factor in deciding whether to stick to their own diagnosis or to rely on the aid's. In this case, the reliance decision also depended on the user's self-confidence in being able to diagnose the patient. As suggested by Lee & Moray (1994), if their trust in the aid exceeds their self-confidence in their own ability, they are more likely to rely on the assistance from the aid. Some participants even reported that after having distrusted the aid and making a wrong

diagnosis, they said that they felt less confident of themselves and were more likely to switch their diagnosis to that of the aid's recommendation if there was mismatch in a later trial.

Transparency may have had a greater influence on trust and the reliance decision in the situation where there is a mismatch between the aid and the decision maker as compared to when there is a match. The additional information from the transparency features provides the user with a means comparing the validity of his independent diagnosis against the aid's recommendation. If the aid is performing correctly, this provides the opportunity to notice diagnostic cues that the aid had used but the user had missed. This was the case for a number of participants who reported changing their initial diagnosis to match the aid when they saw a piece of information on the Key Diagnostic Cues that they had overlooked.

It is noted as a limitation that the study did not differentiate whether the participant checked the aid after making an independent decision or not during each trial. As such, it was not able to isolate the effect of this variable on the trust. Future similar studies should take this factor into account when measuring the trust in the decision aid.

The displayed values of the Likelihood Ratings for the different scenarios were between the 60 and 70% range because it was a variable that was controlled across all the conditions as the minimum reliability for a user to accept the aid. However because of this, participants reported that the changes in the Likelihood Ratings were too small between the different trials that they felt that it did not matter. Some participants reported choosing to ignore the likelihood ratings altogether. They felt that it did not help them in the decision to agree or disagree with the aid because the value was

neither high enough nor low enough. This could have been a contributing factor to why the Likelihood Ratings had no significant effect on trust ratings.

7.2 Transparency and Confidence

When there was no automation failure, although there was no significant difference in trust across the different transparency configurations, participants were marginally less confident when appropriately relying on the aid during the session where the Likelihood Ratings were present (Hypothesis H1B). This decrement in confidence may be due to the participants being more aware of the vulnerability of the imperfect aid. While the aid was being more transparent about its performance in producing the given recommendation, it also made its level of reliability more salient to the user. The user did not need to recall past system performance to judge its reliability as the system itself was dynamically presenting a variable that represented it. The rating gave them a number for them to base their confidence on. This awareness of the system's imperfect reliability may have made the participants more mindful about accepting the system's recommendation without careful consideration of the accuracy of the outcome. Hence, this may have made them less sure about the correctness of their diagnosis, leading to the marginally lower confidence ratings.

However, considering the cases where the participants decided to disagree with the aid when it was actually performing correctly, the confidence ratings during these cases were found to be marginally lower when the Key Diagnostic Cues were presented (Hypothesis H1C). Participants who had access to the Key Diagnostic Cues and chose the wrong diagnosis had made the erroneous diagnosis with a lower confidence that they were correct. The availability of the aid's underlying rationale was likely to have

raised some doubts in the participant's reasoning. They may have noticed the aid's rationale, realized that the reasoning was compelling and felt that their own reasoning may not be valid as it may seem. Even though they ultimately relied on their own reasoning and committed an error in diagnosis, in actual practice, having lower confidence in a wrong diagnosis would make them more likely to be prepared for the possibility of error.

7.3 Transparency and the Experience of Automation Failure

The experience of automation failure in the second trial of the session expectedly caused the trust in the aid's recommendation to be lower in the first post-failure trial immediately after it, when compared to the trust in the pre-failure trial. Hypothesis H3A and Hypothesis H3B predicted that the presence of the transparency features would reduce this drop in trust by affording the user the ability to understand why and how the aid had committed the error. The study did not show this reduction of the trust drop but instead, it was found that the presence of the transparency features had caused the trust to drop significantly more than when it was not present.

Some participants in the Cues-ON condition who over-trusted the aid during the failure trial felt that the Key Diagnostic Cues had convinced them that its recommendation was correct when actually it was wrong. Hence, when they went into the first post-failure trial, they not only distrusted the recommendation, they also mistrusted the individual details provided by the Key Diagnostic Cues display. The compounded distrust of multiple elements led to much lower levels of trust compared to the aid that did not display the Key Diagnostic Cues.

The presence of the Key Diagnostic Cues had also caused the diagnostic confidence among participants who appropriately agreed with the aid to drop significantly more than when the transparency feature was absent. They felt much less confident even though they were correct as compared to when they were correct before experiencing the failure. This is probably because the diagnostic confidence was closely tied to the trust in the aid's recommendation. When the Key Diagnostic Cues caused an increased drop in trust, confidence in the final diagnosis correspondingly dropped as well. The participants had agreed with the aid but had felt less confident about the accuracy of their decision because they trusted the aid less and thus had a stronger impression that the aid could be wrong.

There was also a significant interaction effect between the presence of the Key Diagnostic Cues and the presence of the Likelihood Ratings. The Likelihood Ratings had no significant effect on the trust drop when the Key Diagnostic Cues were absent but its effect became significant when the Key Diagnostic Cues were present, further intensifying the trust drop. The trust drop was the greatest in the Cues-ON/Like-ON configuration. Appropriate reliance on the aid during the first post-failure trial also had the least percentage of participants who agreed with the aid.

According to some of the participants' comments, they felt that the Key Diagnostic Cues and the Likelihood Ratings contradicted one another. During the first post-failure trial, the participants felt that the rationale shown in the Key Diagnostic Cues was convincing enough and seemed to support the recommendation very well. However, they were unable to fathom why the Likelihood Rating value was not higher than what it was in the failure trial, which meant that the probability of an automation error was just as likely. When the underlying reasoning was supportive enough, they could not understand why there was a remaining 30-40% rating of uncertainty. They were unable

to rationalize the possible factors that could result in that much uncertainty in the aid. This may have led to a perception of low competence in the aid which then caused trust to drop further.

7.4 Transparency and the Experience of Return to Correct

Automation Performance

The presence of the transparency features was expected to improve the recovery of trust after the user has experienced a return to non-failure performance. The study provided some evidence of this as the presence of the Likelihood Ratings had marginally improved the recovery of trust (Hypothesis H4). The trust in the aid's recommendation was lower in the final trial of the session compared to the trust ratings in the first trial for participants who did not have access to the Likelihood Ratings. However, for participants in the Like-ON condition, the trust ratings in the final trial were higher than that of the initial trial.

The Likelihood Ratings had provided a salient cue for the reliability of the system which could then be observed over multiple uses of the system. The presence of the transparency feature had probably facilitated the build-up of trust as the user utilized the system repetitively. They may have felt that they could better understand how the system was performing. On the other hand, participants who did not have the Likelihood Ratings may have felt that the system was more unpredictable and therefore more difficult to trust.

However, it has to be noted that after the massive drop in trust due to the experience of error that was observed in the Cues-ON/Like-ON condition, there was little difference

in the trust ratings for that condition during the final trial as compared to the other configurations. Even though the participants may have felt that the aid with both the transparency features was the least trustworthy among the other transparency configurations immediately after the automation failure, it did not continue to feel as untrustworthy after the participant experienced a return to proper performance. The presence of the two transparency features had caused the trust to return to levels that were similar to the other transparency configurations after experiencing one non-failure trial.

7.5 Overall Assessment of Transparency Features

Despite there being no significant difference in trust between the different transparency configurations and a potential confidence decrement, participants still reported a higher preference for aids with transparency features. Participants who used both the aids in the CUES-OFF/LIKE-OFF and CUES-OFF/LIKE-ON conditions reported preferring to use the aid with the Likelihood Ratings if they had to use the aid in real life. However, some participants added that if the variation in the value of the Likelihood Ratings is too little, it will not be useful in helping them decide on whether it is safe to rely on the aid or not. This is a noted limitation of the study as there was a need to control the perceived reliability level of the aid so that the effect of the presence of the feature itself could be studied. If the Likelihood Ratings had varied more, it was expected to affect the trust and confidence levels accordingly and it would have been a more useful piece of information to the decision maker. Instead, the similar values during the session had confused the participants because it was not different enough between the failure and non-failure trials.

They also agreed that they would prefer it more if they could also see the actual rationale behind the aid's recommendations. This was echoed by some participants in the CUES-ON group who agreed that they preferred that the aid showed the Key Diagnostic Cues over not showing it. One of the advantages of the Key Diagnostic Cues, as noted by some of the participants, was that it helped them notice certain pieces of evidence that they may have not taken into consideration or may have missed.

This was supported by their responses in the questionnaire. Although trust during the actual scenario trials may not have been affected by the transparency configuration, participants may have perceived a difference at the lower specificity (i.e. trust in the overall system as measured in the questionnaire) when they were considering how much they trusted the whole system after having interacted with it during the session. In the presence of Key Diagnostic Cues, participants were found to have a significantly higher agreement for the questionnaire items "*I was confident in the diagnostic aid's recommendations*" and "*The recommendations from the diagnostic aid were reliable*". Likelihood Ratings was found to have a marginally significant increase in participant agreement when Key Diagnostic Cues were absent for the item "*I trusted the diagnostic aid to provide me with accurate recommendations*". The presence of the Key Diagnostic Cues was also found to have a significantly higher agreement on the item "*The recommendations from the diagnostic aid were reliable*" when the Likelihood Ratings were absent. Aggregating the component items for overall trust, there was no significant main effects found for either of the transparency features but for simple effect, Key Diagnostic Cues when the Likelihood Ratings were absent, was found to make the perceived overall trust significantly higher.

The results of the questionnaire seem to suggest that there was a higher overall trust in the system associated with the transparency features. Participant in the Cues-ON group

felt more confident with their aid's recommendation and perceived the aid to be more reliable with the Key Diagnostic Cues present. However, as expressed in their post-experiment debriefing, the positive effect the Key Diagnostic Cues had on trust was tempered by the presence of the Likelihood Ratings which seem to contradict the underlying rationale and making the system seem less competent. Therefore, individually, the transparency features may generate some benefits but combinations of them may erode the benefits if not executed properly.

7.6 Summary of Analysis

Figure 7-1 below illustrates the key findings of the study on the effects of transparency features on the evolution of trust over successive usage of an imperfect automated diagnostic decision aid.

Figure 7-1. Illustration of effects of transparency features on trust dynamics for an imperfect automation

The presence of diagnostic cues as a means of increasing the transparency of the system intensified the drop in trust that occurs immediately after an experiencing a failure by the automation. Diagnostic confidence was also found to have been decreased. Instead of reducing the trust drop as initially predicted, users trusted a transparent system much lesser even if it was performing properly if they had just experienced the system making a critical error in the usage in the immediately previous usage.

However, a potentially useful finding, although marginally significant, was that transparency in the form of the presence of Likelihood Ratings could improve the recovery of trust that occurs as the user experiences a return to correct performance. The difference in trust from the initial levels to the level found in the usage after experiencing one correct performance post-failure was smaller for the more transparent system.

