

City University London  
MSc in Human-Centred Systems  
Project Report  
2015

## **“This thing knows more than me.”**

Understanding how clinicians make decisions using a CDSS

Adrian M. Bussone  
Supervised by: Dr. Simone Stumpf  
09 January 2015

*By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.*

*Signed:*

Adrian M. Bussone

**Abstract:**

Clinical decision support systems (CDSS) have existed for over four decades and are acknowledged as being largely beneficial in regards to quality of care and clinical decision-making. CDSSs are designed for a vast array of purposes, from preventing overdose to suggesting diagnoses. However, these systems often fail to be truly effective due to their frequent *misuse* due to automation bias or *disuse* due to poor intelligibility. System explanations have been suggested as a solution in other domains.

This report presents an experimental research study that takes an empirical look at the use of explanations within a CDSS, exploring the effect of completeness on a user's decision-making, trust, confidence, and workload. Additionally, the information used and desired by clinicians during a decision-making task was analysed. The results of this study suggest that explanations with high completeness have positive effects on intelligibility, trust, and confidence. They also appear to have little impact on workload, yet may be unsuccessful in preventing automation bias. Finally, four new explanation types for CDSSs are identified: Certainty Details, More Than a Match, Disease Details, and Differential Diagnosis. The implications of these findings are discussed in context of related literature and the need for additional research.

**Keywords:** *Decision making, Automation Bias, Explanations, Medical Systems, Decision Support*

## TABLE OF CONTENTS

<b>Chapter 1 Introduction and Objectives .....</b>	<b>8</b>
Section 1.1 Problem Description.....	8
Section 1.2 Research Questions and Objectives .....	10
Section 1.3 Products of the Work and Beneficiaries.....	10
Section 1.4 Project Scope .....	11
Section 1.5 Outline of Methods .....	11
Section 1.6 Report Structure .....	12
<b>Chapter 2 Critical Context.....</b>	<b>13</b>
Section 2.1 Overview.....	13
Section 2.2 Intelligent Systems: Automation and Decision Support.....	13
Section 2.3 Clinical Decision Support Systems .....	14
Section 2.4 CDSS Reliability and Automation Bias .....	14
2.4.1 Automation bias .....	15
Section 2.5 Intelligibility through System Explanations .....	16
2.5.1 Intelligibility .....	16
2.5.2 System Explanations .....	16
2.5.3 Explanation Completeness .....	18
Section 2.6 Critical Context Summary .....	18
<b>Chapter 3 Methods.....</b>	<b>19</b>
Section 3.1 Overview .....	19
Section 3.2 Research Strategy .....	20
3.2.1 Between-Group Experiment .....	20
3.2.2 Observing through Simulated Use .....	20
3.2.3 Ethical Considerations and Risks .....	21
Section 3.3 Participants .....	21
3.3.1 Target Population .....	21
3.3.2 Recruitment and Sampling Strategy .....	21
3.3.3 Screening with Inclusion and Exclusion Criteria .....	22
3.3.4 Recruited Participants and Group Assignment.....	23
Section 3.4 Study Session Outline.....	24
3.4.1 Session Locations.....	24
3.4.2 Materials and Equipment used .....	24
3.4.3 Informed Consent .....	26
3.4.4 Confidentiality .....	27
3.4.5 System Description .....	27
3.4.6 Workload Introduction .....	27

3.4.7	Pre-Use Questionnaire .....	28
3.4.8	Decision-Making Tasks.....	28
3.4.9	NASA TLX .....	34
3.4.10	Post-Use Questionnaire.....	34
3.4.11	Incentive for Participation .....	35
3.4.12	Completion.....	35
<b>Section 3.5</b>	<b>Data Collection and Analysis .....</b>	<b>36</b>
3.5.1	RQ1: How does explanation completeness impact clinical decision-making? .....	36
3.5.2	RQ2: How does explanation completeness impact a clinician's confidence in their ability to diagnose patients? .....	37
3.5.3	RQ3: How does explanation completeness affect a clinical user's trust in a CDSS? .....	38
3.5.4	RQ4: What is the impact of explanation completeness on a clinical user's workload? .....	38
3.5.5	RQ5: What information do clinician's desire from a CDSS's explanation when making a diagnostic decision? .....	39
<b>Chapter 4</b>	<b>Results .....</b>	<b>42</b>
Section 4.1	<b>Introduction to Results .....</b>	<b>42</b>
Section 4.2	<b>RQ1: How does explanation completeness impact clinical decision-making? .....</b>	<b>42</b>
4.2.1	Impact of Completeness on making 'Right' and 'Wrong' decisions .....	43
4.2.2	Impact of Completeness on decision to accept or decline.....	44
Section 4.3	<b>RQ2: How does explanation completeness impact a clinician's confidence in their ability to diagnose patients? .....</b>	<b>46</b>
4.3.1	Changes in Confidence .....	46
4.3.2	Confidence and Explanations of High-Completeness: .....	47
4.3.3	Confidence and Explanations of Low-Completeness: .....	48
Section 4.4	<b>RQ3: How does explanation completeness affect a clinical user's trust in a CDSS?.....</b>	<b>49</b>
4.4.1	Changes in Trust .....	49
4.4.2	Trust in a CDSS with Explanations of High-Completeness: .....	50
4.4.3	Trust in a CDSS with Explanations of low-completeness .....	52
Section 4.5	<b>RQ4: What is the impact of completeness on a clinical user's workload? 54</b>	
4.5.1	Impact of Explanation Completeness on Workload .....	54
4.5.2	Analysis of Tasks Associated with Lowest Overall Workload .....	57
4.5.3	Analysis of Tasks Related to Highest Overall Workload.....	59
Section 4.6	<b>RQ5: What information do clinicians desire from a CDSS's explanation when making a diagnostic decision? .....</b>	<b>62</b>

4.6.1	Explanation Completeness, the Necessary Patient Information .....	62
4.6.2	Additional Explanations Desired .....	70
<b>Chapter 5</b>	<b>Discussion .....</b>	<b>75</b>
Section 5.1	<b>The Impact of Completeness on Decision-Making .....</b>	<b>75</b>
Section 5.2	<b>The Impact of Completeness on Confidence and Trust .....</b>	<b>76</b>
5.2.1	The Impact of Explanations of Low Completeness on Confidence and Trust .....	76
5.2.2	The Impact of Explanations of High Completeness on Confidence and Trust .....	77
Section 5.3	<b>The Impact of Completeness on a Clinical User's Workload ....</b>	<b>78</b>
Section 5.4	<b>Information Used and Desired .....</b>	<b>78</b>
5.4.1	Patient Information Wanted in an Explanation .....	78
5.4.2	Additional Explanations Desired .....	79
Section 5.5	<b>Future Work .....</b>	<b>82</b>
<b>Chapter 6</b>	<b>Evaluation, Reflections, and Conclusion .....</b>	<b>84</b>
Section 6.1	<b>Review of Objectives .....</b>	<b>84</b>
Section 6.2	<b>Literature examined .....</b>	<b>84</b>
Section 6.3	<b>Planning and Method .....</b>	<b>85</b>
6.3.1	Changes to Goals and Methods .....	85
6.3.2	Case study validation .....	86
6.3.3	Troubles with Recruitment .....	86
6.3.4	Balanced groups .....	87
6.3.5	Data Analysis .....	87
Section 6.4	<b>Conclusions .....</b>	<b>87</b>
<b>Chapter 7</b>	<b>References .....</b>	<b>Error! Bookmark not defined.</b>

## FIGURES

FIGURE 1: LOG-IN SCREEN.....	29
FIGURE 2: CREATE NEW PATIENT RECORD .....	30
FIGURE 3: MEDICAL HISTORY WITH EXAMPLE OF INPUT SCREEN .....	30
FIGURE 4: SYMPTOMS SCREEN.....	31
FIGURE 5: EXAMINATION SCREEN .....	32
FIGURE 6: PHOTOGRAPH OF PAPER CASE STUDIES BEING USED DURING A STUDY SESSION .....	32
FIGURE 7: SUGGESTED DIAGNOSIS WITH EXPLANATION OF LOW COMPLETENESS .....	33
FIGURE 8: SUGGESTED DIAGNOSIS WITH EXPLANATION OF HIGH COMPLETENESS .....	33
FIGURE 9: NASA TLX RATING SCALE (NASA HUMAN PERFORMANCE RESEARCH GROUP, 1987) .	34
FIGURE 10: PORTION OF DECISION-MAKING SPREADSHEET.....	37
FIGURE 11: EXAMPLE SECTION OF WORKLOAD RATING SPREADSHEET .....	39
FIGURE 12: EXAMPLE OF TRANSCRIPTS .....	40
FIGURE 13: PERCENT OF RIGHT AND WRONG DECISIONS, BY GROUP .....	43
FIGURE 14: PERCENTAGE OF CDSS SUGGESTIONS ACCEPTED OR DECLINED .....	44
FIGURE 15: PRE-USE AND POST-USE RATINGS OF CONFIDENCE .....	46
FIGURE 16: PRE-USE AND POST-USE RATINGS OF TRUST.....	50
FIGURE 17: MENTAL DEMAND OF HC GROUP AND LC GROUP .....	56
FIGURE 18: PHYSICAL DEMAND OF HC GROUP AND LC GROUP.....	56
FIGURE 19: TEMPORAL DEMAND OF HC GROUP AND LC GROUP .....	56
FIGURE 20: PERFORMANCE OF HC GROUP AND LC GROUP .....	56
FIGURE 21: EFFORT OF HC GROUP AND LC GROUP .....	56
FIGURE 22: FRUSTRATION OF HC GROUP AND LC GROUP .....	56
FIGURE 23: WORKLOAD OF HC GROUP AND LC GROUP.....	57
FIGURE 24: PERCENTAGE OF DECISION-MAKING TASKS IN WHICH MEDICAL HISTORY, SYMPTOMS, AND/OR EXAMINATIONS WERE MENTIONED .....	63
FIGURE 25: FREQUENCY OF REFERENCES SHOWN/NOT SHOWN, BY GROUP .....	64
FIGURE 26: FREQUENCY OF ADDITIONAL INFORMATION TYPES REQUESTED/MENTIONED .....	71

## TABLES

TABLE 1: DETAILS OF RECRUITED PARTICIPANTS .....	23
TABLE 2: BREAKDOWN OF CASE STUDY MANIPULATION (HIGH AND LOW CERTAINTY, TRUE AND FALSE POSITIVES) .....	26
TABLE 3: COUNTERING FOR ORDERING EFFECTS; THE ORDER IN WHICH THE CASE STUDIES WERE RECEIVED.....	29
TABLE 4: NUMBER OF CORRECT AND INCORRECT DIAGNOSTIC SUGGESTIONS, PER GROUP .....	42
TABLE 5: 'RIGHT' AND 'WRONG' DECISION FREQUENCIES, PER GROUP.....	43
TABLE 6: PROPORTION OF 'RIGHT' AND 'WRONG' DECISIONS.....	45
TABLE 7: CHANGE IN CONFIDENCE AND IMPACT OF EXPLANATIONS .....	47
TABLE 8: CHANGE IN TRUST AND IMPACT OF EXPLANATIONS.....	50
TABLE 9: MEAN AND STANDARD DEVIATIONS OF WORKLOAD SOURCES AND WORKLOAD .....	54
TABLE 10: DETAILS OF LOWEST OVERALL WORKLOAD, BY PARTICIPANT .....	58
TABLE 11: DETAILS OF HIGHEST OVERALL WORKLOAD, BY PARTICIPANT .....	60
TABLE 12: FREQUENCY OF MENTIONS TO VARIOUS MEDICAL HISTORY INFORMATION.....	65



# Chapter 1 Introduction and Objectives

## Section 1.1 Problem Description

Computer systems have evolved a long way from the now seemingly simple technological tools of the past. What once was the theme of futuristic science-fiction films has become reality - intelligent and automated computer systems now exist and are capable of 'thinking', 'acting', and 'suggesting' on their own. These intelligent systems are designed with the intent of aiding the human user, whether by automatically performing tasks for them or by supporting decision-making tasks by providing suggestions based on algorithmic reasoning. Because the fully- or semi-automated systems have the capability to respond to a user's needs in real-time, they are commonly applied to safety-critical domains such as aviation, manufacturing, and medicine (Parasuraman & Riley, 1997).

Decision support systems, an example of a semi-automated technology, have been used within the medical domain for over forty years. A decision support system receives data and applies a series of rules or reasoning processes to determine a logical inference, which is then suggested to the user (Seong & Bisantz, 2008). Across the healthcare field, clinical decision support systems (CDSS) are used for a variety of purposes; alerts or reminders, therapeutic planning, medication decision-making, and diagnostics are just a few examples (Sambasivan et al., 2012). It is widely believed that the use of CDSSs bring improvements to quality of care and clinical decision-making (Kong et al., 2008).

Whilst clinical decision support systems and other intelligent technologies are designed with the intention of reducing human error, they frequently have to reason around uncertainty. Thus, these systems are not perfectly reliable; they make suggestions or perform actions that are incorrect. However, users are not always sensitive to the reliability of automated systems (Wiegmann et al., 2001) and often believe them to be more reliable than themselves or other humans (Lee & See, 2004) (Madhavan & Wiegmann, 2007). Additionally, users tend to rely more on an automated system in times of high stress or high workload (Parasuraman & Riley, 1997). As a result, users of automated and intelligent technology are prone to *automation bias*, an over-reliance on the system (Alberdi et al., 2005).

The negative effects of automation bias can be seen in the user's judgment and decision-making. A user with bias towards the automated system will *misuse* the system, accepting the actions or suggestions of the system with little or no critical thinking or monitoring (Parasuraman & Riley, 1997). Automation bias is of particular concern when considering clinical decision

support systems; the medical domain is rife with uncertainty and the suggestions made by a CDSS are not immune to this (Kong et al., 2008). As such, a clinical user with bias towards the CDSS is likely to accept and act upon incorrect suggestions made by the system, resulting in potentially dangerous outcomes.

In order to mitigate the effects of automation bias, users must be able to ascertain whether or not the system has erred. Researchers have proposed system explanations to be a means to an end (Lim & Dey, 2010).

Explanations are designed to inform the user of the system's behaviour – why the system performed *x*, how the system came to suggest *y*, etc. This information is intended to help the user understand how the system works and when it has failed. That is, explanations are intended to make the system's inner reasoning processes transparent, or intelligible, to the user.

Numerous types of explanations have been proposed for various system-types, most famously for context-aware applications (Lee & See, 2004). One type, *Certainty*, addresses the issue of reliability explicitly; *Certainty* explanations display to the user how certain, or confident, the system is of the output (Lee & See, 2004). With this, a user is able to better gauge whether or not a suggestion should be accepted. Another explanation type, *Why...*, informs users *why* the system suggested *x*, grounding the reason for the suggestion in the information known to the system.

Explanations, particularly *Why...* explanations, can be more or less complete. That is, a *Why...* explanation can provide *all* of the reasons a suggestion was made (resulting in a complete explanation) or only *some* (resulting in an explanation of low-completeness) (Kulesza et al., 2013). The optimal level of completeness is not prescribed; it is dependent on the domain and the needs of the users. However, there are pros and cons that result from the use of a more complete explanation or a less complete one. A more complete explanation provides more intelligibility and improves the trust a user has with the system. A less complete explanation is less transparent, lowering trust in the system, but yet requiring less time for reading on the user's behalf. Additionally, research regarding the impact of information and decision-making is conflicting; some authors suggest that more information leads to better decisions (Dzindolet et al., 2003) whilst others argue that it leads to worse ones (Alberdi et al., 2009). Despite the prevalence of research on explanation-use in other intelligent systems, there remains a dearth of knowledge regarding the impact of explanations on decisions made with a DSS (Gonul et al., 2006).

## Section 1.2 Research Questions and Objectives

This research takes an empirical look at the effects of explanation completeness on the users of a clinical decision support system. The aim of the research is to answer the following five research questions that emerged from a review of the related literature:

**RQ1: How does explanation completeness impact clinical decision-making?**

**RQ2: How does explanation completeness impact clinician's confidence in their ability to diagnose patients?**

**RQ3: How does explanation completeness affect a clinical user's trust in a CDSS?**

**RQ4: What is the impact of explanation completeness on a clinical user's workload?**

*Null Hypothesis: Explanation completeness and a clinician's workload are independent of each other.*

**RQ5: What information do clinician's desire in a CDSS's explanation when making a diagnostic decision?**

The objectives of this research are to:

1. Clarify the role of decision support systems within the medical domain
2. Identify issues experienced by clinical users of decision support systems
3. Examine how system explanations might provide intelligibility and mitigate the identified issues
4. Design an experimental study to explore the impact of explanations on clinicians using a CDSS
5. Evaluate the results of the study, as pertaining to the five research questions

Objectives 1, 2, and 3 are reached through a review of the related literature and, thus, the value of these objectives lie in a synthesis of existing knowledge. Objective 4, of course, is a step that is necessary in order to achieve Objective 5, which will answer the five research questions above.

## Section 1.3 Products of the Work and Beneficiaries

The products of this research are an improved understanding of:

- The effect of explanation completeness on clinical users' decision-making, trust, confidence and workload
- The information that clinicians consider when assessing a CDSS's suggestion
- The explanations clinicians desire from a CDSS

Additionally, a Wizard-of-Oz CDSS prototype was produced. This prototype served as a means to incorporate the intelligibility explanations mentioned above to facilitate the assessment of trust.

The results of this study will benefit researchers, healthcare professionals, system developers, and patients. Regarding the Human Computer Interaction community, the results will contribute an improved understanding of information required when non-specialized or inexperienced clinicians are faced with making critical decisions when aided by a specialized CDSS.

A more generalizable contribution will be an improved understanding of the effects of applying an existing intelligibility method, *Why...* explanations, to decision support systems. Looking beyond the benefits to academia, the results may be used for practical applications. CDSS developers could incorporate the findings to design more effective support aids to be used by healthcare professionals, ultimately improving the quality of care that patients receive.

## Section 1.4 Project Scope

This research is focused on clinical decision support systems designed to aid clinical users with a decision-making task in a medical domain they are not specialized in. This specific focus comes from this author's work on a project commissioned by the European Union entitled 'EMBalance' (<http://www.embalance.eu>). The EMBalance project required the design of a clinical decision support system for diagnosing balance diseases to be used by general practitioners. The wireframes designed for the EMBalance project were done so by the author, and were then used during this research. Thus, the CDSS wireframes and Wizard-of-Oz prototype used in this research are designed to aid non-specialized clinicians in the diagnosis of *balance* diseases, a specialist field of medicine.

Only the factors associated with a general practitioner's assessment of a CDSS' suggestion were examined, rather than impressions or interactions of the system as a whole. These impressions include perception, interpretation, and believability. Additionally, clinical guidelines were not examined. Group use of the system was not explored, nor was patient care, quality of care, economical benefits, organizational integration, or professional autonomy.