Despite the larger trust drop, based on the analysis of individual items in the post-usage trust questionnaire and post-experiment comments of the participants, participants preferred a more transparent automated decision aid over one that provided its aid without any supporting information.

CHAPTER 8: CONCLUSION

8.1 Summary of Findings

Automation transparency had been identified in the literature as a potential design dimension for automated decision support systems to promote trust, acceptance and appropriate reliance. While the literature on transparency and trust has previously centred mainly on recommender systems in non-critical decision making tasks such as choosing movies and artwork, there has been a lack of empirical studies investigating the relationship between transparency and trust in decision making domains that involve time pressure and higher risks. This research has addressed this gap by providing empirical evidence of the effect in such a domain, namely in the area of emergency medical diagnostic decision making. It has done so through an experiment that closely emulates a naturalistic decision making setting through the use of a high-fidelity patient simulator. The study had aimed to investigate if the presence of automation transparency features would lead to more appropriate trust and reliance and ultimately better performance through higher confidence in appropriate reliance and lower confidence in non-appropriate reliance. It has also shed some light on the interaction between transparency and the effect of the experience of failure and the subsequent return to non-failure performance, which is an area that has not been sufficiently addressed in the literature but has importance relevance for the interaction with imperfect automation.

The study was not able to show significant evidence that manipulating the transparency of the automation through the availability of Key Diagnostic Cues and Likelihood Ratings would affect trust and high specificity (trust in aid's recommendation during the actual trials) during failure and non-failure performance. It

could not be shown that the presence of the transparency features would lead to higher trust during correct performance or lower trust during automation failure. There was also some marginal evidence that the Likelihood Ratings had instead led to an adverse effect on performance, reducing the diagnostic confidence during appropriate reliance.

Another key finding of the study was that the transparency features could intensify the drop in trust and confidence immediately after the experience of an automation failure. Experiencing an automation failure with a system that was more transparent could lead to unwillingness to use the system immediately after the experience even though the system had returned to normal performance and collaboration with it was beneficial.

Although the presence of the transparency features may have caused some drawbacks, the study has shown that there may be some potential benefits. Despite the larger drop in trust due to the presence of the transparency features, the transparent system quickly recovered trust to similar levels of the less transparent system after the participants experienced the return to non-failure performance. The presence of the Likelihood Ratings even caused the trust to be greater than its initial levels after multiple uses. There was also some marginal evidence that the presence of the Key Diagnostic Cues led to a lower diagnostic confidence when inappropriately choosing to not rely on an aid that was actually performing properly. Lower confidence when making a misdiagnosis is beneficial to performance as it could make the user more prepared to handle the possibility of a misdiagnosis. Furthermore, questionnaire responses at the lower trust specificity (overall trust in system) seem to reflect that the participants trusted the system more when one of the transparency features was available, but not necessarily when both were present. This could lead to a higher acceptance to use the system in a more practical setting.

8.2 Practical Implications of Findings

This research has provided some useful insights into how trust and performance is affected by user interface design features that provide transparency information to the user in a human-automation collaborative decision making task.

When an automated diagnostic decision aid is imperfect and is able to make mistakes, it is important to not just design for higher trust, but to design for appropriate trust. Its design should support trust calibration by providing some degree of transparency. As a guideline for designing for trust calibration, the developer of an automated decision aid should allow the user to inspect, to some extent, the back-end processes that lead to the decision recommended by the system in order to compare it against the user's own decision. Then the user is able to decide at that particular moment of operation, whether the recommendation is reliable or not reliable. At the very least, the feature should prompt the user to recognize that an error was possible. The study has shown that the overall trust over the whole system, which leads to acceptance, can be improved if the automated decision aid is designed to provide either a rationale behind its recommendation or a confidence rating for its recommendation.

System developers also have to accept that the user of a more transparent system may suffer a greater trust and confidence detriment immediately after experiencing an automation error. However, the results of the study suggest that the user is able to regain the trust back faster when the design is more transparent. This has implications on the long-term acceptance and usage of the system. A less transparent system may become too untrustworthy to continue using after the user has experienced his first automation error because it takes too long to recover the trust that it may end up being disused in the long-term.

The developer of the automated decision aid also needs to be careful with the means of increasing system transparency. In the designs used for the study, increasing transparency by providing the key diagnostic cues as well as the likelihood ratings led to a lower overall trust in the system, compared to when they were used individually. This was because the system was perceived to be not working properly when the given cues and the likelihood rating did not seem to tally to the user. Thus, it is important that if multiple means of showing transparency is utilized, they should be well integrated with one another. One unexplored alternative of combining the two transparency features is to show the individual confidence rating for the diagnosticity of each cue, which is how strong each diagnostic cue supports the given recommendation.

8.3 Limitations of Study

The findings of this research study are limited to human-automation collaborative decision making by novice physicians. It is acknowledged that more experienced physicians with higher expertise may interact differently with the decision aid and the transparency features. It is possible that their reservoir of domain knowledge and more sophisticated decision making heuristics may make them less dependent on the decision aid. Furthermore, their expertise may allow them to scrutinize the transparency information more rigorously, which may have different effects on trust when using the aid.

The applicability of the study may be relevant for decision making domains that are similar to diagnostic decision making, where the decision maker needs to gather evidence from the decision making environment in order to form an understanding of

what is going on. It may also apply to human-automation collaborative decision making tasks which involve the selection and execution of a response after an assessment of the situation, because in such situations, providing transparency information underlying the logic behind the action is of relevance to the user. However, in this study, participants have full access to the raw data to be used for decision making. They can reasonably consider all the necessary and available information in order to make a decision independently from the decision aid. In certain decision making domains, informational cues may be as available and accessible but may too voluminous to completely consider all or even most of them. In other domains, some of the necessary raw data may not be directly perceivable by the user but can only be sensed and processed meaningfully by the automation. In these cases where the user may not be able to perceive all the required raw data to form a decision, there will be a greater reliance on the automated decision aid or even no choice but to rely on the aid's assistance. The dynamics of the interaction and how trust plays a part in the decision making may be different, leading to possible differences in the effects of transparency not covered by this study.

Lastly, although efforts were made in the experiment to emulate as close as possible to the naturalistic decision making environment involved in emergency diagnosis, it is acknowledged that the decision making task is still only a simulation where the risk associated to the decision outcome is not as high as in actual practice. In practice, where the potential risks such as the possibility of physical risk to the patient is more real, the reliance decision and its outcome may be even more critical and thus, the effects of trust and the effect of transparency on that trust may be more acute. Nonetheless, the study does provide a reasonable representation of the possible dynamics involved.

8.4 Future Directions

While other studies in non-emergency decision making domains have shown significant improvements in automation trust due to introduction of transparency features, this study was not able to replicate the findings, especially at high trust specificity, in a domain that involves more time pressure and higher risks. Despite this, participants did report a higher preference for decision aids with transparency features over those without.

Future studies should look into investigating the other factors involved in this higher preference for transparency features. One possible direction is to introduce more complexity into the diagnostic decision making task, which increase the need to rely on the aid. It is expected that transparency will have a greater impact in such a task. Future studies should also replicate the study in other decision making domains such as process control and military where decision making based on complex evaluation of environmental data could also lend relevance to transparency features. Another variable that should be considered in other studies should be the number of trial repetitions. As a logistical constraint of the decision making trust in this study, the number of repetitions within a session was limited to four. As such the interaction between transparency and the effects of failure and the effect of the return to non-error performance could not be studied over a longer time-scale, which could reveal more about the trend of trust recovery.

In conclusion, this study has provided useful insights into how automation transparency affects trust and the dynamics of collaborative decision making between human and automation. It has highlighted some potential pitfalls that may be caused by introducing the transparency features while also revealing some potential benefits.

Therefore designers of automated decision support need to carefully consider such factors when deciding the appropriateness of certain transparency features for their user interfaces.

[INTENTIONALLY LEFT BLANK]

REFERENCES

- Barnett, G. O., Cimino, J. J., Hupp, J. A., & Hoffer, E. P. (1987). An evolving diagnostic decision-support system. *JAMA: The Journal of the American Medical Association*, 258, 67-74.
- Barnett, G. O., Famiglietti, K. T., Kim, R. J., Hoffer, E. P., & Feldman, M. J. (1998). *DXplain on the Internet*. Paper presented at the Proceedings of the AMIA Symposium.
- Barrows, H. S., & Feltevich, P. J. (1987). The clinical reasoning process. *Medical Education*, 21(2), 86-91.
- Bauer, B. A., Lee, M., Bergstrom, L., Wahner-Roedler, D. L., Bundrick, J., Litin, S., ... & Elkin, P. L. (2002). Internal medicine resident satisfaction with a diagnostic decision support system (DXplain) introduced on a teaching hospital service. In *Proceedings of the AMIA Symposium* (p. 31). American Medical Informatics Association.
- Beller, J., Heesen, M., & Vollrath, M. (2013). Improving the Driver–Automation Interaction: An Approach Using Automation Uncertainty. *Human Factors: The Journal of the Human Factors and Ergonomics Society*.
- Boguševičius, A., Maleckas, A., Pundzius, J., & Skaudickas, D. (2003). Prospective randomised trial of computer-aided diagnosis and contrast radiography in acute small bowel obstruction. *European Journal of Surgery*, 168(2), 78-83.
- Bond, W. F., Schwartz, L. M., Weaver, K. R., Levick, D., Giuliano, M., & Graber, M. L. (2012). Differential diagnosis generators: an evaluation of currently available computer programs. *Journal of General Internal Medicine*, 27(2), 213-219.

- Bornstein, B. H., & Emler, A. C. (2001). Rationality in medical decision making: a review of the literature on doctors' decision-making biases. *Journal of Evaluation in Clinical Practice*, 7(2), 97-107.
- Buchanan, B. G., & Shortliffe, E. H. (1984). *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project*: Addison-Wesley.
- Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4), 331-370.
- Charlin, B., Tardif, J., & Boshuizen, H. P. A. (2000). Scripts and Medical Diagnostic Knowledge: Theory and Applications for Clinical Reasoning Instruction and Research. *Academic Medicine*, 75(2), 182-190.
- Coderre, S., Mandin, H., Harasym, P. H., & Fick, G. H. (2003). Diagnostic reasoning strategies and diagnostic success. *Medical Education*, 37(8), 695-703.
- Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., et al. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. [Article]. *User Modeling and User-Adapted Interaction*, 18(5), 455-496.
- Crandall, B., Klein, G., & Hoffman, R. R. (2006). *Working Minds: A Practitioner's Guide to Cognitive Task Analysis*. Cambridge, MA: The MIT Press.
- Croskerry, P., & Sinclair, D. (2001). Emergency medicine: A practice prone to error? *Canadian Journal of Emergency Medicine*, 3(4), 271-276.
- De Dombal, F., Leaper, D., Staniland, J. R., McCann, A., & Horrocks, J. C. (1972). Computer-aided diagnosis of acute abdominal pain. *British Medical Journal*, 2(5804), 9.

- Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6), 697-718.
- Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 44(1), 79-94.
- Elstein, A. S. (2009). Thinking about diagnostic thinking: a 30-year perspective. *Advances in Health Sciences Education*, 14, 7-18.
- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1), 32-64.
- Fordyce, J., Blank, F. S., Pekow, P., Smithline, H. A., Ritter, G., Gehlbach, S. et al (2003). Errors in a busy emergency department. *Annals of Emergency Medicine*, 42(3), 324-333.
- Garg, A. X., Adhikari, N. K., McDonald, H., Rosas-Arellano, M. P., Devereaux, P., Beyene, J., et al. (2005). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes. *JAMA: The Journal of the American Medical Association*, 293(10), 1223-1238.
- Graber, M. (2005). Diagnostic errors in medicine: a case of neglect. *Joint commission journal on quality and patient safety*, 31(2), 106-113.
- Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic error in internal medicine. *Archives of Internal Medicine*, 165(13), 1493-1499.
- Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). *Explaining collaborative filtering recommendations*. Paper presented at the Proceedings of the 2000 ACM conference on Computer supported cooperative work.

- Holohan, T. V., Colestro, J., Grippi, J., Converse, J., & Hughes, M. (2005). Analysis of Diagnostic Error in Paid Malpractice Claims with Substandard Care in a Large Healthcare System. *Southern Medical Journal*, 98(11), 1083-1087.
- Höst, M., Regnell, B., & Wohlin, C. (2000). Using students as subjects—a comparative study of students and professionals in lead-time impact assessment. *Empirical Software Engineering*, 5(3), 201-214.
- Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1), 53-71.
- Kachalia, A., Gandhi, T. K., Puopolo, A. L., Yoon, C., Thomas, E. J., Griffey, R., et al. (2007). Missed and delayed diagnoses in the emergency department: A study of closed malpractice claims from 4 liability insurers. *Annals of Emergency Medicine*, 49(2), 196-205.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3), 430-454.
- Kantowitz, B. H., Hanowski, R. J., & Kantowitz, S. C. (1997). Driver Acceptance of Unreliable Traffic Information in Familiar and Unfamiliar Settings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39(2), 164-176.
- Khasawneh, M. T., Bowling, S. R., Jiang, X., Gramopadhye, A. K., & Melloy, B. J. (2004). Effect of Error Severity on Human Trust in Hybrid Systems. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 48, No. 3, pp. 439-443).
- Koenderink, N. J., Broekstra, J., & Top, J. L. (2010). Bounded transparency for automated inspection in agriculture. *Computers and Electronics in Agriculture*, 72(1), 27-36.

- Kohn, L. T., Corrigan, J., & Donaldson, M. S. (2000). *To err is human: building a safer health system*. The National Academies Press.
- Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10), 1243-1270.
- Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human Computer Studies*, 40(1), 153.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50-80.
- Lee, R. B., Vicki, R. L., & Geoffrey, R. N. (2000). On the Difficulty of Noticing Obvious Features in Patient Appearance. *Psychological Science*, 11, 112-117.
- Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). *Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems*. New York: Assoc Computing Machinery.
- Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors*, 48(2), 241-256.
- Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience. *Journal of Cognitive Engineering and Decision Making*, 6(1), 57-87.
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *Academy of Management Review*, 20(3), 709-734.

- McGuirl, J. M., & Sarter, N. B. (2006). Supporting Trust Calibration and the Effective Use of Decision Aids by Presenting Dynamic System Confidence Information. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 48(4), 656-665.
- Mortensen, T., Fisher, R., & Wines, G. (2012). Students as surrogates for practicing accountants: Further evidence. In *Accounting Forum* (Vol. 36, No. 4, pp. 251-265). Elsevier.
- Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27(5), 527-539.
- Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3), 429-460.
- Musen, M. A., Shahar, Y., & Shortliffe, E. H. (2006). Clinical decision-support systems. *Biomedical informatics*, 698-736.
- Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology*, 29(1), 85 - 95.
- Newman-Toker, D. E., & Pronovost, P. J. (2009). Diagnostic errors—the next frontier for patient safety. *JAMA: The Journal of the American Medical Association*, 301(10), 1060-1062.
- Norman, S. M., Avolio, B. J., & Luthans, F. (2010). The impact of positivity and transparency on trust in leaders and their perceived effectiveness. *The Leadership Quarterly*, 21(3), 350-364.

- O'Connor, P. M., Dowey, K. E., Bell, P. M., Irwin, S. T., & Dearden, C. H. (1995). Unnecessary delays in accident and emergency departments: Do medical and surgical senior house officers need to vet admissions?. *Journal of Accident & Emergency Medicine*, 12(4), 251-254.
- Orasanu, J., & Connolly, T. (1993). The Reinvention of Decision Making. In G. A. Klein, J. Orasanu, R. Calderwood & C. E. Zsombok (Eds.), *Decision Making in Action: Models and Methods* (pp. 3-20). New Jersey: Ablex Publishing Corporation.
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 52(3), 381-410.
- Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, 47(4), 51-55.
- Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance Consequences of Automation-Induced 'Complacency'. *The International Journal of Aviation Psychology*, 3(1), 1-23.
- Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230-253.
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems Man and Cybernetics Part a-Systems and Humans*, 30(3), 286-297.
- Pozen, M. W., D'Agostino, R. B., Selker, H. P., Sytkowski, P. A., & Hood Jr, W. B. (1984). A predictive instrument to improve coronary-care-unit admission practices in acute ischemic heart disease. *New England Journal of Medicine*, 310(20), 1273-1278.

- Rasmussen, J., & Goodstein, L. (1987). Decision support in supervisory control of high-risk industrial systems. *Automatica*, 23(5), 663-671.
- Rauch, H. (1984). *Expert systems as automated decision aids*. Paper presented at the Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84.
- Reisman, Y. (1996). Computer-based clinical decision aids. A review of methods and assessment of systems. *Informatics for Health and Social Care*, 21(3), 179-197.
- Rogers, D. P. (1987). The development of a measure of perceived communication openness. *Journal of Business Communication*, 24(4), 53-61.
- Röttger, S., Bali, K., & Manzey, D. (2009). Impact of automated decision aids on performance, operator behaviour and workload in a simulated supervisory control task. *Ergonomics*, 52(5), 512-523.
- Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a General Theory of Planning. *Policy Sciences*, 4(2), 155-169.
- Sandhu, H., & Carpenter, C. (2006). Clinical decisionmaking: Opening the black box of cognitive reasoning. *Annals of Emergency Medicine*, 48(6), 713-719.
- Schiff, G. D., Kim, S., Abrams, R., Cosby, K., Lambert, B., Elstein, A. S., et al. (2005). Diagnosing Diagnosis Errors: Lessons from a Multi-institutional Collaborative Project. In K. Henriksen, J. B. Battles, E. S. Marks & D. I. Lewin (Eds.), *Advances in Patient Safety: From Research to Implementation* (Vol. 2, Concepts and Methodology). Rockville, MD: Agency for Healthcare Research and Quality.

- Selker, H. P. Beshansky, J. R., Griffith, J. L., Aufderheide, T. P., Ballin, D. S. et al (1998). Use of the Acute Cardiac Ischemia Time-Insensitive Predictive Instrument (ACI-TIPI) To Assist with Triage of Patients with Chest Pain or Other Symptoms Suggestive of Acute Cardiac. *Annals of Internal Medicine*, 129(11).
- Seong, Y., & Bisantz, A. M. (2002). Judgment and Trust in Conjunction with Automated Decision Aids: A Theoretical Model and Empirical Investigation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 46(3), 423-427.
- Sheridan, T. B. (1975). *Considerations in modeling the human supervisory controller*. Paper presented at the International Federation of Automatic Control, Triennial World Congress, 6 th, Boston and Cambridge, Mass.
- Sheridan, T. B. (1988). *Trustworthiness of command and control systems*. Paper presented at the IFAC Man-Machine System.
- Singh, H., Thomas, E. J., Wilson, L., Kelly, A., Pietz, K., Elkeeb, D., et al. (2010). Errors of Diagnosis in Pediatric Practice: A Multisite Survey. *Pediatrics*, 126(1), 70-79.
- Sinha, R., & Swearingen, K. (2002). *The role of transparency in recommender systems*. Paper presented at the CHI'02 extended abstracts on Human factors in computing systems.
- Sorkin, R. D., Kantowitz, B. H., & Kantowitz, S. C. (1988). Likelihood alarm displays. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 30(4), 445-459.
- Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. [Review]. *Psychological Review*, 115(1), 155-185.

- Tintarev, N., & Masthoff, J. (2007). *A survey of explanations in recommender systems*. Paper presented at the Data Engineering Workshop, 2007 IEEE 23rd International Conference on.
- Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2), 207-232.
- Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty - Heuristics and Biases. *Science*, 185(4157), 1124-1131.
- van Dongen, K., & van Maanen, P.-P. (2013). A framework for explaining reliance on decision aids. *International Journal of Human-Computer Studies*, 71(4), 410-424.
- Wallsten, T. S., & Barton, C. (1982). Processing probabilistic multidimensional information for decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(5), 361-384.
- Wears, R. L. (2009). What makes diagnosis hard? *Advances in Health Sciences Education*, 14, 19-25.
- Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, 8(3), 201-212.
- Wickens, C. D., & Hollands, J. G. (2000). *Engineering Psychology and Human Performance* (Third ed.). Upper Saddle River, NJ: Prentice-Hall.
- Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: the effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science*, 2(4), 352-367.

- Wilson, R. M., Runciman, W. B., Gibberd, R. W., Harrison, B. T., Newby, L., & Hamilton, J. D. (1995). The Quality in Australian health care study. *Medical Journal of Australia*, 163(9), 458-471.
- Wyatt, J., & Spiegelhalter, D. (1991). Field trials of medical decision-aids: potential problems and solutions. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 3-7. American Medical Informatics Association.
- Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, 83(2), 260-281.
- Yeh, M., & Wickens, C. D. (2001). Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 43(3), 355-365.