## Section 1.5 Outline of Methods

This research project took place over the course of six months, from 01 July 2014 until 09 January 2015. The first three months of the project were spent conducting the literature review and designing the CDSS wireframes, whilst

the author worked full-time on the aforementioned EMBalance project. An experimental study was designed and conducted during the month of October. During the study, participants were provided with case studies and observed using the prototype to achieve a diagnosis for each case study. Seven study sessions were completed in October and the results of these were analysed and written from November to mid-December. The remainder of the time was spent writing the report. The work plan created for this project is included in Appendix A. Changes to the goals and methods are discussed in Chapter 6.

## **Section 1.6 Report Structure**

This dissertation is organized as follows.

### **Chapter 1: Introduction and Objectives**

Chapter 1 introduces the reader to clinical decision support systems and the current problems that surround them. A potential solution is introduced and the need for empirical research is emphasized. The research questions, objectives, project scope and beneficiaries are also identified.

### **Chapter 2: Critical Context**

This chapter defines automated and decision support systems, going into specific detail about the history and benefits of clinical decision support systems. CDSS reliability and issue of automation bias are presented. Providing intelligibility through the use of explanations - a potential solution - is illustrated, along with the potential benefits and risks that could result from this solution.

### **Chapter 3: Methods**

Chapter 3 provides the reader with details on how this research study was conducted and how the collected data was analysed. Particulars regarding recruitment, material design, study session flow, and analytical methods are included.

### **Chapter 4: Results**

Chapter 4 presents the results of the data analysis. Each section in this chapter presents the results that pertain to one of the five research questions.

### **Chapter 5: Discussion**

The results presented in Chapter 4 are discussed with context to the research questions, related literature, and generalizability. This chapter concludes with a discussion of future avenues of research based upon the findings.

### **Chapter 6: Evaluations, Reflections, and Conclusion**

Chapter 6 concludes this research report with considerations on what has been achieved, learnt, and areas for improvement.

## Chapter 2 Critical Context

### Section 2.1 Overview

This chapter presents the critical context of clinical decision support systems. It begins with an introduction to intelligent systems, defining automated systems and decision support systems. Clinical decision support systems are then introduced, with examples of how these systems have been used throughout the last forty years. The problems of CDSS reliability and a user's automation bias are then defined. Following this, intelligibility is introduced and system explanations are discussed as potential solutions; particularly *Certainty* and *Why...* explanation types. The topic of explanation completeness is presented and the literature review concludes with a discussion of the risks and benefits that may occur as a result of different completeness levels.

### Section 2.2 Intelligent Systems: Automation and Decision Support

Computing systems have advanced beyond simple tools that act, or *react*, only when prompted by a user. There now exist intelligent systems that are capable of aiding human users achieve a goal. These systems have become increasingly pervasive in the everyday lives of humans: sorting our e-mail, reminding us of meetings, changing the temperature of our homes when we are not around. These systems are known as 'automated systems' and they are described as technology that executes a task that a human would typically perform (Parasuraman & Riley, 1997).

Automated systems work by accessing data and applying 'rules' to determine what action should be taken. Because automation can reduce the physical or cognitive workload of humans (Parasuraman & Riley, 1997), they are able to bring improvements in safety, efficiency, and performance (Grabowski & Sanborn, 2003). Thus, automated systems are becoming increasingly popular within time- and safety-critical domains such as aviation, manufacturing, automobiles, and medicine (Parasuraman & Riley, 1997).

Decision support systems, a type of automated technology, are designed to facilitate a user's decision-making process by compiling complex information and returning it in a comprehensible and meaningful manner (Lim & Dey, 2010) (Seong & Bisantz, 2008). This returned information, or output, may be in the form of an alert, a suggestion, or filtered information (Alberdi et al., 2005). Because decision support systems are able to provide this output automatically and in real-time, they are often designed specifically for use in domains that are not only safety-critical, but time-critical as well. Users maintain decision-making authority, but with the enhanced experience of intelligent and timely support from a decision system (Alberdi et al., 2005).

## Section 2.3 Clinical Decision Support Systems

This research focuses on the use of decision support systems within the medical domain. Clinical decision support systems (CDSS) have been used in the medical field for nearly four decades and are highly regarded as beneficial aids when used by healthcare professionals making clinical decisions (Kawamoto et al., 2005) (Kong et al., 2008). Clinical decision support systems offer improvements to evidence-based care, prescribing practices, preventive care, and adherence to best practices (Kawamoto et al., 2005). There are several definitions of clinical decision support systems (Kong et al., 2008), but all are more or less the same. This research will use Osheroff et al's definition: CDSSs 'provide clinicians or patients with computer-generated clinical knowledge and patient-related information, intelligently filtered or presented at appropriate times, to enhance patient care' (Osheroff et al., 2005).

There are four types of clinical decision support systems: those that respond to clinical data with alerts, those that analyze decisions already made and suggest alternatives, those that perform quality audits, and those that review patient data and suggest appropriate clinical actions in real time (Alexander, 2006). With this breadth of capabilities, CDSSs are used for a variety of purposes across healthcare and by a variety of roles (Sambasivan et al., 2012). For example, one of the first CDSSs was MYCIN, which was designed in the 1970s to identify risks for bacterial infections and recommend appropriate antibiotic prescriptions based upon the patient data (Shortliffe, 1976). PSIP, another European commissioned project, resulted in a prototype designed to prevent clinicians from making medication errors by providing alerts of adverse drug events (Kanstrup et al., 2010). More recently, Povyakalo et al conducted a study on a CDSS that is used in radiography; the system was designed to aid lab technicians detect breast cancer by highlighting areas of interest on patient slides (Povyakalo et al., 2013).

Of course, there are problems that surround clinical decision support systems. A large issue, which has been a popular topic in academia, is that these systems often fail to be adopted or fully utilized because of trouble with organizational integration or the end-user's desire for professional autonomy (Kawamoto et al., 2005) (Sambasivan et al., 2012). But the scope of this research study looks beyond the challenge of adoption towards a more open-ended issue: decision support systems can be unreliable but users may not notice.

## Section 2.4 CDSS Reliability and Automation Bias

While they are designed with the intention to reduce or prevent human error, automated systems are not perfect. It's not uncommon for these systems to



make decisions under uncertainty, and this means that the actions or suggestions they make can sometimes be wrong.

Reasoning under uncertainty is of particular concern within the medical domain. The process of diagnosing a patient requires balancing unknowns and dealing with ambiguity. Patients may inaccurately explain their symptoms, clinicians may misunderstand parts of human physiology, lab results include degrees of error, and accurately determining a patient's prognosis is impossible (Kong et al., 2008). Whilst a CDSS can apply complicated rules or reasoning to help a clinician reach a diagnosis amidst this uncertainty, the system is not immune to it, nor is the system perfect. A CDSS may suggest an incorrect diagnosis or recommend an incorrect drug dosage, but it is still the clinician's responsibility to make the correct decision (Kong et al., 2008).

#### **2.4.1 Automation bias**

Acknowledging the issue of reliability, engineers may try to reduce system errors or failures by designing more robust algorithms. However, it is not possible to develop a system that is perfectly reliable (Parasuraman & Riley, 1997). Users must monitor the system's output and critically analyse the veracity of a decision support system's advice. Still, users may come to rely too much on these systems and may not always be sensitive to their reliability (Wiegmann et al., 2001). In fact, automated systems are commonly perceived as highly accurate or reliable, even more so than other humans (Madhavan & Wiegmann, 2007)(Parasuraman & Riley, 1997). Thus, their users may become biased towards the automation, believing the system to be more reliable than it actually is. This over-reliance is most common among users with low self-confidence in their personal abilities or judgment, leading them to believe that an automated system's actions or a decision support system's suggestions are correct and more reliable than their own (Lee & See, 2004). The over-reliance of a user on an automated system's advice or activities is known as *automation bias* (Alberdi et al., 2005).

Healthcare professionals who use CDSSs are not immune to automation bias (Alberdi et al., 2009) (Kong et al., 2008) (Pu & Chen, 2007). Although designed and implemented with the best intent, the use of computer aids in a clinical setting may not always have beneficial results on the user's judgment or resulting quality of care (Kong et al., 2008) (Povyakalo et al., 2013). For example, Alberdi et al (2004) conducted an experiment comparing the detection of cancerous areas made by radiographers when using a computer aided detection system and when not using the system. The results of this study showed that the critical eye of the radiographers lapsed when aided by the system; the participants came to rely on the system to detect areas of cancer, subsequently failed to identify cancerous areas that they noticed when not using the system (Alberdi et al., 2004).



## Section 2.5 Intelligibility through System Explanations

### 2.5.1 Intelligibility

Intelligible systems are those that make their behaviour, reasoning, or inner-workings transparent to the user (Lim & Dey, 2011). The complexity of a CDSS's reasoning can range from simplistic rules (such as 'if x, do y') to black-box methods that are difficult for end-users to comprehend (such as neural networks) (O'Sullivan et al., 2014). The more straight-forward CDSSs are, the easier it is for a clinical-user to form an understanding of how they work. While the more complex ones (such as diagnostic aids) are designed with the clinical user in mind, they often fail to incorporate the necessary level of transparency that would allow its users to understand the reasoning process (O'Sullivan et al., 2014). Perhaps this explains why CDSSs that are designed to provide alerts have been shown to be more effective than CDSSs that assist diagnosing a patient (Bates et al., 2003).

Alberdi et al write that the most frequently given reason for lack of acceptance of a CDSS 'has been a failure in the system to incorporate an adequate knowledge of the cognitions and working practices of the eventual users' (Alberdi et al., 2001). Whilst not explicitly stated, it can be suggested that this is the result of lack of transparency – the clinical users are not being provided adequate information to determine if the system's reasoning is on par with their own. Thus, a key factor in the success of a CDSS is the effective communication, or intelligibility, of the reasoning (Alexander, 2006) (Gonul et al., 2006). O'Sullivan et al agree, calling for more research to build an understanding of how sophisticated CDSSs can provide transparency to meet the needs of clinical users (O'Sullivan et al., 2014)

### 2.5.2 System Explanations

Explanations are descriptions provided by the system to help the user understand what it is doing (e.g., justify its reasoning, certainty, or behaviour) (Gregor & Benbasat, 1999) (Hasling et al., 1984). Explanations provided alongside a system's output could be effective solutions to the three problems highlighted above. In fact, explanations are the most common method for providing intelligibility (Pu & Chen, 2007) and are considered by some to be the most effective means for influencing a user's acceptance in a decision support system's suggestion (Gonul et al., 2006).

There are many different kinds of explanations that can be implemented in a system, though some are more appropriate for specific systems than others (Gonul et al., 2006) (Lim & Dey, 2010). For example, e-commerce recommender systems use explanation types that aim at persuasion, satisfaction, and trust (Tintarev & Masthoff, 2007). These explanations are geared more towards influencing end-users to accept the recommended product and are inappropriate for clinical decision support systems. There are

also terminological explanations that are used to define information (Gregor & Benbasat, 1999), info-graphic type explanations to show location-awareness (Lim & Dey, 2011), and strategic explanations that define the system's problem solving strategy (Gonul et al., 2006). However, research on decision support systems suggests that users prefer 'justification' type explanations (Gonul et al., 2006).

Justification explanations provide the user with the rationale behind a suggestion (Gonul et al., 2006). Whilst developed specifically for context-aware automation, three of Lim and Dey's defined explanation types suit this purpose: *Certainty*, *Why...*, and *Why Not...* (Lim & Dey, 2010). *Certainty* and *Why...* explanations are already used within clinical decision support systems (refer to Hasling et al's work on NEOMYCIN for additional detail). Work by Kong et al suggests that these two explanation types are key features to a successful CDSS as they include intelligible representations of the system's uncertainties and knowledge base (Kong et al., 2008). However, they have not yet been studied empirically in regards to decision-making (Gonul et al., 2006).

*Certainty* explanations "inform users how (un)certain the application is of the output value produced" (Lim & Dey, 2010). A decision support system will use the available information to calculate and return to the user a suggestion, or set of suggestions, that best match. Therefore, the system output is not always 100% appropriate, or a 100% match to the information. This is especially true within the field of medicine, as every stage of a diagnostic process is fraught with degrees of uncertainty, particularly the final determination of a diagnosis (Kong et al., 2008). Researchers have shown that use and decision-making are improved when users are provided with an indication of the system's certainty (Price, 2012). This approach is logical, for it provides the user with a rational basis upon which they may begin determining whether or not they should accept the system's suggestion. An experiment performed by Lim and Dey (2011) confirms this, showing that a user's perception of system certainty is better calibrated when the system actually displays the certainty level.

Despite the logic of providing *Certainty* explanations, some researchers suggest that they may not be enough to improve clinical decision-making. Studying the diagnostic decisions made by 17 dentists using a CDSS, Choi et al found that the majority do not adequately understand the certainty provided by the system, resulting in inconsistent and inaccurate decision making (Choi et al., 1998). Kong et al (2008) wrote that key features to a successful CDSS include intelligible representations of the system's uncertainties as well as its knowledge base, suggesting that *Why...* explanations should be paired with *Certainty* ones.

A *Why...* explanation informs users ‘why the application derived its output value from the current (or previous) input values’ (Lim & Dey, 2010). Without an explanation of the system’s reasoning, users are more likely to distrust the suggestion and seek external sources (such as a book or colleague) to verify it (Glass et al., 2008). When applied to a decision support system, a *Why...* explanation indicates to the user what data or information was used to determine the suggestion. Research has shown that *Why...* explanations improve system intelligibility, help the user adequately understand the system’s reasoning, and determine if the output is appropriate (Hasling et al., 1984) (Lim et al., 2009).

### 2.5.3 Explanation Completeness

While *Why...* explanations are commonly used in systems, there is no prescribed way to use them. How much information does the system need to show? Are users interested in a list of *all* the reasons why the system suggested x? Or are they only looking for the major reasons?

Explanation completeness is “the extent to which an explanation describes all of the underlying system” (Kulesza et al., 2013). For a complete *Why...* explanation, a system must show *all* of the reasons why a suggestion was made. While a more complete explanation may provide transparency, reduce frustration, and improve trust (Kulesza et al., 2013), it does so at cost to the user. More information may actually be over-information, a lengthy list of reasons rather than a succinct statement taking up more space on a screen than it is worth. But an explanation with low completeness may appear to have missing information and be perceived negatively, leading to reluctance to act (Baron, 2000). Additionally, while some researchers have shown that users with more information make better decisions (Seong & Bisantz, 2008), some have shown it makes no impact (Gonul et al., 2006), and others have shown that it makes decisions worse (Alberdi et al., 2004).

## Section 2.6 Critical Context Summary

Reviewing the literature related to clinical decision support systems and decision-making, a gap in academic knowledge appears. *Certainty* and *Why...* explanations may be the solution to automation bias, intelligibility, and CDSS uncertainty. However, an explanation with high completeness could lead to misuse of the system, but low completeness could lead to distrust or disuse. The completeness of the explanation effects the decisions made as well as the user’s trust and confidence, but also may demand more cognitive faculties than the clinical user prefers. This research project aims at reducing this gap by empirically studying the impact of explanation completeness on clinical users of CDSSs.

## Chapter 3 Methods

### Section 3.1 Overview

As outlined in Chapter 2, there is a lack of empirical research on the use of explanations and appropriate use of a clinical decision support system. In response to Objective 4, this chapter presents the construction of a between-group experimental study. This experiment was designed to answer the five research questions and contribute knowledge to the HCI community by gathering both qualitative and quantitative data.

To perform this study, a Wizard-of-Oz prototype was developed based upon the design of a CDSS developed for the European Union commissioned project, EMBalance. This was used as a tool to observe the participants and acquire the necessary data for analysis. Two conditions were tested; one version that provided explanations of low-completeness, and another that provided explanations of high-completeness. The conditions are as follows:

- a) Suggested diagnosis and percentage of system certainty supported by an explanation of *high* completeness (listing the medical history, symptoms, and examination results that match the disease profile). This is referred to as System A and was assigned to the High Completeness group.
- b) Suggested diagnosis and percentage of system certainty supported by an explanation of *low* completeness (listing only the examination results that match that disease profile). This is referred to as System B and was assigned to the Low Completeness group.

Seven participants were involved in the study. In an attempt to form balanced groups, each participant was assigned to a group based upon their stated expertise with balance disorders. Before using the prototype, participants were asked to rate their trust in a CDSS, as well as their confidence in their ability to diagnose patients with balance disorders. Participants were provided with case studies describing patients suffering from balance diseases. Using these case studies, the participants were observed using the assigned prototype to diagnose the patients. For each case study, participants were asked to determine whether or not they accepted the CDSS's suggested diagnosis; their verbal utterances of the decision-making process and final decisions were recorded. After each case study was completed, participants rated their workload using a NASA Task Load Index (NASA TLX) rating sheet. Upon completion of the case study tasks, participants were asked to rate their trust and confidence again, as well as describe what – if any – impact the explanations had on their ratings.

This chapter begins with Section 3.2, which introduces the research strategy, provides justification for the approach and outlines the ethical considerations

that were made. Section 3.3 describes the target population, recruitment effort, and participants that took part in the research. Section 3.4 provides step-by-step details of how the study sessions were conducted, whilst 0 illustrates how the collected data was analysed for each research question.

## **Section 3.2 Research Strategy**

### **3.2.1 Between-Group Experiment**

As mentioned, the conditions were tested using a between-group design. A between-group design, in this research, was required because the similarity of the two conditions would incur a learning effect on users (Purchase, 2012). A between-group design is commonly used by other HCI researchers focused on appropriate use of automation, such as (Choi et al., 1998) (Lim & Dey, 2011).

Acknowledging that the level of prior experience with balance diseases would impact personal certainty and belief, as well as introduce judgment bias, it was important to create balanced groups (Purchase, 2012). Thus, participants were sorted into one of two groups based upon their self-stated level of knowledge of balance diseases. The final question in the screener (Appendix D) required the interested volunteer to rate their level of knowledge in the domain of balance diseases. This four-point self-assessment scale was used to create the balanced-groups in the experiment. Whilst this decision was necessary, the downside is that it is impossible to ensure that groups are perfectly balanced.

### **3.2.2 Observing through Simulated Use**

Some research on explanations and decision-making has been conducted without the use of real, physical prototypes (for example, research conducted by Lim and Dey in 2011). However, this author believes that creating a notion of realism for the participants involved is necessary in order to effectively gather generalizable data. While observing participants use the prototype in the wild while interacting with actual patients would have ensured a high amount of realism, it would be unethical, as well as nearly impossible within the timeframe, to do so.