APPENDIX I – Exploratory Study: Selected Participant Utterances

Participant 01 (Scenario: Severe Asthma Attack)

“I was watching the monitor, I noticed the blood pressure was dropping. So at that point I thought that because he had a fracture, he was having blood loss, hypovolemia. So I decided to give him fluids...(That was) based on the blood pressure plus the heart rate...Plus the central venous pressure was very low” [01.1]

“I was fairly certain he had some form of hypovolemia. Just that I think I didn’t give enough fluids... I think throughout the thing, I was certain he had (hypovolemia) but that he had more than one diagnosis... I think throughout, there was something more. Usually hypovolemia alone should not cause breathlessness. There should be other things. Likely to be other things...” [01.2]

Participant 03 (Scenario: Hypovolemia due to bleeding from fracture site)

“I think there was shortness of breath, as well as SPO2 95. So that’s why we gave oxygen. Then the BP was low and the CVP was too low... (This led to giving) him more fluids. The goals and concerns was mainly to stabilize him first...And then, we were unsure whether to give more fluids, because we weren’t sure of what the guidelines were.” [03.1]

“So because the patient said that the patient did not eat anything and feeling giddy, so we started to think about the (history of) diabetes... and then we suspected hypoglycaemia or DKA...We weren’t sure because they didn’t give us the results (of the hypocount) back, we couldn’t decide if we should... we were unsure if we could give glucose or insulin to treat the suspected hypoglycaemia or DKA.” [03.2]

“...we gave and did stuff already but he’s still unstable, so we had to look for causes but we could not find anything. So that’s why right now we go back our diabetes thing. Because we don’t know what else to do already...” [03.3]

Participant 05 (Scenario: Anaphylaxis)

“We suspected pulmonary embolism ... Because he had a fractured leg so, it’s a known complication. But at this time, his saturation was still very good. He wasn’t very tachycardic either.” [05.1]

"... I guess I was thinking of pulmonary embolism but then it wasn't so bad... As in, it's keeping with the picture (of somebody suffering a fractured leg after falling off a bike) but it's not that bad. Like, it could be other things... (The conflicting clues were) his saturation, the oxygen was still very good... and his heart rate wasn't that fast..." [05.2]

"Someone said he was stridorous... I think it makes it more complicated, because in pulmonary embolism you don't get stridor. You get chest pain, you get shortness of breath but you don't get stridor. So that was very off" [05.3]

"He was coughing a lot so we asked him if he was coughing out blood... I think we also asked like does he have any phlegm?... We are concerned that we were missing a diagnosis because... it didn't match the whole picture. Like, fractured leg and cough." [05.4]

Participant 07 (Scenario: Severe Asthma Attack)

"... We knew somehow that it was a lung cause. Because when we auscultated, we heard something. But we didn't know what the exact problem of his breathlessness." [07.1]

"... The patient started desaturating, and so, without having in mind what the cause was, but we knew something was wrong with the lungs... everyone was focused on trying to give him more oxygen so that his saturation would go up. We weren't focused on the root cause of it. We weren't trying to find it out. Though after some examination, we were thinking this time about some COPD, chronic obstructive pulmonary disease. We were thinking about that, and probably pneumothorax. But then we kept doing the physical examination. We were still not so sure... It was not very convincing what the diagnosis was." [07.2]

"...We didn't know what the root cause was. Then we auscultated and found that there was unequal breath sounds. But we just thought, unequal breath sounds, so what?... We actually thought of pneumothorax at that time. Some people were saying pneumothorax. Some people were just quiet. I guess some of the team members didn't know what to do next and some of the team members thought it was pneumothorax. But no one acted on it." [07.3]

Participant 08 (Scenario: Hypovolemia due to bleeding from fracture site)

"I asked one of my teammates to check the wound site ... I guess there was an issue here because the thing is that, not knowing what to expect from the mannequin... I suppose in a real life patient, if you were to see a draped up wound which is actively bleeding versus a draped up wound which is not actively bleeding, it would look more realistic. So we didn't know what was going on. But right now, it just looks red. Any wound bandage would look red. It didn't look like it was actively bleeding. We didn't know if we could go about actually undressing the wound or anything. So I guess we just assumed that the wound was alright... (When I saw the red stained bandage,) I thought that it was normal because most bandages would not be too clean." [08.1]

"I looked at the stats and then we decided more or less that he was in a case of going towards hypovolaemic shock... (I knew that because) He was tachycardic, high heart rate. And then his blood pressure was a bit low. His CVP was low as well. And also he was desaturating." [08.2]

"...By right, it looked like it was hypovolemia and if he had (it, then when) we put in fluids to him, he should actually respond. And then his giddiness should get better. But it didn't seem that he was getting better." [08.3]

"Actually looking at this scenario, I wouldn't have been too afraid of the AMI just yet, but rather, this is a very clear picture of hypovolaemia so I should actually in my mind list out what are the causes of hypovolaemia and then just run through them. And if I had done that, I would have figured out that continual blood loss and the wound site would have been one of them and then we would have done something about the wound, which would probably have solved the problem instead of getting distracted by the AMI." [08.4]

APPENDIX II – Detailed Views of Transparency Configurations

No Transparency Features (Cues-OFF/Like-OFF)

Suggested Diagnosis

Congestive Cardiac Failure

Patient Information

Name: [REDACTED]
Age/Gender: 70/Male

Non-Smoker;
History of Hypertension;
History of Ischemic Heart Disease;
History of Diabetes with neuropathy,
retinopathy and nephropathy

C/O Breathlessness

Likelihood Ratings only (Cues-OFF/Like-ON)

Patient Information

Name: [REDACTED]
Age/Gender: 70/Male

Non-Smoker;
History of Hypertension;
History of Ischemic Heart Disease;
History of Diabetes with neuropathy,
retinopathy and nephropathy

C/O Breathlessness

Suggested Diagnosis

Congestive Cardiac Failure

Diagnosis Likelihood

64%

Key Diagnostic Cues only (Cues-ON/Like-OFF)

Patient Information	Suggested Diagnosis
Name: [REDACTED] Age/Gender: 70/Male Non-Smoker; History of Hypertension; History of Ischemic Heart Disease; History of Diabetes with neuropathy, retinopathy and nephropathy C/O Breathlessness	Congestive Cardiac Failure Key Risk Factors <ul style="list-style-type: none">> Hypertension> Ischemic Heart Disease> Diabetes with end organ involvement> CVP high (CVP:15)> Hypotensive (BP:90/50)> Tachycardic (HR:115)

Key Diagnostic Cues and Likelihood Ratings (Cues-ON/Like-ON)

Patient Information	Suggested Diagnosis
Name: [REDACTED] Age/Gender: 70/Male Non-Smoker; History of Hypertension; History of Ischemic Heart Disease; History of Diabetes with neuropathy, retinopathy and nephropathy C/O Breathlessness	Congestive Cardiac Failure Diagnosis Likelihood 64% Key Risk Factors <ul style="list-style-type: none">> Hypertension> Ischemic Heart Disease> Diabetes with end organ involvement> CVP high (CVP:15)> Hypotensive (BP:90/50)> Tachycardic (HR:115)

CONFIDENTIAL

APPENDIX III – Details of Scenario Designs

Sce.	Actual Diagnosis	Aid's Recommended Diagnosis	Likelihood Rating
A	Haemorrhagic Shock	Haemorrhagic Shock	60
B	Congestive Cardiac Failure	Congestive Cardiac Failure	64
C	Acute Myocardial Infarction	Acute Myocardial Infarction	68
D	Exacerbation of COPD	Exacerbation of COPD	66
E	Hypoglycemic Attack	Hypoglycemic Attack	69
F	Diabetic Ketoacidosis	Diabetic Ketoacidosis	61
G	Tension Pneumothorax	Cardiac Tamponade (ERROR)	64
H	Hyperkalemia	Hypoglycemic Attack (ERROR)	67
I	Septic Shock	Septic Shock	63
J	Cerebrovascular Accident	Cerebrovascular Accident	67
K	Patient Normal	-	-

SCENARIO A – Haemorrhagic Shock

The patient is [REDACTED], a 30 year old man, a smoker, and a motorbike rider. He met with road traffic accident and sustained a fracture in the right femur and the right ilium with bruising on the right chest. You have been called to attend to him as he is complaining of backache and increasing drowsiness.

P/E	Tachycardia, Hypotension, Low CVP, lungs clear
ParaM	HR 118/min, BP 89/40mmHg, CVP – 2 cmH ₂ O, RR 24/min, SpO ₂ 95% Temp 36.7°C
Inx	Hb 10g/dl
Moulage	Bruising right chest and flank

ADA Patient Info	ADA Key Cues
Name: [REDACTED] Age/Gender: 30/Male Smoker(10pack-years); Admitted after Road Traffic Accident; Fractured right femur and right ilium; Bruising on right chest; Previous Hb 9g/dl; C/O increasing drowsiness, pain & bruising in right flank	> Fractured femur > Fractured ilium > CVP low (CVP:0) > Hypotensive (BP:89/40) > Tachycardic (HR:118)

CONFIDENTIAL

SCENARIO B – Congestive Cardiac Failure

The patient is [REDACTED], a 70 year old man, a non-smoker, and a retired teacher. He has a past history of hypertension, ischaemic heart disease and diabetes mellitus with neuropathy, retinopathy and nephropathy. You have been called to attend to him as he is complaining of breathlessness.

P/E	Tachycardia, Hypotension, High CVP, Lungs inspiratory crepitations
ParaM	HR 115/min, BP 90/50mmHg, CVP 15cmH2O, RR 26/min/ SPO2 93%, Temp 36.7°C, weak pulse
Inx	CXR: bat's wing appearance, ECG: Sinus tachycardia (Both not given)
Moulage	-

ADA Patient Info	ADA Key Cues
Name: [REDACTED] Age/Gender: 70/Male Non-Smoker; History of Hypertension; History of Ischemic Heart Disease; History of Diabetes with neuropathy, retinopathy and nephropathy C/O Breathlessness	> Hypertension > Ischemic Heart Disease > Diabetes with end organ involvement > CVP high (CVP:15) > Hypotensive (BP:90/50) > Tachycardic (HR:115)

SCENARIO C – Acute Myocardial Infarction

The patient is [REDACTED], a 45 year old man, a smoker, and an accountant. He has a past history of bronchial asthma, hyperlipidemia. He was admitted after experiencing nausea and epigastric discomfort after lunch at work. You have been called to attend to him as he is complaining of nausea and chest discomfort.

P/E	Tachycardia, Normal BP, CVP, Lungs clear
ParaM	HR 105/min, BP 110/60mmHg, CVP 9 cmH2O, RR 23/min, SpO2 95% Temp 37.4°C
Inx	ECG – sinus tachycardia with ST changes and ectopics, TWC 11, 500 (not given)
Moulage	Sweating

ADA Patient Info	ADA Key Cues
Name: [REDACTED] Age/Gender: 45/Male Smoker(20pack-years); History of Bronchial Asthma; History of Hyperlipidemia; Admitted for nausea and lower chest discomfort	> Male > Smoker > Hyperlipidemia > Chest discomfort with nausea > ECG sinus tachycardia with ST Changes and ectopics

CONFIDENTIAL

SCENARIO D – Exacerbation of COPD

The patient is [REDACTED], a 68 year old man, a smoker and a retiree. He has a past history of pulmonary tuberculosis which is treated, hypertension and hyperlipidemia. He had undergone an operation for a fractured neck of femur and has been in post-op for 1 day. You have been called to attend to him as he is complaining of chest tightness.