Conducting a remote un-moderated study was another option, one that has been performed by other researchers, such as Lim and Dey (2011) and Kulesza et al (2012). While this option may have resulted in a higher number of participants, it was determined not to be viable as there would be no way to ensure that the participants were clinically trained individuals. Thus, an appropriate alternative to this method was chosen: the research would be conducted in person with the use of case studies to simulate the interaction between the user, the patient, and the system.

### **3.2.3 Ethical Considerations and Risks**

An ethical consideration checklist was completed as a part of the project proposal and the approved form is provided in the appendix of this document (Appendix B). This author acknowledges the importance of ethics, confidentiality, and safety. Section 3.4.1.1 describes the approach to ensuring safety during the sessions while Sections 3.4.2 and 3.4.4 detail Informed Consent and Confidentiality.

## **Section 3.3 Participants**

### **3.3.1 Target Population**

The scope of this research is on the use of a CDSS designed to aid clinically trained users with no special knowledge on balance disorders reach a diagnosis. The target populations for the research are individuals who are clinically trained in the field of *general* medicine, such as general practitioners and practice nurses, as well as medical students training to fill one of those roles. While the prototype used in this system is regarding balance disorders, the generalizability of the results are still high, given that the participants observed are general in their knowledge.

General Practitioners, or GPs, are the main target population for this system and other CDSSs like this, as these end-users deal must apply their broad (non-specialized) medical knowledge to the vast array of symptoms and diseases they are faced with on a day-to-day basis. Nurse practitioners, practice nurses, and nurses of similar roles, working within a general practice surgery are also a target population. Whilst the nursing role may sometimes involve less diagnostics than their general practitioner co-workers, it often includes a higher level of activity in performing the examinations that lead up to the diagnosis. As such, it is entirely realistic to expect a practice nurse to use a system, such as the one designed, to have support in the process of examining and diagnosing patients. Additionally, medical students and residents with experience working in clinics are another target population. These individuals are considered to be a likely user of such a CDSS as their work involves the examination and diagnosis of patients; yet their confidence, experience, and knowledge is considerably lower than that of a GP and even some practice nurses.

### **3.3.2 Recruitment and Sampling Strategy**

This research used a random sampling approach. Recruitment advertisements, included in Appendix C, were written and used to recruit volunteers for this study. The advertisement, which was created in various formats, gave a summary of the study purpose, description of the desired participants, outline of the session, and defined the compensation to be

provided to participants as thanks. The advertisement explained that the aim was to observe how clinically trained individuals made decisions when using a clinical decision support system.

Those with experience in research involving clinicians will agree that recruiting this population is indisputably difficult. Therefore, attracting the attention of potential participants required a major effort. The recruitment effort began in late September and ended on the first of November. During this time, the advert was distributed digitally across various channels; personal networks, various university mailing lists, medical network groups, and numerous clinician-specific forums. Flyers were also posted in medical schools, job fairs, and events throughout London. The information on the flyer was also sent in the form of personal letters addressed to GPs in neighborhood surgeries. Additionally, flyers were posted in local establishments frequented by nearby medical students and professionals, such as pubs, coffee shops, and hospital-staff gymnasiums.

### **3.3.3 Screening with Inclusion and Exclusion Criteria**

There are exclusions to the target populations. Given the focus on non-specialist users, it is rather obvious that experts of balance diseases such as those in the field of Ear Nose Throat (ENT), neuro-otology, audiology, physical therapy, and otolaryngology, must be excluded from participating in the study. Other exclusions include clinicians who have general medical knowledge but do not work in an outpatient setting. While paramedics, visiting nurses, and care providers may come across patients with balance diseases, the settings in which they would do so involve very different time constraints for examinations and diagnoses, very different environments of use, and may not allow for the use of a computerized system such as the one being researched.

A set of inclusion/exclusion criteria was written into a screening document to ensure participants in the study were representative of the target populations. Screening the volunteers was conducted verbally, over the phone, prior to scheduling a session time. The screener contains five questions about the interested volunteer's experience and role, ensuring non-representative individuals were excluded prior to a study session being conducted. The answers to these questions were recorded and used as demographic information, and not used for analysis. Criteria such as age, gender, and years of experience were left open. Allowing a range of ages, genders, and experience was intended to increase generalizability of the findings, so that the results can apply to other decision support systems. The inclusion/exclusion screener can be found in Appendix D.



### 3.3.4 Recruited Participants and Group Assignment

The allotted recruitment time period (the month of October) yielded seven participants total: four in the High Completeness group and three in the Low Completeness group. Each participant was assigned a group (Low Completeness group or High Completeness group). An attempt was made to construct balanced groups based upon each participant's experience in the specific domain. Thus, participants were assigned a group based upon their response to Q5 in the Screener, which asked them to rate their level of balance disease expertise on a four-point scale.

Table 1 presents the participants assigned to each group, along with their title, practice setting, and self-rated expertise in balance diseases.

<b>High-Completeness Group (System A)</b>		
<b>ID</b>	<b>Occupation, practice setting, and title</b>	<b>Self-rated balance expertise (1 = none, 4 = expert)</b>
A01	CT1 SHO, works in A+E, has GP experience	3
A02	Medical officer, similar to GP	2
A03	Nurse practitioner in GP surgery	2
A04	1st year foundation doctor in GP surgery	2
<b>Low-Completeness Group (System B)</b>		
<b>ID</b>	<b>Occupation, practice setting, and title</b>	<b>Self-rated balance expertise (1 = none, 4 = expert)</b>
B01	Student, final year of graduate medicine	3
B02	1st year foundation doctor in GP surgery	2
B03	General Practitioner in GP surgery	3

**Table 1: Details of Recruited Participants**



## Section 3.4 Study Session Outline

Each volunteer participated in one session only. Study sessions were designed to last 60 minutes, though participants were not rushed. Each session followed the outline described in the following sub-sections.

### 3.4.1 Session Locations

The seven research sessions were held in quiet environments within Greater London. The research sessions were held in various locations around the city to allow greater convenience and incentive for participation to the clinical volunteers. One session was held at a participant's home, one in a private meeting room in the Town Hall of Middlesex University, one in the examination room of a London surgery, one in a quiet corner of a coffee shop, and three were held in private meeting rooms at City University London.

In each of these sessions, a small table was shared between the participant and the researcher. The table held the laptop and external wireless mouse, as well as serving as a surface for placing the various papers used throughout the session.

#### 3.4.1.1 Risks and Safety in session locations

As mentioned, the seven research sessions took place in various locations around London. Six of these were in public locations, reducing the risk of any discomfort or harm coming to either the researcher or participants. Only one study session took place in a participant's home. However, this participant and the researcher shared a close friend who made the introduction, thus reducing the risk coming to either party. Had this participant not been a trusted individual by the shared friend, the meeting would have taken place in a public location.

### 3.4.2 Materials and Equipment used

Each study session required various materials and equipment to conduct. Paper documents included the Information Sheet, Consent Form, Pre-Use Questionnaire, Interview Guide, NASA TLX rating forms, and Post-Use Questionnaire. Each of these will be described later in this chapter. Additionally, the case studies were provided to the participants on small slips of paper. Details on the construction of the case studies are provided in the subsection that follows.

The Wizard-of-Oz prototypes were created in Axure RP and were presented on the researcher's own MacBook Air (13"). To allow the reader to interact with these prototypes, they have each been uploaded for public access.

System A (with explanations of high-completeness) can be accessed at:

<http://a8qomn.axshare.com>

System B (with explanations of low-completeness) can be accessed at:  
<http://igzu13.axshare.com>

The study sessions were recorded using Camtasia for Mac, which captured video recordings of the computer screen and audio recordings of the verbal utterances.

#### 3.4.2.1 Case Study Design

Eight case studies were created for this research project. This subsection will first describe how *correct* case studies were constructed. Following this will be details of how the case studies were manipulated to create false positives. Finally, the creation of the explanations of high and low completeness is made clear. An example will be given to illustrate this process.

Each case study includes a patient identification letter (rather than a name), age, gender, a short list of medical history and symptoms, four suggested examination results, and one diagnosis. The content for the case studies was generated using diagnostic decision-trees and descriptions of balance diseases (specifically: (NIH National Institute on Deafness and Other Communication Disorders (NIDCD), 2010) (Vestibular Disorders Association, 2014) (Shupert & Kulick, 2013) (Bamiou et al., 2014)). The first step to creating a case study was the selection of a balance disease. Then, using the decision-trees and descriptions, the examinations that are used to confirm the diagnosis were identified along with the examination results. These details were recorded in a Microsoft Word document and labeled as “Examinations”. Following this, the typical symptoms presented by patients with the selected disease were identified and added to the document under the label “Symptoms”. If the description of the disease included any risk factors or common causes, such as smoking or drinking, these were identified as well and recorded as “Medical History”. Building off of this structure, various bits of detail were added to the case studies to develop the character of each patient described. Examples of these details include gender, age, and occupation. These details were also added to the word document, completing the construction of a correct case study.

In order to examine the decisions made by the participants it was necessary to add a certainty percentage to each and to manipulate some of the diagnoses to create false positives. Adding the certainty percentage was a rather straightforward process. Half of the case studies were given low-certainty (between 19 and 28% system certainty), while the other half were given high-certainty (between 77 and 84%). The certainty percentage assigned to each case study was then recorded in the word document.

Next, two of the high-certainty and two of the low-certainty case studies were selected to be the *incorrect* case studies, or ‘false positives’. To create the false positive, the correct diagnosis was replaced with an incorrect one.

However, one cannot simply swap diseases and confidently state that a false positive has been created, the incorrect diagnosis must be carefully chosen to ensure that it is, in fact, incorrect. For example, Case Study J was manipulated to be a false positive. The suggested diagnosis is Age-Related Imbalance, while the correct diagnosis is Labrynthitis. In this instance, Age-Related Imbalance was appropriate because the symptoms presented by patient are common of age-related diseases, yet actually symptoms of a viral infection (Shupert & Kulick, 2013). Using true positives (correct diagnostic suggestions) and false positives (incorrect diagnostic suggestions) in this experiment would allow us to identify when participants made 'right' and 'wrong' decisions. True and false positives have also been used in research conducted by Alberdi et al (2005). Table 2 presents the results of the manipulations.

	<b>Correct Diagnostic Suggestion</b> (True Positive)	<b>Incorrect Diagnostic Suggestion</b> (False Positive)
<b>High System Certainty %</b>	Case Study L	Case Study K
	Case Study G	Case Study Q
<b>Low System Certainty %</b>	Case Study N	Case Study J
	Case Study D	Case Study R

**Table 2: Breakdown of Case Study Manipulation (High and Low Certainty, True and False Positives)**

Finally, the explanations for each diagnosis were written. The explanations with high-completeness listed items in the patient's history, symptoms, and exam results that matched the profile of the suggested disease. The explanations with low-completeness listed only the items in the patient's examination results that matched the disease profile.

Each case study was then printed out onto sheets of A4 paper and cut up into strips to be handed to the participant at the appropriate time during the study session. The completed case studies and related explanations are included in Appendix H.

### **3.4.3 Informed Consent**

Using a template provided by City University London, an information sheet and a consent form were created. Volunteers who were successfully recruited were sent an information sheet and consent form to review prior to

the study session. This was done to allow the volunteers ample time to review the information. The information sheet included a description of the research purpose, a summary of the study session, and what would be done with the data upon completion. Additionally, it informed participants that the study was not focused on *their* abilities, but on the use of the prototype and how they made decisions with it. This point was also verbalized in the study introduction in order to ensure no negative emotional impact occurred as a result of participation.

All recruited participants were also provided with a printed copy of the sheets during the introductory section of the study session. Participants were offered the choice of reading the details privately or reviewing the document together before being asked to sign. A signed copy was provided to each participant for his or her records. Appendix E shows the information sheet used.

#### **3.4.4 Confidentiality**

As part of the informed consent, all participants were informed of the data that would be gathered, stored, and used for the purposes of the research. Signed consent forms and pre-screeners were placed in a sealed envelope stored in a private locked filing cabinet. Participants were assigned a unique identification number that indicates their assigned group and session number. For example, the first participant assigned to the High Completeness group (which used System A) was given the ID 'A01'.

Identifiable information (such as names or specific workplaces) captured during the session recording was not included in the transcripts. All digital files were stored on a personal laptop in a password-protected folder, as well as on a password-protected external hard-drive. All physical materials, such as completed rating sheets, were stored in a locked cabinet and kept separate from the signed consent forms. All files related to this study will be kept for two years, and then destroyed.

A copy of an unsigned Consent form is included in Appendix F. The completed pre-screeners and signed consent forms are included in the Appendix K, the Confidential Appendix.

#### **3.4.5 System Description**

The participants were introduced to the purpose and design of the CDSS prototype. To ensure each participant received the same information, the researcher read directly from the script written in the interview guide (Appendix G.1).

#### **3.4.6 Workload Introduction**

Following this introduction, the participants were told that part of the research was examining the impact that explanations had on workload. Each

participant was handed a copy of the NASA TLX workload source definitions (Appendix G.3) and given time to read before continuing.

The decision was made to measure each participant's workload by using the NASA TLX technique. The NASA Task Load Index measures the workload of users by collecting subjective ratings on six subscales (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration) and using these ratings to calculate an overall workload score (NASA Human Performance Research Group, 1987). This method was deemed appropriate for use in this study as it is highly rigorous; it is the result of more than three years of research (NASA Human Performance Research Group, 1987). Additionally, it has been used in other related research, such as the recent publication by Kulesza et al (2011). Additional details on the NASA TLX method and instructions on how to apply it to a research project can be found in the Paper and Pencil Package (NASA Human Performance Research Group, 1987).

#### **3.4.7 Pre-Use Questionnaire**

Participants were handed a piece of paper containing a two-question Pre-Use Questionnaire regarding self-confidence and trust. These two questions asked participants to indicate their response on the seven-point rating scale provided. The questions were read aloud to the participant, who was then asked to mark the rating they felt best applied.

Q1: How confident are you, currently, in your ability to diagnose patients with balance diseases?

*(1 is 'Not at all confident' while 7 is 'Extremely confident')*

Q2: How much do you feel you would trust a system like the one I've described to help you diagnose your patients with balance diseases?

*(1 is 'Distrust completely', 7 is 'Trust completely')*

This paper was then set out of sight so that the participant could not later refer to the original rating given during the Post-Use Questionnaire. The purpose of gathering pre-use ratings is to gather a benchmark with which the post-use ratings can be compared. The Pre-Use Questionnaire is included in Appendix G.2.

#### **3.4.8 Decision-Making Tasks**

Following the Pre-Use Questionnaire, participants were informed that the diagnostic decision-making portion of the study session would begin.

Before beginning, the order in which the participant received the case studies was manipulated to counter for ordering effects. It has been shown that trust and confidence levels are affected by changes in a system's reliability or certainty (de Vries et al., 2003). Given this, if all participants were to receive

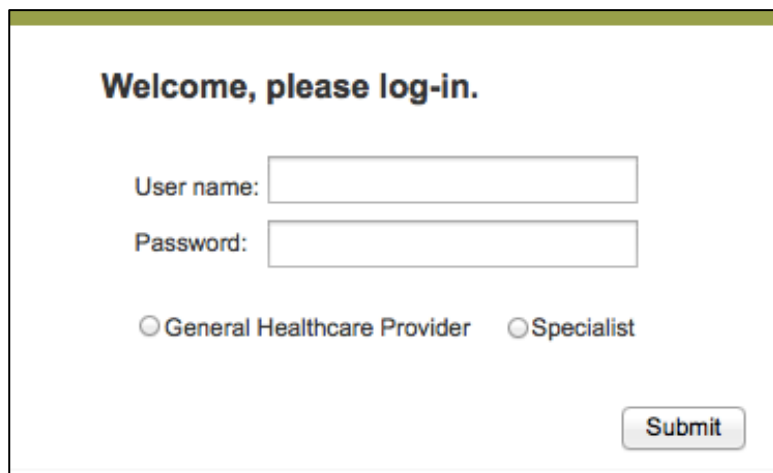
the case studies in the exact same order, it could result in a confounding variable. Thus, the order in which the case studies were presented needed to be rotated for each participant. With two groups and eight case studies each, a Latin Square Design was adopted to rotate the order that the case studies were presented. Table 3 presents the order in which each participant received the case studies.

ID	Group Assigned	Case Study Order
A01	High Completeness	KGDLNJQR
B01	Low Completeness	KGDLNJQR
A02	High Completeness	DLNJQRKG
B02	Low Completeness	DLNJQRKG
A03	High Completeness	NJQK
B03	Low Completeness	NJQRKGD
A04	High Completeness	NJQRKGD

**Table 3: Countering for Ordering Effects; the order in which the case studies were received**

It should be noted that participant A03 only completed four decision-making tasks. The study session with this participant ran over the allotted time, and only half of the decision-making tasks could be completed.

Once the order had been determined, it was recorded in the interview guide. The CDSS prototype assigned to that participant was then opened to the log-in page (Figure 1), where the researcher entered a username and password. The log-in screen was not a necessary feature of the prototype, but the decision was made to include it as it added a layer of realism to the experiment.



**Welcome, please log-in.**

User name:

Password:

☐ General Healthcare Provider ☐ Specialist

**Figure 1: Log-in screen**

After 'logging in', control of the laptop and wireless mouse was turned over to the participant. A wireless mouse, in this experiment, was an important accessory given that the laptop used was an Apple computer, one that has unique scrolling techniques that not all participants would be familiar or comfortable with.

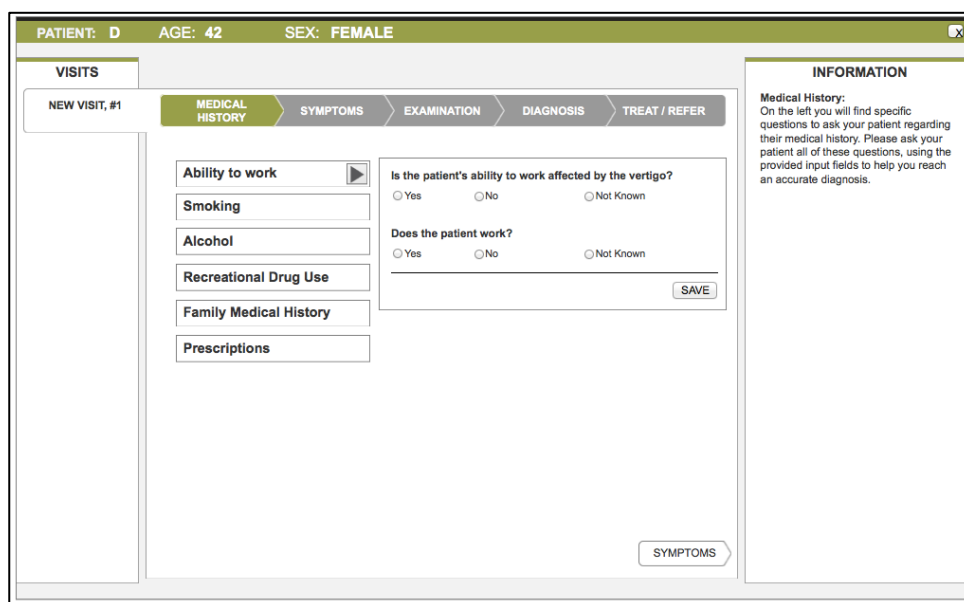


The screenshot shows a web form titled "Create new patient record" with a green header bar. Below the header, there are three input fields: "Patient ID:" followed by a text box containing "enter ID", "Age:" followed by a text box, and "Sex:" followed by a dropdown menu showing "- Select -". At the bottom right of the form is a button labeled "Create record".

**Figure 2: Create new patient record**

As the laptop was turned to them, the prototype displayed a screen to create a new patient record (Figure 2). The participant was then given a piece of paper with the demographic details and medical history of their first case study. They were instructed to use the demographic details to create a new patient record. To create a new patient record, the user must enter a patient identifier, age, and gender. This is a simplified version of the information a clinician would be required to enter into a CDSS in order to create a new patient record, but additional details (such as full name, date of birth, etc.) would have needlessly required additional time.

After creating a new patient record, the user is brought to the medical history page. Figure 3 shows an example of the interface where a user would input the patient's ability to work on the medical history page.



The screenshot shows a complex web interface for medical history. At the top, a green header bar displays "PATIENT: D", "AGE: 42", and "SEX: FEMALE". Below this, there are two main sections: "VISITS" on the left and "INFORMATION" on the right. The "VISITS" section has a sub-header "NEW VISIT, #1" and a series of tabs: "MEDICAL HISTORY" (active), "SYMPTOMS", "EXAMINATION", "DIAGNOSIS", and "TREAT / REFER". The "MEDICAL HISTORY" tab is expanded, showing a list of input fields: "Ability to work" (with a play button), "Smoking", "Alcohol", "Recreational Drug Use", "Family Medical History", and "Prescriptions". To the right of these fields, there are two questions with radio button options: "Is the patient's ability to work affected by the vertigo?" (Yes, No, Not Known) and "Does the patient work?" (Yes, No, Not Known). A "SAVE" button is located below these questions. The "INFORMATION" section on the right contains a "Medical History:" heading and a paragraph of instructions: "On the left you will find specific questions to ask your patient regarding their medical history. Please ask your patient all of these questions, using the provided input fields to help you reach an accurate diagnosis." At the bottom right of the interface is a button labeled "SYMPTOMS".

**Figure 3: Medical History with example of input screen**

After entering the medical history, the participants were asked to click on the 'Symptoms' button, which brought them to Figure 4, the symptoms input page. When this page appeared, the participant was handed a slip of paper with the patient's symptoms listed. Again, once the participant had completed entering the symptoms, they were asked to click on the button at the bottom of the screen to proceed. Participants were reminded that the system would suggest a series of examinations to perform, based upon the patient information that had been entered.

The screenshot shows a web-based medical interface. At the top, a header bar displays 'PATIENT: D', 'AGE: 42', and 'SEX: FEMALE'. Below this, a navigation bar contains several tabs: 'VISITS', 'MEDICAL HISTORY', 'SYMPTOMS' (which is highlighted), 'EXAMINATION', 'DIAGNOSIS', and 'TREAT / REFER'. On the left side, under the 'SYMPTOMS' tab, there is a section titled 'NEW VISIT, #1' followed by a list of input fields: 'Notes', 'Recent falls', 'Hearing loss', 'Tinnitus', and 'Vertigo / Instability'. On the right side, there is an 'INFORMATION' panel with the heading 'Symptoms:' and a paragraph of instructions: 'On the left you will find specific questions to ask your patient. Please ask your patient all of these questions, using the provided input fields to help you reach an accurate diagnosis.' At the bottom of the screen, there are two buttons: 'HISTORY' on the left and 'EXAMINATION' on the right.

**Figure 4: Symptoms Screen**

The examinations page (example shown in Figure 5) presents four 'recommended' examinations to perform; participants were told that the system recommends these examinations based on the medical history and symptoms of the patient. When the examinations page appeared, participants were handed individual slips of paper, each with one examination result written on them (photographic example in Figure 6). While a CDSS used in industry would allow the user to conduct additional examinations, the prototype used here does not.



PATIENT: D AGE: 42 SEX: FEMALE

VISITS

NEW VISIT, #1

MEDICAL HISTORY SYMPTOMS EXAMINATION DIAGNOSIS TREAT / REFER

SUGGESTED

Gaze Test Start

Straight-head Hang Start

Smooth pursuit Start

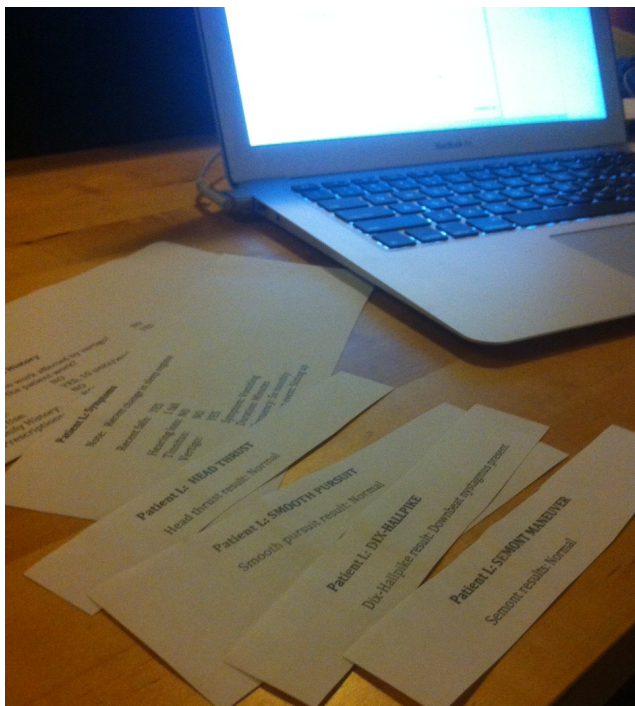
Gait test Start

SYMPTOMS DIAGNOSIS

INFORMATION

Examinations:  
On the left you will find a list of clinical examinations to perform on your patient. Please perform these exams, using the instructions and input fields, to help you reach an accurate diagnosis.

**Figure 5: Examination Screen**



**Figure 6: Photograph of paper case studies being used during a study session**

As with the previous screens, the participant was asked to click the button on the right hand side to advance to the system's suggested diagnosis. Some participants chose to pause and consider the information they had entered before advancing, and this was allowed.

Advancing from the examination page brought the participant to the suggested diagnosis. If the participant was assigned to the High Completeness group, they were shown a screen like Figure 8, which had an explanation of high-completeness. If they were assigned to the Low

Completeness group, a page like Figure 7 was shown. As the screen with the suggested diagnosis appeared, participants were reminded that they should think out loud about the suggestion and if they agree with it. They were asked what information they used to make their decision, and what information on the page was helpful or unhelpful.

PATIENT: D AGE: 42 SEX: FEMALE

VISITS  
NEW VISIT, #1

MEDICAL HISTORY SYMPTOMS EXAMINATION DIAGNOSIS TREAT / REFER

INFORMATION  
Suggested Diagnosis:  
Based on the information you have entered, the system suggests the most likely diagnosis, and the certainty.

Suggested Diagnosis:

Diagnosis: **Vestibular Migraine** Certainty: **26%**

This diagnosis is suggested because of matches to the disease profile

Examination results:  
Gaze test results: Motion intolerance  
Smooth pursuit results: Motion intolerance

☐ ACCEPT Diagnosis ☐ DECLINE Diagnosis

EXAMINATIONS CONFIRM

Figure 7: Suggested diagnosis with explanation of low completeness

PATIENT: D AGE: 42 SEX: FEMALE

VISITS  
NEW VISIT, #1

MEDICAL HISTORY SYMPTOMS EXAMINATION DIAGNOSIS TREAT / REFER

INFORMATION  
Suggested Diagnosis:  
Based on the information you have entered, the system suggests the most likely diagnosis, and the certainty.

Suggested Diagnosis:

Diagnosis: **Vestibular Migraine** Certainty: **26%**

This diagnosis is suggested because of matches to the disease profile

Medical history:  
Ability to work affected by vertigo

Symptoms:  
Recent falls  
Presence of Tinnitus  
Vertigo symptom: Migraine  
Vertigo triggering event: Diet

Examination results:  
Gaze test results: Motion intolerance  
Smooth pursuit results: Motion intolerance

☐ ACCEPT Diagnosis ☐ DECLINE Diagnosis

EXAMINATIONS CONFIRM

Figure 8: Suggested diagnosis with explanation of high completeness

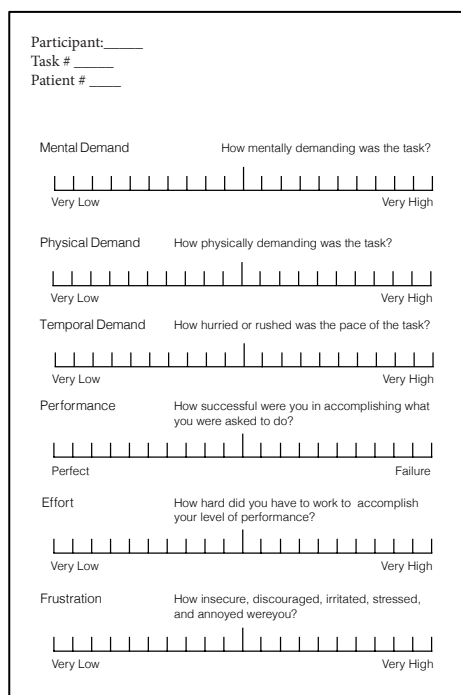
Once the participant had verbally declared his or her decision to accept or decline the suggestion, they were asked to indicate their decision on the screen, using the two radio-buttons, and click 'Confirm' (bringing the screen back to Figure 2). The participant's decision was recorded on the interview

guide.

There are small notes that are important to point out. First, participants were allowed to keep all of the current patient information in front of them until a decision had been made. Second, if participants made errors or had questions about how to use the system, they were answered. Finally, to allow full concentration participants were not probed or required to think aloud during the data entry activity, and the interviewer did not intervene unless to provide clarification or assistance with the prototype.

### 3.4.9 NASA TLX

Once the participant had indicated their decision to accept or decline the suggestion, the case study papers were removed and they were handed a NASA TLX rating sheet, like the one shown in Figure 9. Participants were asked to use the sheet to evaluate the decision-making task they had just completed. Participants were allowed to refer to the subscale definitions provided to them, if necessary.



The form is titled 'Participant: \_\_\_\_\_', 'Task # \_\_\_\_\_', and 'Patient # \_\_\_\_\_'. It contains six subscale rating scales, each with a 10-point horizontal line and 'Very Low' and 'Very High' labels at the ends. The subscales are: 1. Mental Demand: 'How mentally demanding was the task?'. 2. Physical Demand: 'How physically demanding was the task?'. 3. Temporal Demand: 'How hurried or rushed was the pace of the task?'. 4. Performance: 'How successful were you in accomplishing what you were asked to do?' with 'Perfect' and 'Failure' labels at the ends. 5. Effort: 'How hard did you have to work to accomplish your level of performance?'. 6. Frustration: 'How insecure, discouraged, irritated, stressed, and annoyed were you?'.

**Figure 9: NASA TLX rating scale (NASA Human Performance Research Group, 1987)**

Once this was done, all papers were removed and a new case study task began. This process was repeated until all eight case studies were completed or until the participant wished to stop.

### 3.4.10 Post-Use Questionnaire

After completing the decision-making tasks and NASA TLX ratings, participants were handed the Post-Use Questionnaire. The questions were read aloud to the participant, who was then asked to indicate on the scale the

rating they felt best applied. As shown below, these questions have slight variations to Q1 and Q2:

Q3: How confident are you, when aided by this system, in your ability to diagnose patients with balance diseases?

Q4: How much do you trust this system to help you diagnose your patients with balance diseases?

*(The scales used for Q3 and Q4 match those used for Q1 and Q2, respectively.)*

Once the participant had indicated their rating for each question (Q3 and Q4), they were asked an open-ended question aimed at determining if the explanations had an impact on their confidence or trust.

The use of rating scales to measure trust and confidence are common within the field and have been used by researchers such as Lim and Dey (in 2011), Kulesza et al (in 2012), and Glass et al (2008). It has been shown that trust levels fluctuate over the course of an interaction (Dzindolet et al., 2003), suggesting that trust ratings – and confidence ratings – should be acquired after each decision-making task. However, research has also suggested that participant may simply rate their trust, confidence, or certainty based upon the presented System Certainty just seen in the task (Lim & Dey, 2011). By collecting ratings at only two points in the experiment rather than immediately following each case study task, we ensure that participants will not simply regurgitate the System Certainty displayed before them, as has happened in the past.

#### **3.4.11 Incentive for Participation**

All volunteers that participated in a research session were given compensation for their time spent. This compensation, and £8 Amazon gift voucher, was provided at the completion of the study session.

#### **3.4.12 Completion**

The study concluded with the post-use questionnaire and open-ended questions. Participants were thanked again for their involvement and received their incentive for volunteering. Recording was stopped and saved.

## **Section 3.5 Data Collection and Analysis**

To answer the research questions, both quantitative and qualitative data was gathered. The quantitative data was collected through the pre- and post-use questionnaire ratings and NASA TLX workload source scales. Verbal utterances recorded during the decision-making portion of each task as well as during two open-ended questions at the end of the study. These utterances include the decision to accept or decline a suggested diagnosis. The verbal utterances were transcribed in full and entered into an excel file along with the ratings and decisions made. The paper rating scales were digitally scanned for record-keeping and are included in Appendix I, along with the completed interview guides and questionnaires for each participant.

### **3.5.1 RQ1: How does explanation completeness impact clinical decision-making?**

One way of examining the impact of explanation completeness on clinical decision-making is to look at the number of 'right' and 'wrong' decisions made by our participants. This research study defines a 'right' decision as one in which a participant accepts a correct diagnostic suggestion or declines an incorrect one. A 'wrong' decision is one in which a participant either accepts an incorrect suggestion or by declining a correct one.

Every participant's decision to accept or decline a suggested diagnosis was recorded on the interview guide along with the patient's ID. After the study session was completed, the participant's decisions for each case study were entered into an Excel spreadsheet. The participant's ID and assigned group were listed, along with the order of patient case studies, whether the suggested diagnosis was correct or incorrect, whether the system certainty was high or low, and whether the participant accepted or declined the diagnosis. The decisions were then marked as 'right' or 'wrong' based upon the definitions stated above. The decisions were tallied and analysed graphically and through a Chi-Squared test of association. A screen shot of the Excel document is shown in Figure 10 and the spreadsheet is included in Appendix J.1.

ID	SYS	Case Study ID	Order	IN/CORRECT	H/LOW	RIGHT?	# Right	# Wrong	# Accept Correct	# Decline Incorrect	# Decline Correct	# Accept Incorrect
A01	A	G	2	C	H	R	6	2	3	3	1	1
A01	A	D	3	C	L	R						
A01	A	L	4	C	H	R						
A01	A	N	5	C	L	W						
A01	A	J	6	I	L	R						
A01	A	Q	7	I	H	R						
A01	A	R	8	I	L	R						
A01	A	K	1	I	H	W						
B01	B	G	2	C	H	R	4	4	4	0	0	4
B01	B	D	3	C	L	R						
B01	B	L	4	C	H	R						
B01	B	N	5	C	L	R						
B01	B	J	6	I	L	W						
B01	B	Q	7	I	H	W						
B01	B	R	8	I	L	W						
B01	B	K	1	I	H	W						

**Figure 10: Portion of Decision-Making Spreadsheet**

### **3.5.2 RQ2: How does explanation completeness impact a clinician's confidence in their ability to diagnose patients?**

The ratings from Q1 and Q3 of the pre- and post-use questionnaires were entered into an Excel spreadsheet alongside the participant's ID and assigned group. The recording of the verbal response to the open-ended question ("Do you feel that the explanations for why a diagnosis has been suggested have any impact on your confidence in diagnosing patients?") were transcribed in full and entered into the same spreadsheet.

To determine the change in confidence, the rating provided for Q1 (pre-use) was subtracted from the rating provided for Q3. The result was also included in the spreadsheet. These were compared between groups to determine if explanations of high-completeness lead to higher confidence.

The transcripts from the open-ended question were then reviewed to understand what the various participants thought about the explanations, and how they might have impacted their confidence.

No statistical tests were done on the confidence scores, given the low number of participants. The resulting spreadsheet is included in Appendix J.4.

### **3.5.3 RQ3: How does explanation completeness affect a clinical user's trust in a CDSS?**

RQ3 was analysed in much the same way as RQ2. The ratings from Q2 and Q4 of the pre- and post-use questionnaires were entered into an Excel spreadsheet alongside the participant's ID and assigned group. The recording of the verbal response to the open-ended question ("Do you feel that the explanations that were provided impact your trust?") were transcribed in full and entered into the same spreadsheet.

To determine the change in trust, the rating provided for Q2 (pre-use) was subtracted from the rating provided for Q4. The result was also included in the spreadsheet. These were compared between groups to determine if explanations of high-completeness lead to higher trust in the CDSS.

The transcripts from the open-ended question were then reviewed to understand what the various participants thought about the explanations, and how they might have impacted their trust.

No statistical tests were done on the confidence scores, given the low number of participants. The resulting spreadsheet is included in Appendix J.5.

### **3.5.4 RQ4: What is the impact of explanation completeness on a clinical user's workload?**

Each participant filled out the NASA Talk Load Index (NASA TLX) rating scale after each decision-making task. The researcher noted the participant ID, decision-making task number, and case study ID on each of these papers. After each study session was completed, the workload ratings were entered into an Excel Spreadsheet, along with the participant's ID, assigned group, the patient ID, system certainty, whether or not the suggestion was correct, and whether the participant made a right or wrong decision.

Each workload source (Mental Demand, Physical Demand, etc.) was given an equal rating of one. The sum of the workload sources was calculated for each task then divided by 15 as specified in the NASA TLX Paper and Pencil package (NASA Human Performance Research Group, 1987). The outcome of this equation was labeled as Workload. Figure 11, below, shows an example section of the spreadsheet, which is included in full in Appendix J.3.

The mean and standard deviation was calculated for each group's workload sources and overall workload. A Mann-Whitney test was performed on each of the sources and workload to determine if there was a significant effect caused by the completeness of the explanations.

In addition to the statistical analysis of the workload, the verbal decision-making transcripts associated with each participant's highest and lowest overall workload scores were analysed. These transcripts were reviewed with

a focus on finding commonalities in the decision-making behaviour to better understand what may be impacting the workload.