P/E	Lungs bilateral wheezing
ParaM	HR 108/min, BP 166/95mmHg, CVP 8 cmH2O, RR 26/min, SpO2 91% Temp 36.9°C
Inx	CXR: Hyperinflated lungs
Moulage	-

ADA Patient Info	ADA Key Cues
Name: [REDACTED] Age/Gender: 68/Male Smoker(60pack-years); History of Pulmonary Tuberculosis (treated); History of Hypertension; History of Hyperlipidemia; CXR revealed hyperinflated lungs; 1 day Post-op for fractured neck of femur; C/O chest tightness	> Smoker (High pack-years) > Chest Tightness > Hyperinflated lungs

SCENARIO E – Hypoglycemic Attack

The patient is [REDACTED], a 21 year old woman, and a university student. She has a past history of hyperthyroidism and amenorrhea. She had developed acute abdominal pain while sitting for an exam paper. She was diagnosed to have acute appendicitis and was waiting for an operation. You have been called to attend to her because she was found drowsy by the hospital staff.

P/E	Lungs clear
ParaM	HR 98/min, BP 90/45mmHg, CVP 3 cmH2O, RR 22/min, SpO2 96% Temp 38.3°C
Inx	H/C 1.2 mmol/l
Moulage	sweating

ADA Patient Info	ADA Key Cues
Name: [REDACTED] Age/Gender: 21/Female Non-Smoker; History of Hyperthyroidism; History of Amenorrhea; Admitted for acute abdominal pain; Diagnosed with acute appendicitis and awaiting op for 18hrs; Previous H/C 2.0mmol/l C/O drowsiness	> Hypocount low (H/C:2.0) > Prolonged fasting > Hypermetabolic state

CONFIDENTIAL

SCENARIO F – Diabetic Ketoacidosis

The patient is [REDACTED], a 48 year old man, and a durian seller. He has a past history of hypertension, hyperlipidemia and fatty liver disease. He had recently lost 5kg over the last two weeks. You have been called to attend to him as he is complaining of nausea, abdominal discomfort and pain when passing urine.

P/E	Lungs clear
ParaM	HR 95/min, BP 90/45mmHg, CVP 1 cmH2O, RR 22/min, SpO2 93% Temp 37.8°C
Inx	H/C 18mmol/l, U ketones ++, Metabolic acidosis pH 7.28 HCO2 14
Moulage	acetone breath

ADA Patient Info	ADA Key Cues
Name: [REDACTED] Age/Gender: 48/Male History of Hypertension; History of Hyperlipidemia; History of Fatty Liver Disease; Recent weight loss (5kg over 2 weeks); Previous H/C 18mmol/l; C/O nausea, abdominal discomfort and dysuria	> High weight loss > Nausea > Abdominal discomfort > Hypocount high (H/C:18) > CVP Low (CVP:1)

SCENARIO G – Tension Pneumothorax (with misdiagnosis of Cardiac Tamponade)

The patient is [REDACTED], a 20 year old man, a smoker and an army mechanic. He has a recent history of bronchopneumonia. He was admitted after a fall from a ladder at a height of 3 metres. You have been called to attend to him as he is complaining of drowsiness.

P/E	Tachycardia, Lung reduce air entry Right, occasional wheeze
ParaM	HR 100/min, BP 80/45mmHg, CVP 18 cmH2O, RR 28/min, SpO2 90% Temp 37.1°C
Inx	CXR: Right Rib fractures (not given), ECG – sinus tachycardia
Moulage	Bruising right chest

ADA Patient Info	ADA Key Cues
Name: [REDACTED] Age/Gender:20/Male Smoker (3pack-years); History of Bronchopneumonia; Admitted after fall from height 3m; Previous ECG - Sinus Tachycardia; C/O drowsiness;	> Trauma > Tachycardic (HR:100) > Hypotensive (BP:80/45) > CVP high (CVP:18) > ECG - Sinus tachycardia

CONFIDENTIAL

SCENARIO H – Severe Hyperkalemia (with misdiagnosis of Hypoglycemic Attack)

The patient is [REDACTED], a 78 year old, a non-smoker and a retired taxi driver. He has a poor nutrition status because of low appetite and low food intake due to having multiple carious teeth. He has undergone an operation to extract the teeth under GA and had fasted for 10 hours before the op. He was initially complaining of op site pain for which IV Morphine was ordered but he complained of more pain on administration by the ward nurse. You have been called to attend to him as he is complaining of nausea, fatigue and a tingling feeling.

P/E	Eyes half close pupils normal, Lungs clear
ParaM	HR 65 with ectopics/min, BP 85/35mmHg, CVP 5 cmH2O, RR 23/min, SPO2 95%, Temp 36.7 °C,
Inx	Serum K 6.5 (not given), ECG tall peak T waves with Ventricular ectopics (may have been shown on monitor),
Moulage	-

ADA Patient Info	ADA Key Cues
Name: [REDACTED] Age/Gender: 78/Male Non-Smoker; Poor nutrition status; Multiple carious teeth affecting food intake; Loss of appetite; Undergone extraction of teeth under GA; IV Morphine given for op site pain but C/O of more pain; Fasted 10hrs before op	> Prolonged fasting > Low food intake > Poor nutrition status

SCENARIO I – PRACTICE: Septic Shock

The patient is [REDACTED], a 61 year old man, and a cleaner. He has a past history of diabetes mellitus with peripheral vascular disease and ischemic heart disease. You have been called to attend to him as he is complaining of drowsiness, breathlessness and coughing.

P/E	Tachycardia, Hypotension, Low CVP, lungs clear
ParaM	HR 125/min, BP 80/35mmHg, CVP – 5cmH2O, RR 25/min, SpO2 91% ? Lungs clear, Temp 38°C, bounding pulse
Inx	TWBC 18K
Moulage	-

ADA Patient Info	ADA Key Cues
Name: [REDACTED] Age/Gender: 61/Male Non-Smoker; History of Diabetes Mellitus History of Peripheral Vascular Disease History of Ischaemic Heart Disease Previous TWC done: 18000 C/O breathlessness, drowsiness, coughing	> TWC high (TWC:18000) > Temp high (T:38) > Hypotensive (BP:80/35) > Tachycardic (HR:125)

CONFIDENTIAL

SCENARIO J – PRACTICE: Cerebrovascular Accident

The patient is [REDACTED], a 50 year old woman, and a housewife. She has a past history of hypertension, and diabetes mellitus. She was admitted after feeling increasingly drowsy and nauseous after returning from the market. You have been called to attend to her as she is feeling photophobic and experiencing impaired speech.

P/E	Photophobia, Unequal pupils
ParaM	HR 50/min, BP 170/100mmHg, CVP 5 cmH2O, RR 10/min, SpO2 96%, Temp 36.8°C
Inx	-
Moulage	-

ADA Patient Info	ADA Key Cues
Name: [REDACTED] Age/Gender: 50/Female History of Hypertension; History of Diabetes; Admitted for increasing drowsiness and nausea; C/O impaired speech and photophobia	> Hypertension > Diabetes > Photophobia > Impaired Speech > Bradycardic (HR:50) > Hypertensive (BP:170/100)

SCENARIO K – ORIENTATION: Patient Normal

Appendix IV – Training Notes for Participants

REFRESHER NOTES

Haemorrhagic shock

Hemorrhagic shock occurs when the on-going blood loss (hypovolemic state) overcomes the body's compensatory mechanisms e.g. tachycardia to result in hypotension and impaired organ perfusion. Patients become pale, drowsy and dyspneic. Urine output and hemoglobin levels are low and immediate management include large-bore IV cannulae, IV fluid resuscitation, Group and Cross Match and blood/blood product transfusion.

Septic shock

When an overwhelming infection spreads into the circulation, it causes severe peripheral vasodilatation, capillary leak and reduced myocardial contractility. Patients develop fever, chills and rigors, hypotension and altered mental state. Investigations should include blood culture to isolate the pathogen and initial management includes expanding intravascular volume, inotropic support and empiric antibiotic cover.

Congestive Cardiac Failure (CCF)

The failure of the heart to pump blood effectively can be due to myocardial damage, cardiomyopathy, myocarditis or cardiac valve problems. It can cause dyspnea, pitting pedal edema or mental confusion. CCF can be diagnosed with echocardiography or blood tests e.g. B-type Natriuretic Peptide. The CVP is high in patients with fluid overload and Frusemide is often used to cause diuresis. IV Inotropic support is sometimes needed.

Acute Myocardial Infarction (AMI)

AMI results from the occlusion of coronary arterial supply to the heart, leading to myocardial cell death. AMI could be "silent" or have symptoms e.g. chest pain (typically radiating to left arm or side of neck), dyspnea, nausea, palpitations or sweating. It can be divided into ST elevation or Non-ST elevation MI. Risk factors include Male gender, HPT, DM, hypercholesterolemia, smoking and stress. Tests include ECG and cardiac enzymes e.g. creatine kinase-MB, troponin levels. Immediate treatment includes oxygen, aspirin, sublingual nitroglycerin and pain relief.

Pulmonary Embolism (PE)

PE is potentially lethal and occurs when a main pulmonary artery or its branches are blocked. This is often due to a break-away thrombus from the periphery (e.g. deep veins of the calves), i.e. venous thromboembolism. Patients develop dyspnea, chest pain on inspiration and palpitations. Tests include ABG and imaging studies e.g. pulmonary angiography. Treatment may include anti-coagulants e.g. heparin, warfarin; thrombolysis with tPA or pulmonary thrombectomy.

Cerebrovascular Accident (CVA)

CVA occur when blood supply to the brain is disrupted via vessel blockage (thrombosis or embolism) or intracranial hemorrhage. Patients may present with hemiplegia/ paresis, impaired speech/ vision or confusion. Risk factors include HPT, DM, hypercholesterolemia, smoking and atrial fibrillation. Ischemic CVA may need thrombolysis and antiplatelet drugs e.g. aspirin, dipyridamole. Hemorrhagic CVA raises intracranial pressure and may require emergency neurosurgery.

Cardiac Tamponade

Cardiac tamponade is caused by fluid accumulating in the fibrous pericardial sac. This puts pressure on the heart, reducing ventricular filling and causing pulsus paradoxus. Beck's Triad: muffled heart sounds, distended jugular veins and hypotension. Causes include physical trauma, pericarditis or rupture of myocardium. ECG show generalized low-voltage "qrs" complexes and echocardiography is the diagnostic test of choice. Besides supportive therapy, initial management includes pericardiocentesis with subsequent emergency pericardial window surgery.