ID	SYS	Case Study ID	Order	IN/CORRECT	H/LOW	RIGHT?	Mental	Phys	Temp	Perf	Eff	Frustr	Workload
A01	A	G	2 C	H	R		3	4	11	4	4	4	2 Lowest
A01	A	D	3 C	L	R		10	11	11	8	7	14	4.0667
A01	A	L	4 C	H	R		4	4	7	15	8	4	2.8
A01	A	N	5 C	L	W		6	12	13	14	13	10	4.5333 Highest
A01	A	J	6 I	L	R		10	10	7	7	5	4	2.8667
A01	A	Q	7 I	H	R		13	6	9	10	12	13	4.2
A01	A	R	8 I	L	R		11	5	4	7	7	12	3.0667
A01	A	K	1 I	H	W		11	10	5	13	6	13	3.8667
B01	B	G	2 C	H	R		7	1	1	5	5	2	1.4
B01	B	D	3 C	L	R		6	1	1	5	5	1	1.2667
B01	B	L	4 C	H	R		5	1	1	5	4	1	1.1333
B01	B	N	5 C	L	R		6	1	1	3	2	1	0.9333 Lowest
B01	B	J	6 I	L	W		7	1	1	9	3	1	1.4667
B01	B	Q	7 I	H	W		5	1	1	12	3	1	1.5333
B01	B	R	8 I	L	W		13	1	1	12	11	3	2.7333 Highest
B01	B	K	1 I	H	W		12	1	2	7	9	5	2.4

Figure 11: Example section of Workload Rating Spreadsheet

### 3.5.5 RQ5: What information do clinician's desire from a CDSS's explanation when making a diagnostic decision?

All comments made during the decision-making task (from the point in which the participant entered the examination results into the system through to the point in which the participant indicated their decision to accept or decline the suggested diagnosis) were transcribed. These transcripts were entered into an Excel spreadsheet, each sentence occupying a new cell in the document. The participant's ID, assigned group, the case study details (case study ID, system certainty, correctness of suggested diagnosis), and the participant's decision were also included next to the transcripts. To create an accurate representation of the responses, the utterances of the participants was transcribed in full, with great care to use the exact words, punctuation, pauses, and actions of each participant – as recommended by Braun and Clarke (2006). Pauses are shown through ellipses while physical actions are encapsulated in asterisks. Questions or statements made by the researcher, when transcribed, are shown in italics. An example of this is shown below, in Figure 12.



ID	SYS	PT	I/C	H/L	R/W	Verbal
B03	B	N	C	L	W	So i'd need to know more about her history before I could accept this. I'd actually decline this because I don't know enough from her story.
B03	B	J	I	L	R	*laughs* Okay.
B03	B	J	I	L	R	Again, um, *looks at patient info* Again it's to do with the patient information I have here. Because I don't know whether this is new or if he's had previous episodes.
B03	B	J	I	L	R	Um, but, if he's already slightly disorientated, I would be very loathe to put it down to age-related imbalance. I'd decline that one.
B03	B	J	I	L	R	As for what info, I think, you see, that what it all comes down to is the definitions of these things, cause that's kind of at the crux of all these things. So what he actually means by dizziness and what he actually means by all these other bits and pieces and things. and also you've got that he's taking small steps, but what kind of steps? Shuffling? Small steps because.. I think there's just little bits of history that I need more information about. That would probably be one of those things that would lead to this.
B03	B	J	I	L	R	<i>So if the screen said the matches to his history, would that help?</i>
B03	B	J	I	L	R	Yeah, yes, I think so. That would be quite useful I think. So if it said something along the lines... Cause it just says "matches to the disease profile" and it doesn't.. Yeah. Yeah, I tell you what, if it had a bit more information about the disease profile because of the following and what matches within the history is leading it to that conclusion for that disease profile. So if it said 'in the history he's said he had a fall and those sort of things' so you get a little summary, so not just the exam results, but that's just probably because of the way I work.

**Figure 12: Example of transcripts**

The transcripts were coded using a thematic analysis. This research was approached without pre-identified themes to confirm or compare against, making it necessary to let the themes develop from the data. Thematic analysis was chosen as it is widely used for 'identifying, analyzing, and reporting patterns (themes) within data' (Braun & Clarke, 2006). Thematic analysis has also been used in related studies, such as Alberdi et al (2001) and Glass et al (2008). As described by Braun and Clarke, there are two approaches to thematic analysis: inductive thematic analysis is appropriate for instances in which 'the specific research question can evolve throughout the coding process' while a theoretical thematic analysis is more appropriate when coding 'for a specific research question' (Braun & Clarke, 2006). For this research, a theoretical approach was taken as the thematic analysis was used to provide answers to a particular research question, Research Question 5.

The process of thematic analysis began as patterns began to emerge from the verbal utterances were being transcribed, and evolved throughout repeated readings of the transcripts. Two over-arching themes were identified: the first being mentions or references to patient information, the second being additional explanations desired. The former included only information that was provided to the participant in the case study, leading to three specific codes: Medical History, Symptoms, and Examination Results. These codes were later divided between information shown (referring to information that was visually available in the explanation provided to the participant) and information not shown (referring to information that was

included in the case study but was not included in the explanation provided to the participant). The latter of the two over-arching themes – additional explanations - included any questions about how the system worked or requests for more explanation. The codes that were developed out of these statements were adjusted at various points throughout the thematic analysis to clarify meaning.

Appendix J.6 includes the full transcripts, while Appendix J.7 includes the code definitions and examples.

## Chapter 4 Results

### Section 4.1 Introduction to Results

This chapter presents the results of the research study described in Chapter 4. The results are presented in five sections, one for each research question. The following chapter (Chapter 6) discusses these results.

### Section 4.2 RQ1: How does explanation completeness impact clinical decision-making?

Research Question 1 (RQ1) is focused on examining the decisions clinical users make when using a CDSS and how they are impacted by the completeness of the system explanation. Thus, the decisions of interest are those that completed each case study task: the decision to accept or decline the diagnosis suggested by the prototype (System A or System B). To remind the reader, this research uses the terms ‘right’ and ‘wrong’ to define the decisions made, and uses the terms ‘correct’ and ‘incorrect’ to indicate to the reader whether the suggested diagnosis was a true positive (where the suggested diagnosis is the actual diagnosis) or false positive (where the suggested diagnosis is not the actual diagnosis). At the risk of oversimplifying, the results presented here examine the right and wrong decisions made by each group to determine if explanation completeness has an impact. The following paragraphs present the results, beginning with a review of the numbers of decisions made, then a Chi-Square test of association, and finally an examination of the percentages and frequencies of the decisions made.

A total of 52 decision-making tasks were completed by the participants of the study. The High Completeness group completed 28 (A01, A02, and A04 completed all eight tasks while A03 only completed four). The Low Completeness group completed 24 (B01, B02, and B03 each completed all eight tasks). The High Completeness group completed 28 decision-making tasks. Refer to Table 4 for the breakdown of correct and incorrect diagnostic suggestions, per group.

	Completed decision-making tasks with Correct Suggestions (True Positive)	Completed decision-making tasks with Incorrect Suggestions (False Positive)
High Completeness Group	N = 13	N = 15
Low Completeness Group	N = 12	N = 12

**Table 4: Number of Correct and Incorrect diagnostic suggestions, per group**

#### 4.2.1 Impact of Completeness on making 'Right' and 'Wrong' decisions

Of all the decisions made by the High Completeness group, 57% (16 of 28) were 'right' and the remaining 43% were 'wrong' decisions. The Low Completeness group's decisions were 58% (14 of 24) 'right' and 42% 'wrong'. Figure 13, a stacked bar graph, displays the percentages while Table 5 shows the frequencies.

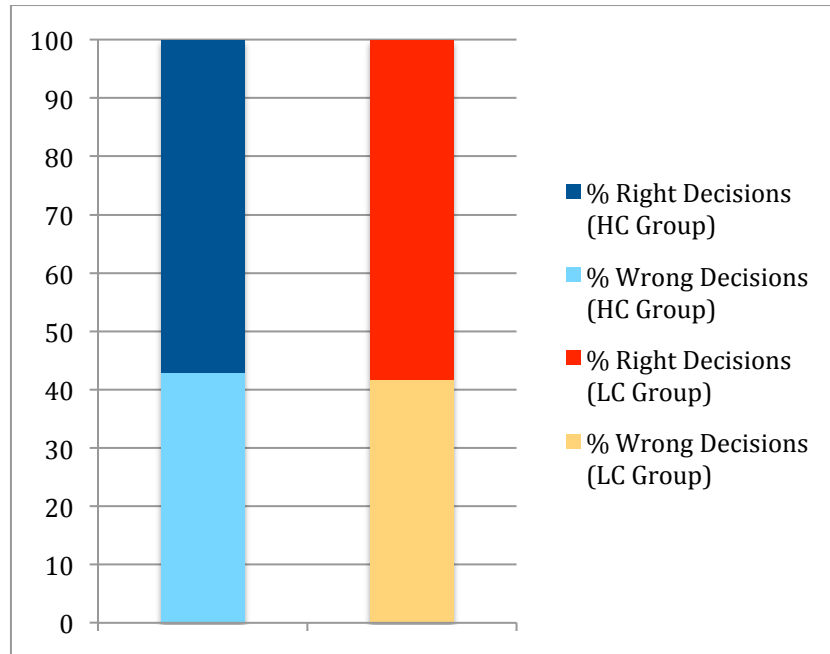


Figure 13: Percent of Right and Wrong decisions, by group

Group	'Right' Decisions		'Wrong' Decisions		Total decisions
	# Accept Correct	# Decline Incorrect	# Decline Correct	# Accept Incorrect	
High-Completeness	12	4	1	11	28
Low-Completeness	9	5	3	7	24

Table 5: 'Right' and 'Wrong' Decision Frequencies, per group

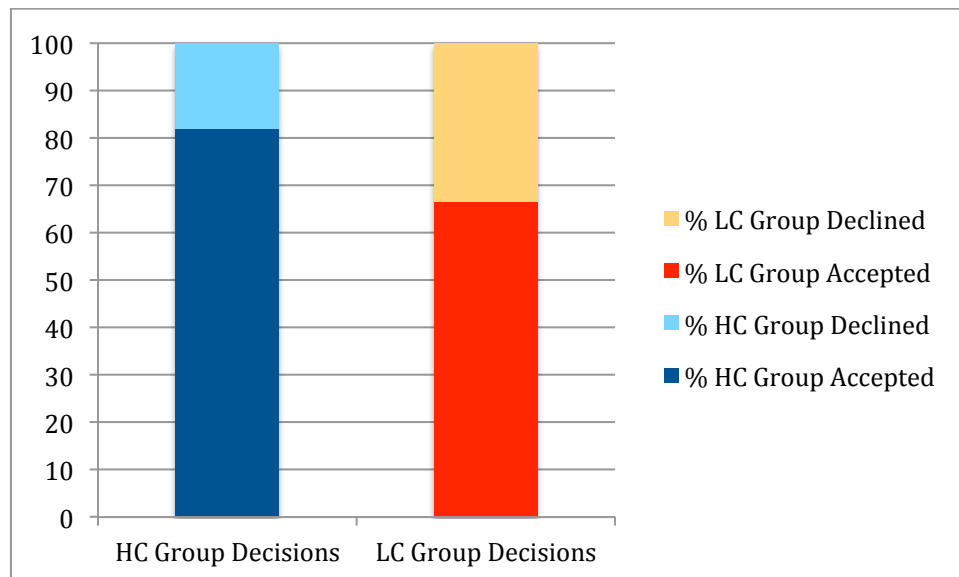
A Chi-Square Test of Association was performed on these frequencies to examine the relation between decisions made and explanation completeness.

The relation between these variables was insignificant ( $\chi^2(3, N = 52) = 2.13$ ,  $p = 0.54516$ ). The decisions made by clinicians are independent to the completeness of the explanation provided.

#### 4.2.2 Impact of Completeness on decision to accept or decline

The decision-making recorded during the study sessions must also be looked at for automation bias. To remind the reader, automation bias is an over-reliance on the system, and users with automation bias will frequently accept of a CDSS's suggestions, even when the suggestions are incorrect.

Figure 14 presents the percentage of suggested diagnoses that each group accepted and declined – regardless of whether or not the suggestion was correct. The High Completeness group accepted 82% of all of the CDSS's suggested diagnoses. The Low Completeness group, in comparison, accepted 67%. This means that the High Completeness group accepted 15% more of the suggested diagnoses than the Low Completeness group.



**Figure 14: Percentage of CDSS suggestions accepted or declined**

As shown in Table 6, the High Completeness group accepted 92% of the correctly suggested diagnoses, 'right' decisions. The Low Completeness group accepted only 75% of these. That is, the Low Completeness group wrongly declined *one quarter* of the correct suggestions made by the CDSS, while the High Completeness group declined 8%. However, when presented with an incorrect diagnosis the Low Completeness group made the right decision in 42% of the instances. The High Completeness group made 27%, meaning that this group accepted 15% more incorrect diagnostic suggestions than the Low Completeness group.

		<b>Correct Suggestions</b> (True Positive)	<b>Incorrect Suggestions</b> (False Positive)
<b>High Completeness Group</b>	% Accepted N = 23	<b>Accepted 92.3%</b> (Right decisions)	<b>Accepted 73.3%</b> (Wrong decisions)
	% Declined N = 5	<b>Declined 7.69%</b> (Wrong decisions)	<b>Declined 26.7%</b> (Right decisions)
<b>Low Completeness Group</b>	% Accepted N = 16	<b>Accepted 75%</b> (Right decisions)	<b>Accepted 41.7%</b> (Wrong decisions)
	% Declined N = 8	<b>Declined 25%</b> (Wrong decisions)	<b>Declined 58.3%</b> (Right decisions)

**Table 6: Proportion of 'Right' and 'Wrong' decisions**

To examine the relation between accepting/declining suggestions and explanation completeness, a Chi-Square test of association was performed.

The relation between these variables was insignificant ( $\chi^2(1, N = 52) = 0.93$ ,  $p = 0.3352$ ). The decision to accept or decline a suggestion is independent to the completeness of the explanation provided.

While the results of both Chi-Square tests show no significance, it is possible that a relationship between decisions and completeness would be seen with a larger sample size. On the surface it appears that there is no difference in the decisions made by the clinicians. However, it seems that the Low Completeness group acted more cautiously or, to put it another way, that the High Completeness group acted more boldly. The High Completeness group accepted 82% of the suggested diagnoses, 15% more than the Low Completeness group. This suggests that a more complete explanation can lead to automation bias.

## Section 4.3 RQ2: How does explanation completeness impact a clinician's confidence in their ability to diagnose patients?

Systems, such as the one prototyped for this research, often fail to be adopted when the users are more confident in their own abilities than they are in the system's (de Vries et al., 2003). Research Question 2 (RQ2) asks whether or not a more complete explanation improves a user's confidence. To answer this question, the pre- and post-use ratings of confidence are used to calculate the change in the user's confidence. The pre-use question acts as a benchmark – how confident is the clinician on their own? The post-use question measures how confident they are *when aided by the system*. Thus, one is able to ascertain whether or not the participants felt more confident in their own abilities or their own abilities combined with the system's. Then, the verbal responses to the open-ended question (*Do the explanations for why a diagnosis has been suggested have any impact on your confidence in diagnosing patients?*) are analysed by group to understand how different levels of completeness impact confidence.

### 4.3.1 Changes in Confidence

Figure 15 shows the pre- and post-use ratings of each participant's confidence in their ability to diagnose patients with balance disorders. The High Completeness group had two participants who rated their confidence as unchanged, and two who rated it as increasing by two points. This results in the High Confidence group as having an average increase in confidence of one point. The Low Completeness group had one participant whose confidence remained unchanged, one whose confidence rose by two points, and one whose confidence dropped by two points, giving the group as a whole an average of zero change in confidence.

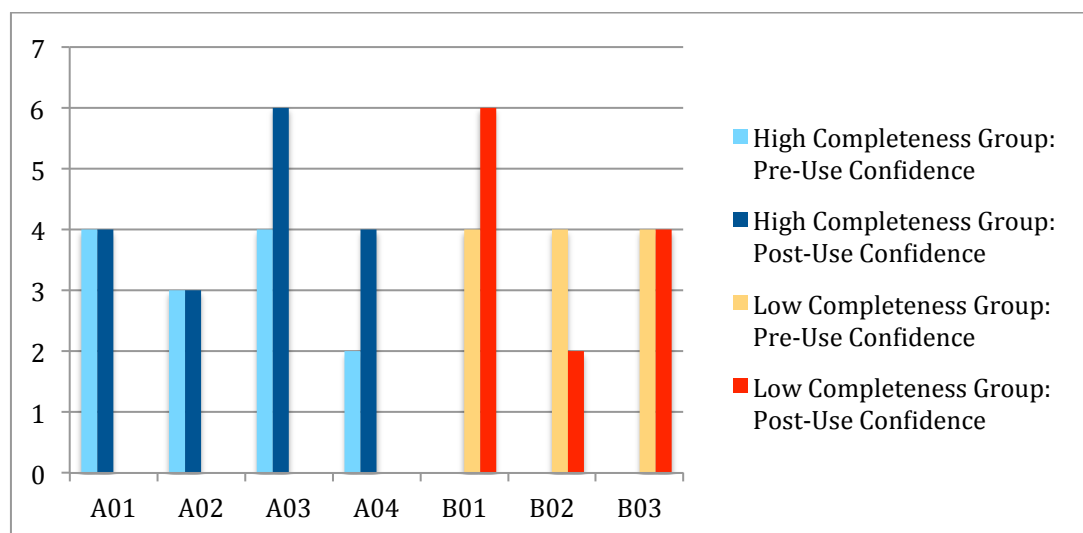


Figure 15: Pre-Use and Post-Use ratings of confidence

Group	Participant ID	Change in Confidence	Did explanations impact confidence?
High-Completeness Group	A01	0	No
	A02	0	Yes
	A03	+2	Yes
	A04	+2	Yes
Low-Completeness Group	B01	+2	Yes
	B02	-2	No
	B03	0	Yes (negatively)

**Table 7: Change in confidence and impact of explanations**

Table 7 shows the calculated change in confidence (post-use confidence rating minus pre-use confidence rating). Table 7 also shows whether or not the participant felt the explanation provided had any impact on their confidence: “Yes” means that the participant’s confidence was positively impacted, “No” means that it was not impacted, and “Yes (negatively)” means that their confidence was negatively impacted by the explanations.

With the low number of participants, it is difficult to confidently declare that a pattern emerges. However, these results do seem to suggest that a CDSS with explanations of high-completeness is more likely to raise a user’s confidence than one with explanations of low-completeness. The following two sub-sections include the responses to the open-ended questions and support this suggestion.

#### **4.3.2 Confidence and Explanations of High-Completeness:**

Participants A02, A03, and A04 stated that the explanations did impact their confidence. These participants said that the explanations provided them with the ability to reflect on what patient information they entered and decide whether or not they agree with the suggested diagnosis.

*“It’s really nice that you can look at all those things you found and look and match with the diagnosis, and think ‘Do I agree with that?’ So that definitely increases my confidence.” A03*



A04 said that the explanations “*definitely*” have an impact because they “*help you find the link*” between the suggested diagnosis and the patient’s information. While A02’s confidence score shows no change, he explains that it is due to his own abilities, not that of the system:

*“My confidence hasn’t changed because I don’t trust myself, having not read – brushed up – on my ENT.”* A02

Only participant A01 said the explanations had no impact on his confidence, stating that this was because they did not expand upon the clinical reasoning. This participant stated that the explanations seemed only to present the information that he had already input into the system.

*“[The explanations are] very much like when you’re a kid in school and you play that game where you match across the lines.”* A01

While A01 found the explanations too simple, his confidence score was not lowered after using the system. Indeed, none of the participants of the High Completeness group showed that their confidence was lower when aided by the system. From this, one can surmise that explanations of high completeness are not detrimental to a user’s confidence.

#### **4.3.3 Confidence and Explanations of Low-Completeness:**

The verbal responses provided by the Low Completeness group suggest that the explanations they received were not adequate and did have an effect on their confidence when using the system.

Participant B01, the only participant in this group whose confidence increased (by 2 points), said that a more complete explanation would be preferred:

*“Yeah, they do [have an impact]. I wish they were a bit more expansive, I will say. I mean, that bit at the end. Because I’ve seen the whole process through then I have more confidence in it, but if I was seeing it for the first time and that [explanation] was all I saw then I wouldn’t trust it at all. But because I’ve seen it through, it’s fine.”* – B01

Participants B02 and B03 also refer to the lack of information provided by the explanations. B02, whose confidence rating dropped by two points, suggested that the lack of information in the explanation meant that the system didn’t consider anything but the examination results:

*“No [they didn’t impact my confidence]. Because, um, the only things that were listed were the examinations and it needs to take into account the history as well.”* – B02

Participant B03 provided a very similar response:

*“Yeah, it does [impact my confidence], because it’s focusing solely on the investigation, the examinations, whereas the way that I work – I focus more on the history, on the story that they give me... So it does – it only says exams – so it does lower my confidence.” – B03*

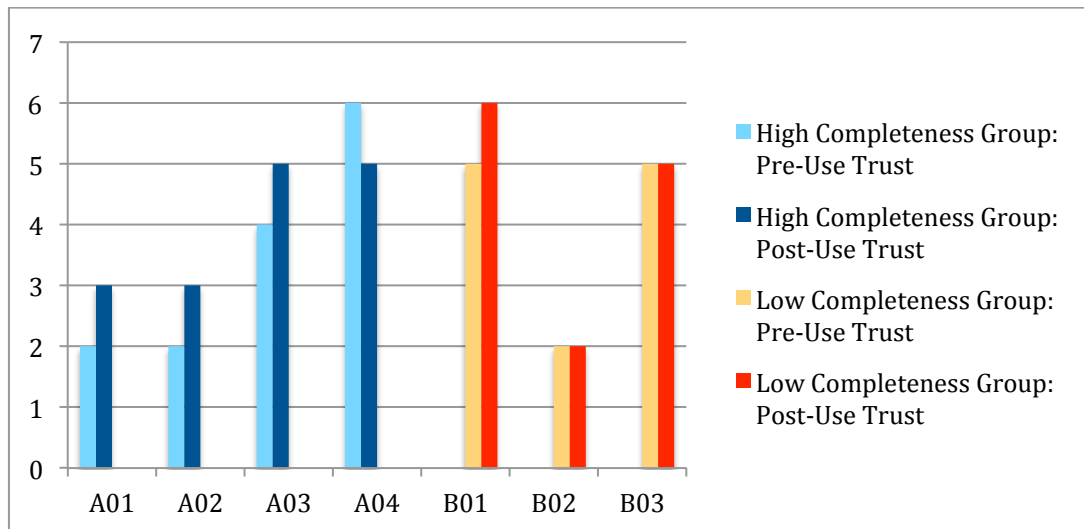
All three participants of the Low Completeness group clearly stated that they desired a more expansive, or complete, explanation. Comparing these responses to those provided by the High Completeness group, the results suggest that explanations with high completeness have a positive impact on a clinician’s confidence in their ability to diagnose patients with balance diseases. The results also suggest that explanations with low completeness have little or no effect on a clinician’s confidence. With this, one could argue that a more complete CDSS explanation leads to a more confident clinical decision-maker.

## **Section 4.4 RQ3: How does explanation completeness affect a clinical user’s trust in a CDSS?**

Do explanations of high completeness have a more positive impact on a user’s trust in a CDSS than an explanation of low completeness? Does completeness not matter at all, when it comes to trust? To answer RQ3, the pre and post-use ratings of trust are used to calculate the change in the user’s trust in a CDSS, from what was hypothesized (pre-use) to what resulted (post-use). The verbal responses to the open-ended question (“Do you feel that the explanations that were provided impact your trust?”) are analysed by group to understand how different levels of completeness impact trust in a CDSS.

### **4.4.1 Changes in Trust**

Of the four participants in the High Completeness group, three rated their trust increased by one point after using the system. The fourth member of the group indicated a decline in trust, dropping from 6 to 5 after using the system – the only participant in the study to have reduced trust. Within the Low Completeness group only one participant showed an increase in trust; a one point increase from a rating of five to a rating of six. B02 and B03 remained unchanged. Refer to Figure 16 for a bar chart showing the pre- and post-use ratings of each participant. Table 8 shows the calculated change of trust in a CDSS, as well as whether or not the participant felt the explanations provided had an impact on their trust.



**Figure 16: Pre-Use and Post-Use Ratings of Trust**

Group	Participant ID	Change in Trust	Did explanations impact trust?
High-Completeness Group	A01	+1	No
	A02	+1	Yes
	A03	+1	Yes
	A04	-1	Yes
Low-Completeness Group	B01	+1	Yes
	B02	0	Yes (negatively)
	B03	0	Yes

**Table 8: Change in trust and impact of explanations**

The results of the ratings suggest that explanations of high-completeness have a positive impact on a user's trust in a CDSS. The following two subsections explore each group's response to the open-ended trust question, providing a more rich understanding of how the explanations have an impact.

#### 4.4.2 Trust in a CDSS with Explanations of High-Completeness:

*"I guess this thing knows more than me. The system knows more than me. I'll accept."* – A02 displaying a high level of trust in the CDSS during a decision-making task

Three of the four participants (A01, A02, and A03) in the High Completeness group had an increased trust in the system after using it. A02, A03, and A04 all stated that the explanations did improve their trust, while A01 stated that they didn't impact his trust at all.

For A01, it seems, the explanations weren't really a consideration to base trust upon, it was more about the system being validated:

*"I don't distrust it completely, I just think if I knew that this system had been used in an evidence-based trial to show that it was equally effective as the opinion of an expert then I'd have more faith in it."* A01

While A01 did not state that the explanations played a role in improving his trust, he also did not state that they had a negative impact. A02 and A03, on the other hand, did not focus on whether or not the system had been validated.

*"My trust in the system has improved a bit. One step further,"* A02 stated while indicating his rating.

A03's response was similar to A02's: *"They do impact my trust. ...It seems to me like it follows the system of, you know, the same system of decision-making that we use. And examinations. It seems really good."*

Both A02 and A03 acknowledged that the CDSS provided some incorrect suggestions, mentioning a particular case study task that they struggled with. For A02, it was the suggested diagnosis of a Vestibular Schwannoma (an incorrect suggestion or 'false positive'). During that particular decision-making task, he stated that there was no way the system could come up with that diagnosis based upon the limited patient information he had input. When asked if the explanations impacted his trust, he replied *"Yeah. Like the Schwannoma thing. That was a big one."* A03 had a similar incident when attempting to determine if a patient's heavy drinking was the true cause of some of the symptoms. Like A02, she mentions this when explaining her trust:

*"You still have to use brain work and knowledge and skills. You still have to look and think. Like with the alcohol."* –A03

However, these two participants stated that the system is not designed to be blindly trusted; it is intended to provide support to a clinician, rather than replace them.

*"I don't distrust it – my trust has actually improved a bit because some things it gets right and others it doesn't get right. Well, that should make it worse. But these systems are not going to be used in isolation. So it's aiding."* A02

A04, despite the one point drop in trust, also said that the explanations make him trust it more.

*“More, [the explanations] make me trust it more. It’s proof that the system is matching the information with digital medical knowledge.” - A04*

The certainty percentage that was shown also seems to have impacted A04’s trust:

*“There is this degree of certainty that makes me trust more in this system because I can see the link. It’s not just like drawing dots, it’s linking the test results with the source in the medical diagnosis.”-A04*

Both A04 and A03 seemed effected by the completeness of the explanations, not just the presence. Their statements suggest that the more complete explanations are indicative of the technological logic that the system uses to provide a suggested diagnosis.

*“[The explanations] do impact my trust. ...It seems to me like it follows the system of, you know, the same system of decision-making that we use, and examinations. It seems really good.” A03*

*“There’s a link behind that – an algorithm them links with my patient’s information. So there is an algorithm that says the latest research in medical knowledge – this test is positive, so literature says that this test is related to specificity or sensitivity with this disease.” A04*

The responses provided by the High Completeness group are highly suggestive that the explanations they received had a positive impact on their trust in the system.

#### **4.4.3 Trust in a CDSS with Explanations of low-completeness**

The explanations provided to the participants of the Low Completeness group had an interesting impact on their trust. Two of the three participants showed no change in trust levels, while the third had an increase in trust of one point.

Participants B01 and B03 replied that they feel that the system is a helpful tool to use, given their limited experience on the domain of balance disorders.

*“Um, the only thing really impacting my trust is the fact that I know that I don’t know that much about balance disorders” - B01*

While the suggestions put forth by the CDSS may sometimes be incorrect, these participants did not find that to be particularly negative:

*“It definitely throws out things. There were maybe one or two that it threw at that weren’t at the top of my list, but it is actually quite helpful as a differential type thing. ...But I suppose it’s kind of helpful in that kind of way, just to see what it is that - especially if you’re really stuck with somebody tough - so that’s quite helpful.” - B03*

Whether or not the explanations impacted the trust of these two participants is not quite clear – it seems, from the transcripts, that the explanations were

secondary in importance to the system's ability to provide a diagnostic suggestion:

*"So I'm glad that the system is able to help me think of things that I wouldn't have thought of in the first place. So I'm glad of that. But the explanation at the end helped, I just need more of it really."* - B01

B02's trust rating did not change after using the system and his explanation for this indicates that it may be due to his disposition towards trusting such a system:

*"I trust very little in any system to help me diagnose anything. So, this one in particular is not anything against this system."* – B02

While this may be true, it does appear as though the explanations provided to the participant had a negative impact:

*"Also, with this, I know it doesn't take into account the things that I would have been looking for, for example the clinical details that I would have thought were relevant. So it doesn't take that into account so I know it's not thinking along the same lines that I am."* –B02

With only the examination results shown in the explanations seen by B02, it seems that he formed an incorrect mental model of how the system works; thinking that the examination results were the sole basis upon which the system formed a diagnostic suggestion. As he himself said, this process of reaching a diagnosis does not align with his own.

It appears as though explanations of high-completeness do have a positive impact on a clinical user's trust in a CDSS. As for the participants in the Low Completeness group, it seems that the explanations provided to them did little to improve their trust. Indeed, one participant's trust in the system was negatively impacted because he perceived the system's reasoning did not emulate his own.

Participants in the High Completeness group felt that the explanations showed the system's reasoning; that the patient information listed in the explanation showed the links made by the CDSS internally. The participants of the Low Completeness group felt that the explanations provided were not as helpful as they could be; the explanations suggested that the system's reasoning process did not reflect their own. The implications of this are discussed in Chapter 5.

## Section 4.5 RQ4: What is the impact of completeness on a clinical user's workload?

Research Question 4 (RQ4) is aimed at determining if the completeness of an explanation has an effect on a clinician's workload. While a more-complete explanation provides more information, does the benefit outweigh the toll that it may take on workload? As Kulesza et al wrote, "information comes at the price of attention – a user's time (and interest) is finite, so the solution may not simply be 'the more information, the better'" (Kulesza et al., 2013).

The NASA TLX ratings that were collected after each decision-making task were used to calculate the overall workload score. A Mann-Whitney test was applied to determine if there was a significant effect of completeness, the results are presented in Section 4.5.1. Sections 4.5.2 and 4.5.3 explore the decision-making tasks associated with the lowest and highest workload ratings of each participant. This was done to determine if there were any common factors across the decision-making tasks that led participants to their highest or lowest workload scores.

### 4.5.1 Impact of Explanation Completeness on Workload

The workload score for each decision-making task completed by a participant was calculated using the method recommended in the NASA TLX package:  $([\text{Mental} + \text{Physical} + \text{Temporal} + \text{Performance} + \text{Effort} + \text{Frustration}] / 15 = \text{Workload})$  (NASA Human Performance Research Group, 1987). Table 9 shows the mean and standard deviation (SD) of each group's source ratings (Mental Demand, Physical Demand, etc.) and workload.

		Mental	Physical	Temporal	Performance	Effort	Frustration	Workload
High Completeness Group (n = 28)	Mean	7.39	3.89	7.42	7.28	6.86	7.86	2.71
	SD	5.29	3.17	4.78	5.14	5.00	5.01	1.27
Low Completeness Group (n = 28)	Mean	7.50	7.45	3.66	8.58	7.38	6.75	2.76
	SD	3.48	6.23	1.99	4.76	4.24	5.74	1.45

Table 9: Mean and Standard Deviations of Workload Sources and Workload

A Mann-Whitney Wilcoxon test was performed on the ratings of each group's participants to determine if there was a significant effect of explanation completeness on the sources and workloads. The results are shown below:

**Mental Demand:**

There is no significant effect of completeness  
( $W = 287.5$ ,  $p = 0.3756$ ) (Figure 17)

**Physical Demand:**

There is no significant effect of completeness  
( $W = 246.5$ ,  $p = 0.09841$ ) (Figure 18)

**Temporal Demand:**

There *is* a significant effect of completeness  
( $W = 479.5$ ,  $p = 0.008023$ ) (Figure 19)

**Performance:**

There is no significant effect of completeness  
( $W = 285.5$ ,  $p = 0.3542$ ) (Figure 20)

**Effort:**

There is no significant effect of completeness  
( $W = 298$ ,  $p = 0.489$ ) (Figure 21)

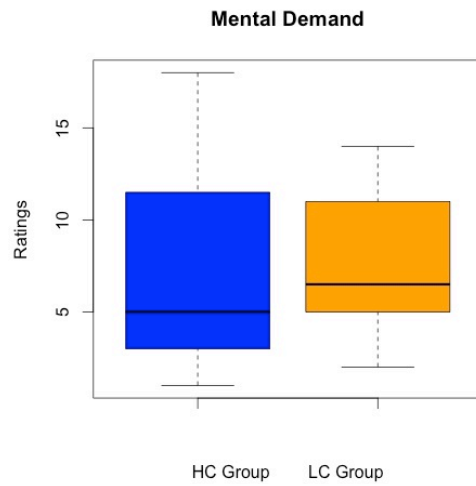
**Frustration:**

There is no significant effect of completeness  
( $W = 381.5$ ,  $p = 0.4064$ ) (Figure 22)

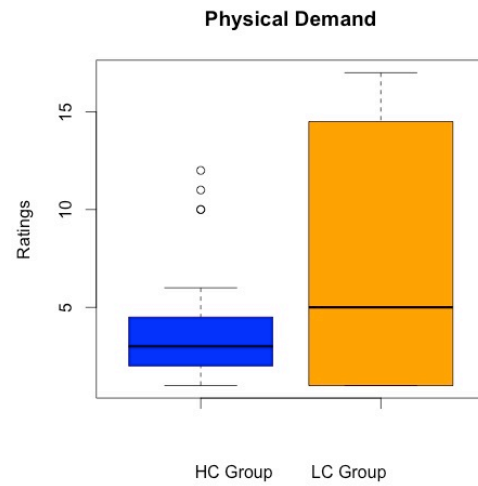
**Workload:**

There is no significant effect of completeness  
( $W = 332.5$ ,  $p = 0.9561$ ) (Figure 23)

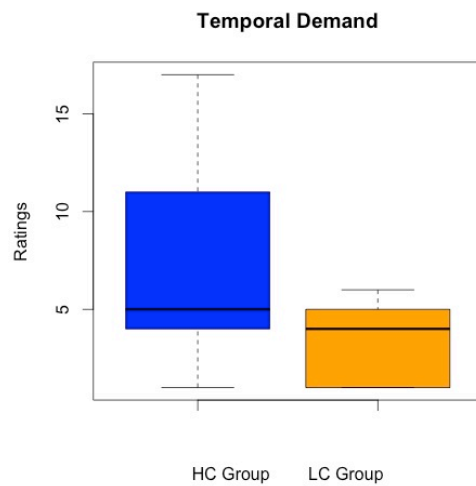




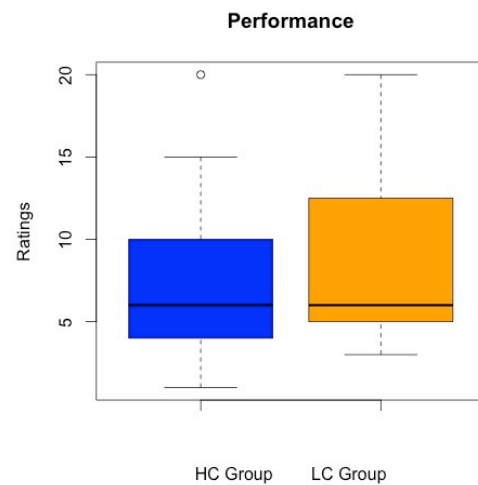
**Figure 17: Mental Demand of HC Group and LC Group**



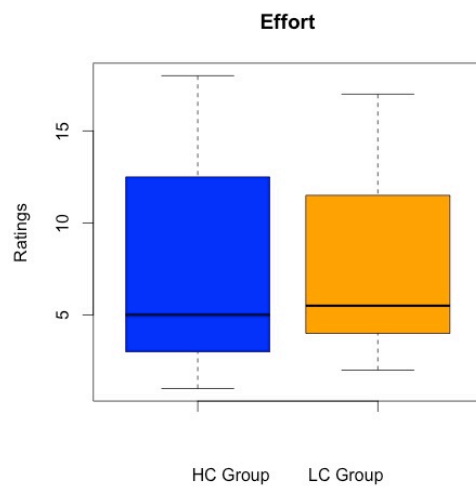
**Figure 18: Physical Demand of HC Group and LC Group**



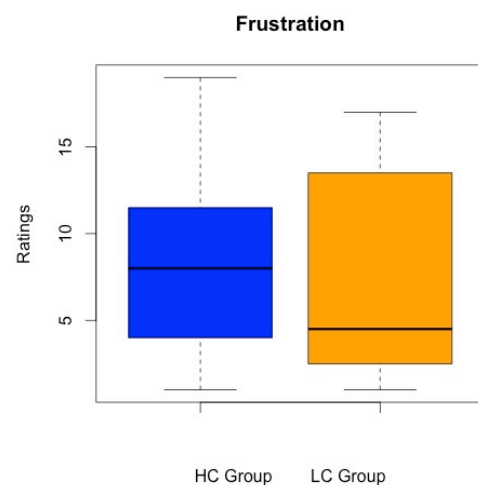
**Figure 19: Temporal Demand of HC Group and LC Group**



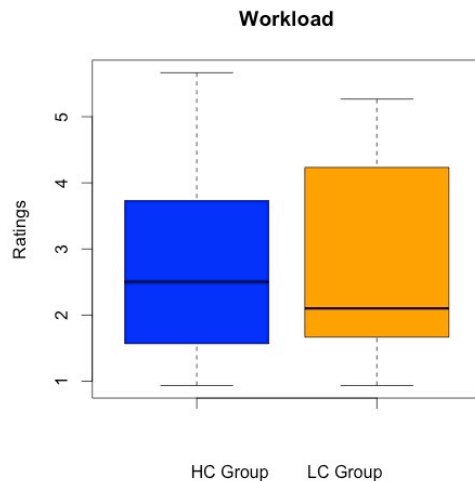
**Figure 20: Performance of HC Group and LC Group**



**Figure 21: Effort of HC Group and LC Group**



**Figure 22: Frustration of HC Group and LC Group**



**Figure 23: Workload of HC Group and LC Group**

The results of the Mann-Whitney test showed that completeness had a significant effect on Temporal Demand. The description of Temporal Demand provided by the NASA TLX instructions reads as follows:

*‘How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?’*

(NASA Human Performance Research Group, 1987)

Again, the resulting *p-value* of the Mann-Whitney test is 0.008023, or  $p < 0.05$ . The mean values of rated Temporal Demand were 7.42 and 3.66 for the High Completeness and Low Completeness groups, respectively. That is, the High Completeness group rated the Temporal Demand two times higher than the Low Completeness group. These results indicate that explanations of high-completeness are more demanding of a user’s time and explanations of low-completeness. However, the Mann-Whitney test showed no significant impact on Workload.