Tension Pneumothorax

The abnormal accumulation of air in the pleural cavity via a one-way valve mechanism can result in high pressure which displaces both the heart and lung. Patients develop dyspnea, hypotension and ipsilateral reduced breath sounds and hyper-resonance on percussion of the chest. There is often tachycardia and the trachea may be displaced contralaterally. Immediate therapy is needle decompression with subsequent chest tube insertion.

Exacerbation of Chronic Obstructive Pulmonary Disease (COPD)

Sudden worsening of COPD symptoms e.g. dyspnea, wheezing and increased phlegm may be triggered by an infection or pollutants in the air. The resulting airway inflammation causes further hyperinflation, reduced expiratory flow rate and worsen gas exchange. The underlying pathology may be chronic bronchitis or emphysema. Treatment includes inhaled bronchodilators e.g. beta-adrenoceptor agonists, anticholinergics, mucolytic agents, corticosteroids or low-dose theophylline.

Hypoglycemic Attack

Hypoglycemia can produce symptoms from dizziness, jitteriness to seizures, unconsciousness or brain damage. Causes include complication of DM treatment, inborn error of metabolism, alcohol, prolonged starvation and altered metabolic states. It is treated by oral glucose drink or IV dextrose; in severe cases, IV glucagon can also be used.

Diabetic Ketoacidosis (DKA)

DKA is life-threatening in DM patients when the body metabolizes fatty acids and produces acidic ketones. It may be triggered by poor compliance to DM therapy or intercurrent illness. Symptoms include vomiting, abdominal cramps, dehydration, Kussmaul breathing and confusion. Test results include high blood sugar levels, urinary ketones and metabolic acidosis. Treatment involves rehydration, insulin therapy and treatment for underlying causes such as infections.

Severe Hyperkalemia

This is a medical emergency as fatal arrhythmias may occur. Symptoms include lethargy, muscle weakness and palpitations. Acute renal failure and medications are often causes. ECG show tall T waves and arrhythmia. Treatment includes IV Calcium for myocardial protection; IV Insulin/dextrose, IV Bicarbonate or Nebulised Salbutamol to lower K⁺; chelating agents or hemodialysis to increase elimination.

APPENDIX V – Post Condition Questionnaire Items

Participants response on agreement scale anchored at 1: Strongly Disagree and 7: Strongly Agree

Perceived Understanding

I did not know how the automated diagnostic aid arrived at its recommendations. (Reverse-Scale)

I understood why the diagnostic aid gave the recommendations that it did.

Trust

The diagnostic aid's recommendations provided me with security.

I was confident in the diagnostic aid's recommendations.

I trusted the diagnostic aid to provide me with accurate recommendations.

Using the recommendations from the diagnostic aid would have a harmful outcome. (Reverse-Scale)

The recommendations from the diagnostic aid were reliable.

The recommendations from the diagnostic aid were deceptive. (Reverse-Scale)

Appendix VI - Experimental Results

Participant	Trial No.	Scenario ID	CUES	LIKE	Correct Diagnosis	Aid's Recommended Diagnosis	Participant's Diagnosis	Automation Non-Error/Error	Agreement	Accuracy	Diagnostic Confidence	Trust in Aid
P01	1	B	ON	OFF	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.0	5.5
P01	2	G	ON	OFF	Tension Pneumothorax	Cardiac Tamponade	Tension Pneumothorax	ERROR	DISAGREE	CORRECT	4.5	4.4
P01	3	E	ON	OFF	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	5.0	5.0
P01	4	A	ON	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.4	6.0
P01	5	C	ON	ON	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.0	5.7
P01	6	H	ON	ON	Severe Hyperkalemia	Hypoglycemic Attack	Severe Hyperkalemia	ERROR	DISAGREE	CORRECT	4.5	2.5
P01	7	D	ON	ON	Exacerbation of COPD	Exacerbation of COPD	Acute Myocardial Infarction	NON-ERROR	DISAGREE	WRONG	1.4	3.0
P01	8	F	ON	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	5.4	5.4
P02	1	A	ON	ON	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.8	6.3
P02	2	H	ON	ON	Severe Hyperkalemia	Hypoglycemic Attack	Acute Myocardial Infarction	ERROR	DISAGREE	WRONG	5.7	5.0
P02	3	D	ON	ON	Exacerbation of COPD	Exacerbation of COPD	Pulmonary Embolism	NON-ERROR	DISAGREE	WRONG	6.1	5.0
P02	4	E	ON	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	6.7	6.6
P02	5	C	ON	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.5	6.8
P02	6	G	ON	OFF	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	6.6	6.7
P02	7	F	ON	OFF	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.3	6.8
P02	8	B	ON	OFF	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.8	6.7
P03	1	A	OFF	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.5	6.5
P03	2	H	OFF	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	4.8	4.8
P03	3	B	OFF	OFF	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.3	6.1
P03	4	C	OFF	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	4.9	5.1
P03	5	D	OFF	ON	Exacerbation of COPD	Exacerbation of COPD	Pulmonary Embolism	NON-ERROR	DISAGREE	WRONG	5.5	2.8
P03	6	G	OFF	ON	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	5.3	5.9
P03	7	E	OFF	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	5.6	5.7
P03	8	F	OFF	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.6	6.6
P04	1	E	OFF	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	5.0	5.4
P04	2	H	OFF	ON	Severe Hyperkalemia	Hypoglycemic Attack	Severe Hyperkalemia	ERROR	DISAGREE	CORRECT	5.4	2.2
P04	3	F	OFF	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	5.8	5.9
P04	4	C	OFF	ON	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.0	6.0
P04	5	D	OFF	OFF	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	6.0	5.9
P04	6	G	OFF	OFF	Tension Pneumothorax	Cardiac Tamponade	Tension Pneumothorax	ERROR	DISAGREE	CORRECT	6.1	2.9

Appendix VI - Experimental Results

P04	7	A	OFF	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	5.9	6.2
P04	8	B	OFF	OFF	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.5	6.0
P05	1	B	OFF	OFF	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.0	6.0
P05	2	G	OFF	OFF	Tension Pneumothorax	Cardiac Tamponade	Tension Pneumothorax	ERROR	DISAGREE	CORRECT	5.0	4.0
P05	3	E	OFF	OFF	Hypoglycemic Attack	Hypoglycemic Attack	Diabetic Ketoacidosis	NON-ERROR	DISAGREE	WRONG	5.0	4.0
P05	4	C	OFF	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	5.5	5.5
P05	5	F	OFF	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	3.0	5.0
P05	6	H	OFF	ON	Severe Hyperkalemia	Hypoglycemic Attack	Pulmonary Embolism	ERROR	DISAGREE	WRONG	6.0	2.0
P05	7	A	OFF	ON	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.0	6.0
P05	8	D	OFF	ON	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	5.5	5.0
P06	1	F	ON	OFF	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	5.7	5.8
P06	2	H	ON	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	4.8	5.8
P06	3	B	ON	OFF	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.3	6.3
P06	4	C	ON	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.6	6.7
P06	5	D	ON	ON	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	5.6	5.0
P06	6	G	ON	ON	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	4.7	5.2
P06	7	E	ON	ON	Hypoglycemic Attack	Hypoglycemic Attack	Septic Shock	NON-ERROR	DISAGREE	WRONG	5.4	1.9
P06	8	A	ON	ON	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	5.1	5.1
P07	1	B	OFF	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.0	6.2
P07	2	G	OFF	ON	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	3.2	5.5
P07	3	D	OFF	ON	Exacerbation of COPD	Exacerbation of COPD	Pulmonary Embolism	NON-ERROR	DISAGREE	WRONG	5.1	4.4
P07	4	E	OFF	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	5.9	6.8
P07	5	C	OFF	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	5.4	5.6
P07	6	H	OFF	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	5.6	5.5
P07	7	A	OFF	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.3	5.4
P07	8	F	OFF	OFF	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	5.6	5.5
P08	1	F	ON	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.0	5.8
P08	2	G	ON	ON	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	6.0	6.0
P08	3	D	ON	ON	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	4.9	4.5
P08	4	E	ON	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	5.2	5.9
P08	5	C	ON	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.1	6.2

Appendix VI - Experimental Results

P08	6	H	ON	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	4.5	4.6
P08	7	A	ON	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.4	6.6
P08	8	B	ON	OFF	Congestive Cardiac Failure	Congestive Cardiac Failure	Pulmonary Embolism	NON-ERROR	DISAGREE	WRONG	4.7	3.1
P09	1	D	OFF	OFF	Exacerbation of COPD	Exacerbation of COPD	Pulmonary Embolism	NON-ERROR	DISAGREE	WRONG	6.9	3.7
P09	2	H	OFF	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Severe Hyperkalemia	ERROR	DISAGREE	CORRECT	6.1	2.2
P09	3	C	OFF	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.2	6.2
P09	4	A	OFF	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	4.5	4.6
P09	5	E	OFF	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	6.5	5.8
P09	6	G	OFF	ON	Tension Pneumothorax	Cardiac Tamponade	Tension Pneumothorax	ERROR	DISAGREE	CORRECT	6.0	3.7
P09	7	F	OFF	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.4	5.9
P09	8	B	OFF	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	7.0	6.2
P10	1	C	ON	ON	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.6	5.2
P10	2	G	ON	ON	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	5.0	5.0
P10	3	E	ON	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	5.9	5.4
P10	4	D	ON	ON	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	6.4	5.4
P10	5	A	ON	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.3	5.3
P10	6	H	ON	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	6.6	5.9
P10	7	B	ON	OFF	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	5.2	4.9
P10	8	F	ON	OFF	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.5	5.4
P11	1	C	ON	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	4.7	5.0
P11	2	H	ON	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	5.5	5.5
P11	3	E	ON	OFF	Hypoglycemic Attack	Hypoglycemic Attack	Septic Shock	NON-ERROR	DISAGREE	WRONG	4.5	3.0
P11	4	A	ON	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	5.6	5.1
P11	5	B	ON	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.0	5.1
P11	6	G	ON	ON	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	5.0	5.4
P11	7	D	ON	ON	Exacerbation of COPD	Exacerbation of COPD	Pulmonary Embolism	NON-ERROR	DISAGREE	WRONG	5.1	1.0
P11	8	F	ON	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.0	5.2
P12	1	F	OFF	OFF	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.9	7.0
P12	2	G	OFF	OFF	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	6.0	6.0
P12	3	D	OFF	OFF	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	6.0	6.0
P12	4	C	OFF	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.9	7.0