#### **4.5.2 Analysis of Tasks Associated with Lowest Overall Workload**

The lowest workload ratings from each participant are shown in Table 10, along with the details of the associated decision-making task (Suggestion, Certainty, and Decision). Six of the suggested diagnoses were correct, one incorrect. All but two were presented with a high-certainty percentage. Only one of the seven decisions made was incorrect (wrong).

Group	Participant ID	Suggestion	Certainty	Decision	Mental	Physical	Temporal	Permanence	Effort	Frustration	Workload
High Completeness Group	A01	C	H	R	3	4	11	4	4	4	2
	A02	C	H	R	2	2	3	1	3	3	0.93
	A03	I	H	W	1	1	17	1	1	1	1.46
	A04	C	L	R	5	1	1	5	3	6	1.40
Low Completeness Group	B01	C	L	R	6	1	1	3	2	1	0.93
	B02	C	H	R	2	17	5	5	3	11	2.86
	B03	C	H	R	6	5	3	5	4	4	1.80

**Table 10: Details of Lowest Overall Workload, by Participant**

**NOTE:** Suggestion: C = Correct, I = Incorrect. Certainty: H = High, L = Low. Decision: R = Right, W = Wrong.

The majority of the decision-making tasks that were rated to have the lowest workload were ones in which the participants rapidly agreed to the suggested diagnosis. Participants A01, A02, B01, B02 and B03 responded immediately that they would accept the suggested diagnosis.

*“Yeah, I’ll go with that.”* B02

*“Yeah, I’ll accept that. Partly because that’s what I was thinking.”*

A01

Before clicking diagnosis, A01 said: *“I think I know what this will be”* and after clicking it: *“Yeah, I accept this diagnosis.”*

*“Yeah. Cool. It’s only 19% sure. But I’m sure just because of that triad.”* B01

The transcripts present a pattern amongst these instances. The participants accepted the suggested diagnosis without deliberation. It seems that, in these instances, the participants had pre-formulated a diagnostic hypothesis whilst entering the patient details into the system. Even B01, who was faced with a low percentage of system certainty, was confident in the answer and immediately agreed with the suggested diagnosis, because of his own confidence. Thus, these participants found it easy to accept the suggestion because it aligned with their own thinking. This suggests that the workload was lighter for these tasks because there was no decision-making required, the system had validated the hypothesized diagnosis of the participants.

Only participants A03 and A04 show different behaviour during their decision-making process.

A03's lowest total workload rating occurred with the case study for Patient K. In this instance, the verbal utterances of A03's deliberation to accept or decline the diagnosis suggest that the decision was not an easy one to make. For example, A03's first response to viewing the suggested diagnosis:

*"Hmm. Well, I was kind of wondering whether it has to do with this person drinking lots of alcohol, because it's kind of difficult to say if the alcohol is affecting this person."* A03

The participant decided to accept the diagnostic suggestion in the end, stating that in a non-hypothetical situation she *"might query something else."* While A03 indicated a high temporal demand, she did not indicate any other negative impact to her workload.

A04's lowest workload score, 1.40, was given for a case study that presented low-certainty. Whilst the participant questioned the basis upon which the system determines its certainty, A04 found the suggested diagnosis to be *"plausible"* and decided to accept it based on the symptoms and examination results shown in the explanation.

This analysis of each participant's lowest overall workload suggests that the explanations were hardly used during these decision-making tasks, if they were used at all. If this is the case, then these scores do not reflect the impact of the explanation itself, but rather the ease of the decision to agree or disagree with the suggested diagnosis.

#### **4.5.3 Analysis of Tasks Related to Highest Overall Workload**

Table 11 presents the details of each participant's highest overall workload score. Two of the diagnoses were incorrect suggestions. All but one was presented with a percentage of low-certainty. Finally, only two of the decisions made were correct.

Group	Participant ID	Suggestion	Certainty	Decision	Mental	Physical	Temporal	Permanence	Effort	Frustration	Workload
High Completeness Group	A01	C	L	W	6	12	13	14	13	10	4.53
	A02	C	L	R	17	4	4	10	4	11	3.33
	A03	I	H	W	3	3	16	1	1	9	2.2
	A04	C	L	R	18	2	8	20	18	19	5.67
Low Completeness Group	B01	I	L	W	13	1	1	12	11	3	2.73
	B02	C	L	W	3	16	6	20	17	17	5.26
	B03	C	L	W	10	8	4	10	7	4	2.86

**Table 11: Details of Highest Overall Workload, by Participant**

**NOTE:** Suggestion: C = Correct, I = Incorrect. Certainty: H = High, L = Low. Decision: R = Right, W = Wrong

The patterns of behaviour observed during these decision-making tasks were very different from that which was described in the previous sub-section. For three of the seven instances, the transcripts show that the participants declined the suggested diagnoses because they preferred to conduct additional examinations or gather more patient information before accepting. With another three, the participants struggled with the decision to accept or decline because they had limited knowledge about the examinations, results, and/or suggested diagnosis. In the remaining instance, the participant (B01) was fatigued and relied on the computer to determine the diagnosis.

Participants A01, B02, and B03 declined the suggested diagnoses. Interestingly, the three diagnoses that were suggested were correct but shown with a low system certainty. These three participants drew upon their clinical experience; rather than accepting or declining the suggestion outright they formed a list of differentials and seemed to establish where the suggested diagnosis sat hierarchically on the list. Running through the patient information, the participants tried to rule out the other differential diagnoses. However, with what patient information they were given, these participants were not able to comfortably rule out the differential diagnoses. Thus, the suggestion was declined.

*“I treat Meniere’s as a diagnosis of exclusion, and she’s got fluctuating hearing loss so I’d want to get some audiology results before to see... I could do some simple tuning fork tests ion the*

*room, and I'd probably send her off to get some audiology before I made the diagnosis of Meniere's. Because is it... Does she have sensorineural hearing loss or conductive hearing loss or is it not a significant thing? So I don't think I would accept that diagnosis yet. So decline."* A01

*"Vestibular Migraine is much further down on the list of balance disorders – they're not the most common thing, they're not the first thing you think of when you think of balance disorders. Only about 5% would be vestibular migraines, so I have to think about how to rule out the top 40% which would be viral causes, and the next 20% which would be Meniere's. So that's the thing."* B02

B03 did not completely disregard the suggested diagnosis, saying that *"it's a differential diagnosis – it's a possibility"* but that she would want more information on the patient before accepting the diagnosis. *"Also she's got problems with her hearing, I would also wonder about things like Acoustic Neuroma. That would be worth looking at."* In the end, B03 declines the diagnosis; *"I'd need to know more about her history before I could accept this. I'd actually decline this because I don't know enough from her story."*

In the instances in which participants A01, B02, and B03 had their highest workload, it was because their clinical experience led them to disagree. The situation with participants A02, A03, and A04 was quite the opposite. These participants seem to have less clinical experience. They were unable to remember the details of the suggested disease, unfamiliar with the examinations, and uncertain of the clinical significance of the exam results.

*"I can't remember how to diagnose Meniere's disease. The symptoms are very relevant, the exam results, very relevant. Probably means that he does not have a balance disorder as such. But I cannot remember if Meniere's is a balance disorder – I mean vestibular disorder - or neuritis."* A02

*"Yeah, maybe I'm not familiar with these tests. They're not tests that I haven't done or don't know about. I know about head thrust. I know Romberg. I know horizontal. It's this one... Or maybe what it means is that I don't know what it means if most of them have positive positional nystagmus. I think that's what I'm trying to say. So I don't know."* A03

*"This is actually the part where I raise my hands because I have very rarely heard of this, so I don't know if it is just my lack of knowledge or if it is something that is not so common."* A04

Unlike the other six participants, B01 had a completely different experience with the decision-making task that he felt had the highest workload. B01 did not have a particular diagnosis in mind. *"I mean, he's a young guy, there*

*shouldn't be that much wrong with him. I don't know. I'm getting tired as well, so let's see what the computer says,"* B01 said before clicking 'Diagnose'. Upon seeing the suggested diagnosis, B01 was quick to agree, citing the physiological factors that supported it. *"Sure. Yeah. Vestibular Neuritis. So he's young, having an infection in his vestibular nerve, I can definitely see that rather than any kind of growth – that makes sense."* So, while he didn't struggle to determine whether or not he agreed, he did put in additional effort than in case study N; finding supporting evidence that was not present on the screen. Thus, high increases in his Mental, Performance, and Effort ratings.

The results presented in Section 4.5.2 suggest that when a clinical user's hypothesis is in agreement with the CDSS's suggestion the decision-making task is made simple, thus removing the need to refer to the explanation for guidance. The results of Section 4.5.3 do not indicate that the overall workload is higher due to the information (explanation) that was shown. Instead, the results indicate areas where additional information could improve the decision-making process and reduce the user's workload.

Finally, while explanations of high-completeness significantly affect a user's Temporal Demand (in comparison to explanations of low-completeness), the overall Workload of both groups is considerably similar.

## **Section 4.6 RQ5: What information do clinicians desire from a CDSS's explanation when making a diagnostic decision?**

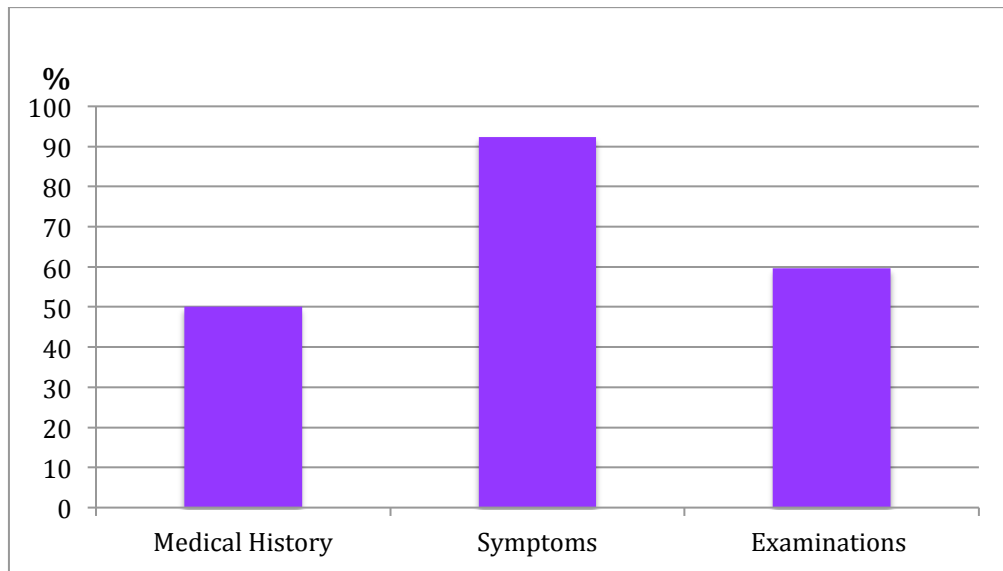
Research Question 5 focuses on the content that should be included in a CDSS's explanations. By examining the patient information that was used during the decision-making tasks, one can determine if the provided high-completeness explanations are appropriate or if the low-completeness ones would suffice. These results are presented in Section 4.6.1. Section 4.6.2 then presents additional information desired by the participants, information that was not included in the description of the patient.

### **4.6.1 Explanation Completeness, the Necessary Patient Information**

For the sake of clarity, the patient information that was used during the decision-making tasks will be referred to as information that was 'mentioned'. That is, 'mentioned' information is that which was provided to the participant in the case study, and was then verbally referred to by the participant during their decision-making.

Of the 52 completed decision-making tasks there were 51 in which participants mentioned a piece of patient information that was provided in the

case study description. The case studies divided the patient information into three categories; Medical History, Symptoms, and Examination Results. Of course, some categories were mentioned more often than others. As shown in Figure 24, the Medical History was mentioned in 48% of all the decision-making tasks, Symptoms mentioned in 92%, and Examination Results in 59%.

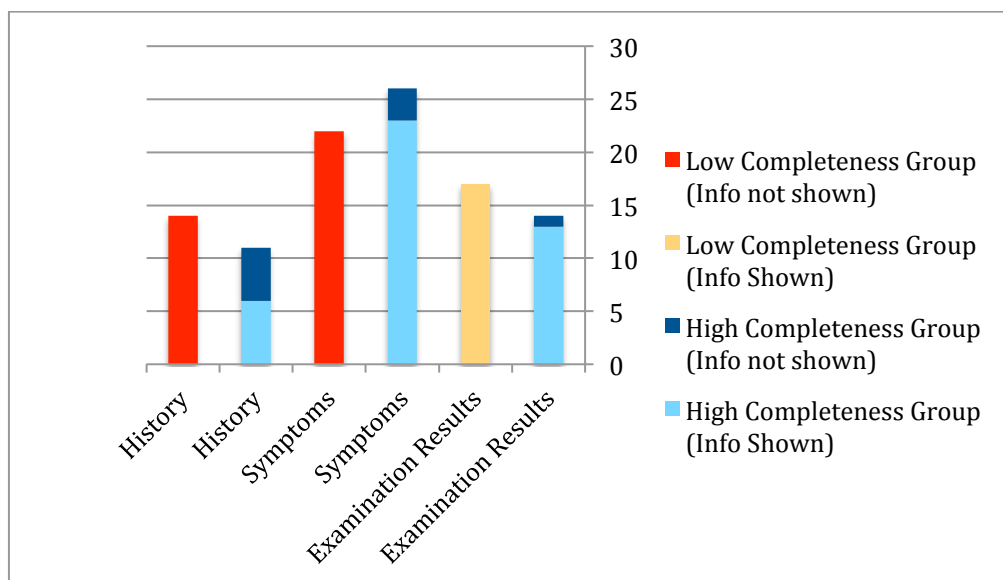


**Figure 24: Percentage of decision-making tasks in which Medical History, Symptoms, and/or Examinations were mentioned**

However, not all of the information mentioned was present at the time of the decision-making task. To remind the reader, the explanations provided were *Why...* explanations, meaning that only the patient information that matched the disease profile were shown. Additionally, the explanations of low-completeness only included the matching examination results, not the medical history and symptoms that the high-completeness explanations contained.

Figure 25 displays the frequency of decision-making tasks in which each category was mentioned. Those tasks in which patient information was mentioned but was not included in the explanation are highlighted and referred to as 'Info not shown'.





**Figure 25: Frequency of references shown/not shown, by group**

It is perhaps unsurprising to find that the participants of the Low Completeness group still mentioned the patient's symptoms and medical history during their decision-making despite the fact that these categories were not included in the explanations. However, it did come as some surprise to find that participants of the High Completeness group did the same.

While these results indicate what information a clinician uses during a diagnostic decision, an analysis of the verbal utterances during these instances reveals whether or not these details are desired in an explanation. The following three subsections present the results of this analysis.

#### 4.6.1.1 Medical History

The patient's medical history was mentioned in 48% of all the decision-making tasks, a frequency of 25 out of 52. Fourteen of these came from the Low Completeness group, which was not provided with any medical history in the explanations. Of the eleven references from the High Completeness group, five referred to information that was not provided in the explanations. The medical history included a variety of information (patient age, ability to work, smoking habits, etc.), and perhaps not all of that information is necessary to include in an explanation. Again, there were 25 mentions of the Medical History. Of these, there were 6 different pieces of patient information that were brought up; Effect on Occupation, Medication, Well-Being, Smoking/Drinking, Gender, and Age. Table 12 displays the frequencies while a discussion of each follows.

	Effect on Occupation	Medication	Well-Being	Smoking and Drinking	Gender	Age
Frequency of Mentions	3	4	5	9	13	18

**Table 12: Frequency of mentions to various Medical History information**

### Effect on Occupation

The information that was least-frequently mentioned was the impact that the patient's disease was having on their life or, more specifically, their ability to work. This was presented in the high-completeness explanation for six of the eight case studies.