Appendix VI - Experimental Results

P12	5	A	OFF	ON	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.9	6.9
P12	6	H	OFF	ON	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	6.9	7.0
P12	7	E	OFF	ON	Hypoglycemic Attack	Hypoglycemic Attack	Septic Shock	NON-ERROR	DISAGREE	WRONG	6.0	4.0
P12	8	B	OFF	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.0	6.0
P13	1	E	OFF	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	3.8	4.7
P13	2	H	OFF	ON	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	6.4	6.4
P13	3	A	OFF	ON	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.6	6.6
P13	4	F	OFF	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.7	6.7
P13	5	B	OFF	OFF	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.8	6.9
P13	6	G	OFF	OFF	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	5.0	5.0
P13	7	C	OFF	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.3	5.8
P13	8	D	OFF	OFF	Exacerbation of COPD	Exacerbation of COPD	Pulmonary Embolism	NON-ERROR	DISAGREE	WRONG	6.0	3.5
P14	1	E	OFF	ON	Hypoglycemic Attack	Hypoglycemic Attack	Septic Shock	NON-ERROR	DISAGREE	WRONG	5.0	4.0
P14	2	G	OFF	ON	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	5.0	5.0
P14	3	F	OFF	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	5.9	5.5
P14	4	D	OFF	ON	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	5.0	5.4
P14	5	A	OFF	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	5.9	5.9
P14	6	H	OFF	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	6.0	6.0
P14	7	C	OFF	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	5.9	6.5
P14	8	B	OFF	OFF	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	5.0	5.0
P15	1	B	ON	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	5.7	5.6
P15	2	H	ON	ON	Severe Hyperkalemia	Hypoglycemic Attack	Diabetic Ketoacidosis	ERROR	DISAGREE	WRONG	5.2	3.3
P15	3	D	ON	ON	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	6.6	6.4
P15	4	F	ON	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.0	6.5
P15	5	E	ON	OFF	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	4.7	5.5
P15	6	G	ON	OFF	Tension Pneumothorax	Cardiac Tamponade	Tension Pneumothorax	ERROR	DISAGREE	CORRECT	6.5	2.4
P15	7	A	ON	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	5.6	6.8
P15	8	C	ON	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	7.0	6.8
P16	1	C	ON	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	5.9	5.9
P16	2	G	ON	OFF	Tension Pneumothorax	Cardiac Tamponade	Tension Pneumothorax	ERROR	DISAGREE	CORRECT	6.0	2.8
P16	3	A	ON	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.4	5.2

Appendix VI - Experimental Results

P16	4	D	ON	OFF	Exacerbation of COPD	Exacerbation of COPD	Cerebrovascular Accidents	NON-ERROR	DISAGREE	WRONG	5.0	3.0
P16	5	E	ON	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	5.9	5.0
P16	6	H	ON	ON	Severe Hyperkalemia	Hypoglycemic Attack	Severe Hyperkalemia	ERROR	DISAGREE	CORRECT	4.9	3.2
P16	7	B	ON	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	5.4	4.8
P16	8	F	ON	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.0	5.6
P17	1	E	ON	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	6.0	6.0
P17	2	H	ON	ON	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	4.7	4.3
P17	3	C	ON	ON	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	5.2	3.3
P17	4	B	ON	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Pulmonary Embolism	NON-ERROR	DISAGREE	WRONG	4.0	5.2
P17	5	D	ON	OFF	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	5.5	5.5
P17	6	G	ON	OFF	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	4.0	3.3
P17	7	A	ON	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	5.5	4.0
P17	8	F	ON	OFF	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	4.5	3.5
P18	1	A	OFF	ON	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	4.9	4.6
P18	2	G	OFF	ON	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	4.6	5.0
P18	3	C	OFF	ON	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	5.9	4.9
P18	4	E	OFF	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	5.9	5.0
P18	5	B	OFF	OFF	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	5.9	5.6
P18	6	H	OFF	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	5.5	5.5
P18	7	D	OFF	OFF	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	5.0	5.5
P18	8	F	OFF	OFF	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	5.0	5.9
P19	1	A	OFF	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	5.0	5.1
P19	2	H	OFF	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	5.9	5.9
P19	3	D	OFF	OFF	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	6.0	5.1
P19	4	C	OFF	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.5	5.5
P19	5	E	OFF	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	5.1	5.4
P19	6	G	OFF	ON	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	6.7	6.4
P19	7	F	OFF	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.9	6.4
P19	8	B	OFF	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.8	6.4
P20	1	C	OFF	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	5.3	4.8
P20	2	G	OFF	OFF	Tension Pneumothorax	Cardiac Tamponade	Haemorrhagic Shock	ERROR	DISAGREE	WRONG	5.5	5.0

Appendix VI - Experimental Results

P20	3	A	OFF	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.0	6.0
P20	4	B	OFF	OFF	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.0	6.0
P20	5	D	OFF	ON	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	5.5	5.0
P20	6	H	OFF	ON	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	5.0	4.0
P20	7	F	OFF	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	5.5	5.0
P20	8	E	OFF	ON	Hypoglycemic Attack	Hypoglycemic Attack	Septic Shock	NON-ERROR	DISAGREE	WRONG	5.5	3.5
P21	1	B	ON	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.2	3.4
P21	2	G	ON	ON	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	6.6	5.1
P21	3	A	ON	ON	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.8	4.4
P21	4	F	ON	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	5.9	5.4
P21	5	C	ON	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	7.0	6.0
P21	6	H	ON	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	5.7	5.5
P21	7	E	ON	OFF	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	6.2	5.1
P21	8	D	ON	OFF	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	6.7	5.1
P22	1	B	ON	OFF	Congestive Cardiac Failure	Congestive Cardiac Failure	Septic Shock	NON-ERROR	DISAGREE	WRONG	5.1	3.0
P22	2	G	ON	OFF	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	5.1	5.1
P22	3	C	ON	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	5.0	5.0
P22	4	A	ON	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.2	6.0
P22	5	F	ON	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	5.5	5.5
P22	6	H	ON	ON	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	4.1	4.4
P22	7	E	ON	ON	Hypoglycemic Attack	Hypoglycemic Attack	Septic Shock	NON-ERROR	DISAGREE	WRONG	5.7	4.1
P22	8	D	ON	ON	Exacerbation of COPD	Exacerbation of COPD	Pulmonary Embolism	NON-ERROR	DISAGREE	WRONG	5.5	3.4
P23	1	C	ON	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	7.0	6.5
P23	2	H	ON	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	6.0	6.0
P23	3	F	ON	OFF	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.0	6.0
P23	4	A	ON	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	7.0	6.4
P23	5	B	ON	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	7.0	6.0
P23	6	G	ON	ON	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	7.0	7.0
P23	7	D	ON	ON	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	5.0	5.0
P23	8	E	ON	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	5.2	6.0
P24	1	F	OFF	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	5.8	5.9

Appendix VI - Experimental Results

P24	2	H	OFF	ON	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	5.1	5.5
P24	3	E	OFF	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	5.5	5.2
P24	4	B	OFF	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	5.7	5.4
P24	5	A	OFF	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	5.9	5.5
P24	6	G	OFF	OFF	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	5.0	5.3
P24	7	C	OFF	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	5.9	5.2
P24	8	D	OFF	OFF	Exacerbation of COPD	Exacerbation of COPD	Pulmonary Embolism	NON-ERROR	DISAGREE	WRONG	5.4	4.0
P25	1	E	ON	OFF	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	5.9	6.4
P25	2	G	ON	OFF	Tension Pneumothorax	Cardiac Tamponade	Tension Pneumothorax	ERROR	DISAGREE	CORRECT	5.0	3.0
P25	3	A	ON	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.0	5.9
P25	4	C	ON	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.6	6.1
P25	5	F	ON	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.0	6.1
P25	6	H	ON	ON	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	5.0	5.1
P25	7	B	ON	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.0	5.1
P25	8	D	ON	ON	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	4.9	5.9
P26	1	F	OFF	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	5.2	5.0
P26	2	G	OFF	ON	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	5.0	5.1
P26	3	B	OFF	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	4.0	5.0
P26	4	E	OFF	ON	Hypoglycemic Attack	Hypoglycemic Attack	Septic Shock	NON-ERROR	DISAGREE	WRONG	5.0	3.0
P26	5	C	OFF	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	5.9	5.1
P26	6	H	OFF	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Severe Hyperkalemia	ERROR	DISAGREE	CORRECT	4.9	3.2
P26	7	A	OFF	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	5.6	4.9
P26	8	D	OFF	OFF	Exacerbation of COPD	Exacerbation of COPD	Pulmonary Embolism	NON-ERROR	DISAGREE	WRONG	5.2	2.5
P27	1	F	OFF	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.3	5.9
P27	2	H	OFF	ON	Severe Hyperkalemia	Hypoglycemic Attack	Severe Hyperkalemia	ERROR	DISAGREE	CORRECT	6.5	3.4
P27	3	E	OFF	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	6.4	6.4
P27	4	B	OFF	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.8	6.0
P27	5	D	OFF	OFF	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	6.1	6.4
P27	6	G	OFF	OFF	Tension Pneumothorax	Cardiac Tamponade	Tension Pneumothorax	ERROR	DISAGREE	CORRECT	6.3	3.0
P27	7	A	OFF	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.8	6.4
P27	8	C	OFF	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.5	6.0

Appendix VI - Experimental Results

P28	1	B	OFF	OFF	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.0	4.9
P28	2	G	OFF	OFF	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	5.0	4.9
P28	3	D	OFF	OFF	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	4.7	3.4
P28	4	F	OFF	OFF	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.8	4.7
P28	5	E	OFF	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	6.1	4.4
P28	6	H	OFF	ON	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	6.0	4.4
P28	7	C	OFF	ON	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.0	3.9
P28	8	A	OFF	ON	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	5.6	4.8
P29	1	E	ON	OFF	Hypoglycemic Attack	Hypoglycemic Attack	Diabetic Ketoacidosis	NON-ERROR	DISAGREE	WRONG	4.0	3.1
P29	2	H	ON	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	4.0	4.0
P29	3	F	ON	OFF	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	4.0	3.9
P29	4	A	ON	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	4.0	4.0
P29	5	D	ON	ON	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	4.0	4.0
P29	6	G	ON	ON	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	3.0	5.0
P29	7	B	ON	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	4.0	5.0
P29	8	C	ON	ON	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	5.0	6.0
P30	1	E	ON	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	5.5	6.0
P30	2	H	ON	ON	Severe Hyperkalemia	Hypoglycemic Attack	Cardiac Tamponade	ERROR	DISAGREE	WRONG	5.9	4.0
P30	3	F	ON	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.0	6.3
P30	4	B	ON	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.3	6.1
P30	5	C	ON	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.0	5.9
P30	6	G	ON	OFF	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	5.9	6.0
P30	7	A	ON	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.0	5.9
P30	8	D	ON	OFF	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	5.3	5.9
P31	1	B	ON	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	5.4	4.4
P31	2	G	ON	ON	Tension Pneumothorax	Cardiac Tamponade	Cerebrovascular Accidents	ERROR	DISAGREE	WRONG	6.0	5.1
P31	3	D	ON	ON	Exacerbation of COPD	Exacerbation of COPD	Pulmonary Embolism	NON-ERROR	DISAGREE	WRONG	6.0	3.0
P31	4	A	ON	ON	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.0	6.1
P31	5	C	ON	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.5	6.0
P31	6	H	ON	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	6.0	6.0
P31	7	E	ON	OFF	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	6.0	6.0
P31	8	F	ON	OFF	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.9	6.0

Appendix VI - Experimental Results

P32	1	A	OFF	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.6	5.0
P32	2	H	OFF	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	6.2	5.1
P32	3	C	OFF	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Cardiac Tamponade	NON-ERROR	DISAGREE	WRONG	5.0	2.0
P32	4	D	OFF	OFF	Exacerbation of COPD	Exacerbation of COPD	Pulmonary Embolism	NON-ERROR	DISAGREE	WRONG	5.9	3.5
P32	5	E	OFF	ON	Hypoglycemic Attack	Hypoglycemic Attack	Septic Shock	NON-ERROR	DISAGREE	WRONG	5.6	3.0
P32	6	G	OFF	ON	Tension Pneumothorax	Cardiac Tamponade	Tension Pneumothorax	ERROR	DISAGREE	CORRECT	5.0	2.5
P32	7	B	OFF	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.8	6.2
P32	8	F	OFF	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.0	6.5
P33	1	F	ON	OFF	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	5.2	4.8
P33	2	G	ON	OFF	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	5.5	5.6
P33	3	D	ON	OFF	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	5.8	4.9
P33	4	C	ON	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.2	5.6
P33	5	B	ON	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.2	5.1
P33	6	H	ON	ON	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	4.8	5.6
P33	7	E	ON	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	5.0	4.7
P33	8	A	ON	ON	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	5.5	5.5
P34	1	B	OFF	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	5.4	5.3
P34	2	G	OFF	ON	Tension Pneumothorax	Cardiac Tamponade	Tension Pneumothorax	ERROR	DISAGREE	CORRECT	5.1	3.1
P34	3	A	OFF	ON	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.3	5.4
P34	4	D	OFF	ON	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	5.3	5.2
P34	5	C	OFF	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.3	5.4
P34	6	H	OFF	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	4.6	5.0
P34	7	E	OFF	OFF	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	6.2	5.9
P34	8	F	OFF	OFF	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.3	6.3
P35	1	D	OFF	OFF	Exacerbation of COPD	Exacerbation of COPD	Pulmonary Embolism	NON-ERROR	DISAGREE	WRONG	5.5	4.0
P35	2	G	OFF	OFF	Tension Pneumothorax	Cardiac Tamponade	Tension Pneumothorax	ERROR	DISAGREE	CORRECT	6.0	2.6
P35	3	E	OFF	OFF	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	6.9	4.9
P35	4	A	OFF	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	7.0	6.0
P35	5	B	OFF	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.5	5.7
P35	6	H	OFF	ON	Severe Hyperkalemia	Hypoglycemic Attack	Acute Myocardial Infarction	ERROR	DISAGREE	WRONG	4.1	4.3
P35	7	C	OFF	ON	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	3.5	4.5

Appendix VI - Experimental Results

P35	8	F	OFF	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Septic Shock	NON-ERROR	DISAGREE	WRONG	5.0	3.8
P36	1	D	ON	ON	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	6.0	5.5
P36	2	G	ON	ON	Tension Pneumothorax	Cardiac Tamponade	Tension Pneumothorax	ERROR	DISAGREE	CORRECT	6.0	3.2
P36	3	B	ON	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Diabetic Ketoacidosis	NON-ERROR	DISAGREE	WRONG	5.4	4.0
P36	4	F	ON	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	4.9	5.7
P36	5	C	ON	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.4	6.5
P36	6	H	ON	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	6.1	6.7
P36	7	A	ON	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	5.9	5.0
P36	8	E	ON	OFF	Hypoglycemic Attack	Hypoglycemic Attack	Septic Shock	NON-ERROR	DISAGREE	WRONG	6.2	2.5
P37	1	A	ON	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	5.9	5.9
P37	2	H	ON	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	6.0	6.0
P37	3	C	ON	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.3	6.3
P37	4	F	ON	OFF	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	5.8	5.9
P37	5	E	ON	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	6.3	6.2
P37	6	G	ON	ON	Tension Pneumothorax	Cardiac Tamponade	Tension Pneumothorax	ERROR	DISAGREE	CORRECT	6.2	3.4
P37	7	B	ON	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	4.9	4.8
P37	8	D	ON	ON	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	5.9	5.0
P38	1	D	OFF	OFF	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	5.0	6.0
P38	2	H	OFF	OFF	Severe Hyperkalemia	Hypoglycemic Attack	Severe Hyperkalemia	ERROR	DISAGREE	CORRECT	6.0	5.0
P38	3	A	OFF	OFF	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.0	6.0
P38	4	B	OFF	OFF	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.0	6.0
P38	5	C	OFF	ON	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.0	6.0
P38	6	G	OFF	ON	Tension Pneumothorax	Cardiac Tamponade	Tension Pneumothorax	ERROR	DISAGREE	CORRECT	6.4	3.0
P38	7	F	OFF	ON	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	5.0	6.0
P38	8	E	OFF	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	6.0	6.5
P39	1	B	OFF	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.6	6.4
P39	2	H	OFF	ON	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	6.0	6.4
P39	3	C	OFF	ON	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	6.4	5.8
P39	4	A	OFF	ON	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.7	6.1
P39	5	D	OFF	OFF	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	5.6	5.0
P39	6	G	OFF	OFF	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	6.6	6.6

Appendix VI - Experimental Results

P39	7	F	OFF	OFF	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.0	5.6
P39	8	E	OFF	OFF	Hypoglycemic Attack	Hypoglycemic Attack	Septic Shock	NON-ERROR	DISAGREE	WRONG	5.7	5.0
P40	1	A	ON	ON	Haemorrhagic Shock	Haemorrhagic Shock	Haemorrhagic Shock	NON-ERROR	AGREE	CORRECT	6.4	6.5
P40	2	H	ON	ON	Severe Hyperkalemia	Hypoglycemic Attack	Hypoglycemic Attack	ERROR	AGREE	WRONG	5.7	6.0
P40	3	E	ON	ON	Hypoglycemic Attack	Hypoglycemic Attack	Hypoglycemic Attack	NON-ERROR	AGREE	CORRECT	5.6	6.0
P40	4	B	ON	ON	Congestive Cardiac Failure	Congestive Cardiac Failure	Congestive Cardiac Failure	NON-ERROR	AGREE	CORRECT	6.6	6.3
P40	5	F	ON	OFF	Diabetic Ketoacidosis	Diabetic Ketoacidosis	Diabetic Ketoacidosis	NON-ERROR	AGREE	CORRECT	6.6	6.4
P40	6	G	ON	OFF	Tension Pneumothorax	Cardiac Tamponade	Cardiac Tamponade	ERROR	AGREE	WRONG	5.3	6.0
P40	7	D	ON	OFF	Exacerbation of COPD	Exacerbation of COPD	Exacerbation of COPD	NON-ERROR	AGREE	CORRECT	5.5	6.0
P40	8	C	ON	OFF	Acute Myocardial Infarction	Acute Myocardial Infarction	Acute Myocardial Infarction	NON-ERROR	AGREE	CORRECT	5.4	5.5

NUS INSTITUTIONAL REVIEW BOARD (IRB) APPROVAL CERTIFICATE

A) Protocol Title: Study on the Use of a Decision Aid during Medical Diagnosis in Time Critical Scenario

Principal Investigator: Assoc Professor Siau Chiang

Department: Anaesthesia

Institution: National University of Singapore

Co-Investigators: A/Prof Chen Fun-Gee, Dr Park Tazeeoon, Mr Mohamed Syahid Bin Hassan

Sponsor (if applicable): -N.A.-

Research Site(s): Dept of Anaesthesia
Yong Loo Lin School of Medicine
National University of Singapore
Centre for Healthcare Simulation
National University of Singapore

B) Documents Reviewed

<u>Documents</u>	<u>Document Date</u>
1. NUS-IRB Application Form	Version 4, 3 May 2012
2. Email Invitation	Version 2, 3 May 2012
3. Participant Information Sheet and Consent Form	Version 4, 3 May 2012
4. Questionnaires	Version 3, 17 Apr 2012
5. Scenario Design	Rec'd on 18 Apr 2012
6. Debriefing Document	Version 1, 19 Apr 2012
7. Monetary reimbursement acknowledgement	Version 1, 1 Apr 2012
8. Investigators' Curricula Vitae	--

C) The above-mentioned documents have been reviewed and have been approved on 15 May 2012. The Board is organized and operated according to GCP guidelines, BAC guidelines and the applicable laws and regulations of Singapore.

D) Please note that:

- (1) No subject should be admitted to the trial before MCRC issues the certificate for the trial (applicable for drug trials only).
- (2) This approval shall remain valid until the completion of the research or notification of termination of the research, whichever is earlier.

- (3) Approval will be withdrawn if there is non-compliance by the Principal Investigator to the regulation on:
- (a) reporting of serious adverse events ("SAEs") on patients in Singapore within the specific time frame;
 - (b) submission of the annual report to the NUS-IRB within the specified time frame.
- (4) No deviation from, or changes of, the protocol should be initiated without prior written NUS-IRB approval of an appropriate amendment, except when necessary to eliminate immediate hazards to the subjects or when the change(s) involve(s) only logistical or administrative aspects of the trial [e.g. change of monitor(s), telephone number(s)].
- (5) The Principal Investigator should promptly inform the NUS-IRB of :
- (a) Deviations from, or changes of, the protocol to eliminate immediate hazards to the trial subjects;
 - (b) Changes increasing the risk to subjects and/or affecting significantly the conduct of the trial;
 - (c) All adverse events and adverse drug reactions (ADRs) that are both serious and unexpected;
 - (d) New information that may affect adversely the safety of the subjects or the conduct of the trial;
 - (e) The completion of the research.
- (6) The Principal Investigator should provide NUS-IRB with a copy of the final summary or final report of the research within 3 months after the completion of the research. The first continuing review report will be due on 31 December 2012. Please use the attached Continuing Review Form. Please note that failure to submit the Continuing Report for the research may result in the IRB's termination of its approval for your research.

E) Serious Adverse Events / Adverse Events reporting

Reports of local or overseas serious adverse events (including Medwatch reports or equivalent) must be accompanied by PI's analysis/evaluation of these events / reports. For local serious adverse event (including Singapore multi-centered trials), please use the attached format in Report of Local SAE Form.

F) Signature of NUS-IRB Chairman:

Professor Lee Hin Peng
NUS Institutional Review Board
Clinical Research Centre
Blk MD 11, #03-02
10 Medical Drive
Singapore 117597
Tel: 6516 4311 / 5453
Fax: 6778 3430

15 May 2012
Date