*"Ability to work not affected by vertigo, the triggering event, yeah."* A01

*"Although she's not working and she's retired, I mean, she's getting significant amounts [of vertigo attacks]..."* B03

The effect of the disease on the patient's lifestyle were mentioned a total of three times. The statements suggest that this information is not of great importance during a diagnostic decision-making task. Perhaps participant A02 best explains why this information lacks clinical importance;

*"Ability to work, as a medical history, I don't know how much... Because that's subjective. I've never used it as a way of doing how serious a disease is, because there are other more objective parameters to use."* A02

### Medication

In all of the case studies, the participants were told that it was 'not known' what – if any – medications were being taken by the patient. However, there were four case study tasks in which the patient's prescriptions were mentioned. All of these came from one participant, A04. This participant felt that a patient's medication was important to consider and – as it was absent from the explanation – should be included along with the diagnostic suggestion to serve as a reminder to the clinician.

*"I would add the fact that she's not on any medication. I put it in, but I would put it here on the screen; 'Not on any medication, not known treatment, not undergoing any treatment right now.' Just the fact that*

*they are not taking any medication, a reminder, because I might take that for granted after a long day.” A04*

## **Well-Being**

The patient’s recent well-being was mentioned in five of the references to medical history. These references were to stress-levels, previous illnesses, and menopause. Interestingly, none of this well-being information (stress, menopause, or recent illnesses) was presented in the explanation for either group.

The participants seem to place value on this information; rather than glazing over it like the effect on occupation, the participants included it in their diagnostic consideration. For example, one of the patients (Patient R) was described as being highly stressed due to recent unemployment. This information was not included in the explanation for the suggested diagnosis. However, participants A01, A04, and B03 all mention it in their decision-making. A statement from A01 helps to explain why stress is important; *“He’s really stressed, understandably, he’s recently unemployed and probably was made redundant. There’s a good association between stress and tinnitus and other symptoms that you don’t always have a cause for.”*

But what about recent illness and menopause? Menopause was lightly touched upon, with one participant discounting it and another wanting to know more.

*“So, she’s menopausal – that doesn’t mean much to me” B02 said. In contrast, A01 wondered aloud, “How is their menopause going?”*

Recent illnesses seemed agreeably more important. Immediately after discounting the menopause, B02 began to question recent changes in the patient’s health; *“I need to know her family history, her prescriptions, recent illnesses, loss of weight, coughs and sniffles, no diarrhea. Those are things I would need to ask to rule out viral.”*

While the effects of menopause may be considered, a recent illness can have a more drastic impact on determining the source of the symptoms:

*“If we said ‘Yeah, he’s 20 years old and he had the flu the other day’ then yeah, boom. But if he’s 70 years old and just had a stroke, I’d definitely go ‘Well, yeah, I doubt it’s Viral Labrynthitis.” B02*

## **Smoking and Drinking**

The patient’s vices (drinking and smoking habits) were mentioned a total of nine times, but only during two different case study tasks: Patient K, who drank 60 units each week, and Patient Q, who smoked two packs each week.

In the explanations with high-completeness, only Patient Q's vice was listed. Patient K's was not, as the drinking did not match the (incorrectly) suggested disease. Patient Q and K were not the only patients who had vices, but they were the only ones whose vices were referred to.

Participants referred to Patient Q's smoking habit as supporting evidence for the (incorrectly) suggested diagnosis.

*"It's just the smoking in this explanation that makes me agree with it. But again, that may be my limited experience playing a role again."*

A04

*"I'm a bit surprised, I mean he's 58, he's a guy, he's a bit older, so he has some of the risk factors of having an ischaemic event. But he only smokes 2 packs a week, which isn't that [much]. But I guess I don't know how long he's been smoking in total."* B01

While Patient Q's smoking habit seemed to support the incorrectly suggested diagnosis, Patient K's excessive drinking did not. Participants struggled to determine if the suggested diagnosis was correct, or if the true cause of Patient K's symptoms was alcohol consumption.

*"I don't know if these falls are related to when she's drinking. Because it completely changes – if it's related to the drinking then it's related to the drinking, but if the falls are separate to the drinking, if it's not related to the drinking, then that's something different."* B03

## **Gender and Age**

Finally, gender and age came up the most frequently. There were thirteen mentions to a patient's gender and eighteen to age.

The statements made to gender were rather straightforward, lacking in inference. Unlike with the other Medical History references, participants did not make statements like "Well, this patient is a female so..." or "He's a male, which means..." Instead, all of the statements were passive references to the patient that used their gender to refer to them. For example, A01's reference to the patient's gender: *"I've got a 22 year old female with vertigo, nausea..."* The transcripts suggest that gender had no implication on whether or not the diagnosis was correct.

Age, however, was treated differently. Age was mentioned eighteen times. The age of the patient seems to have clinical significance. In some instances, a suggested diagnosis seemed more plausible given the age, even if the suggestion was incorrect:

*"So he gets dizzy, but he's 82. ...The disorientation tells me that he's getting a bit old lately, things are just not working the way they should"*

*be. Small steps, balance is gone. It all just works up to an age thing, really.” B01*

In other instances, it seems that a patient's age was taken into consideration when the suggested diagnosis was rare or severe:

*“I mean, he's a young guy, there shouldn't be that much wrong with him.” B01*

*“She's 22? ...Okay, I can definitely see that. She's very young to have it but I can see it. For sure. ...I just think she's very young to have a vestibular schwannoma.” B01*

*“And in view of her age - if she was 34 that's something, it'd be different, if she was 64 and this is the first presentation of her vertigo and she's getting that amount of frequency despite her exams being completely normal, and also she's got problems with her hearing, I would also wonder about things like acoustic neuroma. That would be worth looking at.” B03*

*“She's how old? 42? ...She's young, so it's not too much of an issue. If she was 52 and this was the first onset, then I'd be thinking about something horrible going on. Because anyone that gets a migraine after the age of 50 for the first time, it's not usually a good thing.” B03*

#### 4.6.1.2 Symptoms

Both groups referred to the symptoms presented in the case studies more frequently than the examination results and medical histories. The High Completeness group referred to the symptoms in 92% of the case studies, while the Low Completeness group referred to them in 91%. These numbers alone indicate that the symptom information is highly desirable when determining the veracity of a suggested diagnosis. Statements made by several participants confirm this, showing that a patient's symptoms are the primary basis upon which a diagnosis is formed:

*“80% of the symptoms information you will use to diagnose rather than the exam.” B01*

*“So, before I even did the examination bit if I had a patient who came in and told me that rolling over in bed was what caused the symptoms, that's what my diagnosis would be even without doing the examinations.” A01*

Participants from the Low Completeness group made several explicit statements during their decision-making tasks that the patient's symptoms should be included in the explanation:

*“With BPPV it’s when you move your head and that’s what triggers it. If [the explanation] was to say ‘Triggering event rolling in bed’ that for me would kind of cement it more. That and the fact that her tinnitus is bi-lateral.” B01*

*“So if it said ‘in the history he’s said he had a fall and those sort of things’ so you get a little summary, so not just the exam results, but that’s just probably because of the way I work.” B03*

*“I look at the symptoms a lot, it’d be easier if it said it to me.” B03*

It was expected that the Low Completeness group would refer to symptoms despite the information not being present in the explanation, but it was not so expected for the High Completeness group to do the same. That is, 12% of the references made by the High Completeness group were to symptom information that was presented in the case study but *not* shown in the explanation for the suggested diagnosis. A quote by participant A02 best summarizes the reason for this behaviour:

*“They only put the negatives and not the positives. So that hearing loss was unknown, but they didn’t tell us that there was tinnitus or there was vertigo – they tell us what was not there, not what was there. I don’t know why it chose that. I don’t know why they chose that the hearing loss was not known instead of telling us that there was tinnitus.” A02*

#### 4.6.1.3 Examination Results

The High Completeness group and Low Completeness group both referred to the examinations results in nearly 60% of the decision-making tasks. Of the 52 case studies completed, there were 31 in which participants referred to the patient’s examination results; 17 of these references came from Low Completeness group with the remaining 14 coming from the High Completeness group.

In some of the decision-making tasks, participants mentioned that the examination results clearly pointed to a particular diagnosis. In these instances, the examination results were used as an easy confirmation of the diagnosis.

*“I think for different things it depends on the other exams that are ‘normal’. So for BPPV, when the Dix Hallpike is abnormal, it’s so strongly symbolic of it being BPPV that the rest [of the exams] don’t really matter. But with other things where there is more of a spectrum then you would want the other normal results as well.” B01*

In other instances, the examination results were used to determine the veracity of the suggested diagnosis. Participants referred to the results,

sometimes weighing the outcome of the different tests against each other, other times using the results to see if they support the diagnosis.

*“I guess with the Romberg we’ve ruled out anything cerebellar, and you can see with the Gait Test that his balance is diminished because he’s only taking these small steps. Smooth Pursuit says nothing with... whatever that reflex is called. So it’s not that.” B01*

*“This [exam results] could be... it could be Viral Labrynthitis, Meniere’s, a lot of other things could fit into just these examination results.” B02*

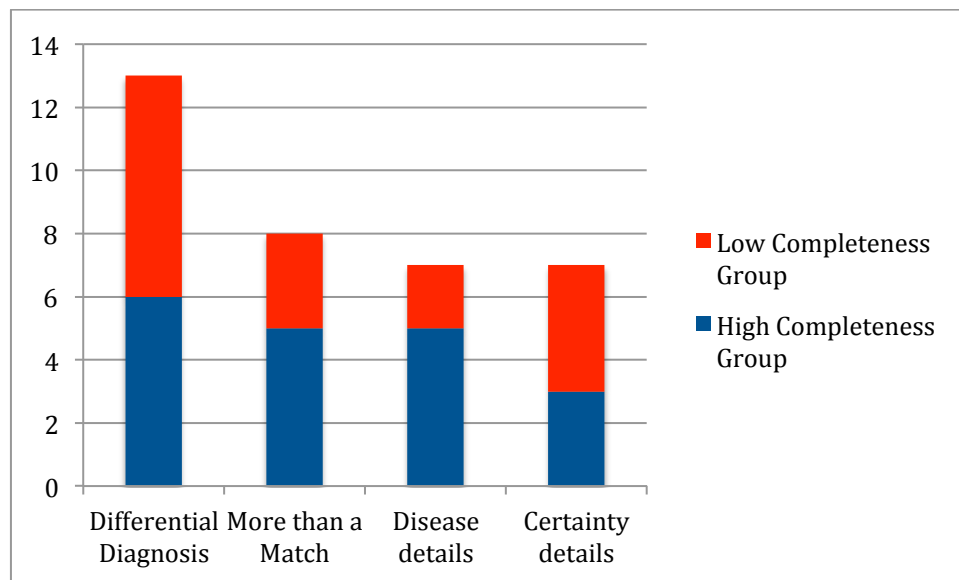
The third way in which the examination results were referenced was more quizzical. In these instances, the participants seemed unfamiliar with the examination or the significance of the result. We can see this with participant A02’s comment made while deliberating the diagnosis for Patient N, *“Yeah, so... I think the Romberg Test is when you stand on one leg.”*

Without knowing how an examination is performed, participants struggled to determine what the result meant clinically.

*“Yeah, maybe I’m not familiar with these tests. I know about Head Thrust. I know Romberg. It’s this one... Or maybe what it means is that I don’t know what it means if most of them have positive positional nystagmus. I think that’s what I’m trying to say. So I don’t know. I don’t know.” A03*

#### **4.6.2 Additional Explanations Desired**

Apart from what patient details are useful to include in an explanation, it’s important to look at what other information clinical users find helpful or necessary. Four information types emerged from the coding of the transcripts: Differential Diagnosis, More Than a Match, Disease Details, and Certainty Details. Figure 26 displays the frequency of each information type in a stacked bar chart. The details of each information type are presented in the subsections that follow.



**Figure 26: Frequency of Additional Information Types Requested/Mentioned**

#### 4.6.2.1 Certainty Details

Each suggested diagnosis was accompanied with a percentage of system certainty and there were seven instances in which participants wanted more information regarding the system's certainty.

*"I'm kind of like, where does this figure come from? Where does the software calculate the low degree of this figure?" A04*

*"My question is if this is the only diagnosis that it has come up with, then why is the certainty so low? ...It sounds like that's what it is, but I don't know why the system is not certain." A02*

Perhaps these inquiries were a result of the study design; the reader may recall that the system certainty percentages were manipulated during the design of the case studies. A04's quote, below, provides a good indication of how the manipulated certainty percentages may have caused confusion; he explains that he is confused by the certainty because he feels that it is low when it should be high, and vice versa.

*"So again, the ones I would agree with are the ones with the lowest degree of certainty and that kind of puzzles me because the only time I had a plausible diagnosis it was something completely unrelated, or something I hadn't even remotely thought about and it had a high degree of certainty. You see what I mean? That was the only time I had something in mind and something completely different comes up with a pretty high degree of certainty." A04*

While the Certainty Details responses received during the study may be partially an effect of manipulating the percentages, they also may be an



indication of users wanting more system transparency and, as suggested by B03's quote (below), a differential diagnosis.

*"So when you say the certainty is 19%, um, is that sort of like kind of - that's basically it and other things are rejected? What does that mean? Because it's got a diagnosis of Meniere's here and so is that sort of like saying that the certainty of the diagnosis is 19% chance of that being correct?" B03*

*"When it says 19%, I think 'is it 19% and the other one is 81%? Is it because there are too many positives for that disease? I'm not sure. I'm not sure what that percentage means. I think I would like to know what the percentage means. Why some of it is 81% and some of it is 17%? Does it mean it's not too sure but it is pointing to this?" A03*

These results suggest that providing a percentage of system certainty is not enough for the clinical users to comfortably decide whether or not a suggestion should be accepted.

#### 4.6.2.2 Disease Details

There were seven instances in which requests for Disease Details were made; five from the High Completeness group, two from the Low Completeness group. When a disease was suggested by the system, the participants of the study were only given a name. In these seven decision-making instances where additional disease detail was sought, the participants asked for a disease description, other common symptoms, and examinations to confirm the diagnosis.

*"I don't know whether it is possible to say 'Meniere's disease, usually a person...' [\*pulls out reference book\*] Okay, like... Having a little summary. Like, if its Meniere's, to say 'Features such as hearing loss and tinnitus suggest Meniere's disease or Acoustic Neuroma. Something like that. Just a quick summary." A03*

*"I'd want to see other things that would cause it. Because it's some sort of stroke. So any other risk factors – patient weight, history of heart disease, history of angina?" A02*

*"So, I don't know what side it's on and I can't remember, I think with Vestibular Schwannoma you tend to get it on one side... That would've helped me make my diagnosis. I'm just trying to remember my medicine as well." B01*

Given that the prototype used in this research was designed to aid non-specialised clinicians diagnose patients with balance disorders – a specialized domain of medicine – it seems likely that the users might not be familiar with all of the suggested diseases.

#### 4.6.2.3 More Than a Match

While the explanations provided reasons that supported why the suggested diagnosis was made, there were still instances in which participants wanted more detail. 'More Than a Match' occurred in eight of the decision-making tasks; three times from the Low Completeness group and five from the High Completeness group. Participants also brought this up in their open-ended answers to the Confidence and Trust questions.

*"All it says is that they've got migraines, which they've obviously come in and told me. It's triggered by diet, which it could be. They've got some tinnitus and vertigo. So, it just tells me what I know already. It doesn't explain why it's come to that conclusion. See what I mean?"*  
A01

*"Cause it just says 'matches to the disease profile'..."* B03

*"If it just explained the results of those examinations as to why it gives me that as a diagnosis..."* A01

So if 'matches' to the disease profile aren't enough, what more do they want? A01 gave a pretty good description when explaining the impact of the explanations on his confidence:

*"It didn't really explain the reasoning. I think also if it explained the significance of positive findings on an examination then I would be more willing to accept it. So, like, if it said 'Dix Hallpike shows rotational nystagmus which is suggestive of autillitis in the ear canal causing patient's vertigo symptoms in combination with positional element' then I would understand that. But it just says 'matches disease profile because it matches this trigger and this exam result.'" A01*

#### 4.6.2.4 Differential Diagnosis:

In one quarter of the decision-making tasks (13 of 52), participants desired a differential diagnosis. A differential was required by the High Completeness Group six times, or in 21% of the case studies, while the Low Completeness group mentioned it seven times (29%).

A differential diagnosis is, simply put, an alternative diagnosis that a clinician tries to 'rule out' or disprove. A differential diagnosis is a key part of the clinical diagnostic process. The absence of such was noted in instances when the participant was finding it difficult to determine the veracity of the suggested diagnosis.

*"As I said before, when I'm not sure of the diagnosis I would want it to suggest other things it could be."* A02

*“So um... So you know when you are making a diagnosis there is something called a differential diagnosis, something else that it could be. Does it have the ability to tell me what else it could be? To suggest other things it could be?” A02*

*“The other thing is what it tells you, as I said this is to do with the differential diagnosis, so that’s one possibility out of two or three that I’m thinking about. And it only gives you one... It might be useful to see the others... So it’s sometimes quite useful to look at differential diagnoses because it kind of prompts you to think ‘Actually, yeah, that might be a possibility.’” B03*

*“It would be helpful to sort of say ‘the second differential is xyz and these are the things which it fulfills. But the Hallpike result is normal and therefore it seems against it.’” B02*

These results suggest that a differential diagnosis is a necessary feature for a CDSS to include.

## Chapter 5 Discussion

The results of this study have shown that completeness of a CDSS's explanation can have an impact on decision-making, trust, and confidence. The results have also shown that explanations with high completeness appear to be more desirable to clinicians, although there are additional explanation types that are also desired. In the following sections, each of these findings will be discussed in the context of the related research questions, literature, and the generalizability of the results. This chapter concludes with consideration of the future work that could build upon the findings of this research.

### Section 5.1 The Impact of Completeness on Decision-Making

Contrary to previous research by Seong and Bisantz (2008), the results of this research study did not suggest that there is any relationship between clinical users making 'right' or 'wrong' decisions and the completeness of the explanation provided; neither group 'performed' better than the other.

The analysis of the decision-making does reveal one phenomenon which can be explained in two ways: based on the findings of this study, it can be argued that explanation completeness and the decision to accept or decline are related.

The first explanation for this is that clinical users with explanations of low completeness will accept a CDSS's suggestion more frequently than if given explanations of high completeness. The explanations provided to the High Completeness group included more justification (more matches to the suggested diagnosis) than those provided to the Low Completeness group. Gonul et al found that longer explanations were no more persuasive than short ones (Gonul et al., 2006), but they were only looking at length, not the content. In the study reported here, an explanation with high-completeness was not just longer than one with low-completeness, it included more justification for the suggestion. Additionally, the participants in the High Completeness group did seem to be sensitive to the reliability of the system, but did not suggest that it had any effect on them. Thus, the participants of this group not only accepted 82% of the system's suggestions, they did so whilst acknowledging that the system was unreliable; essentially epitomizing the definition of automation bias. With these results, one can surmise that *Why...* explanations with high completeness do lead to automation bias, because the explanations provide a high level of justification.

The second explanation for this phenomena is that the explanations of low-completeness seem, to the participants, to be missing information. In

‘Thinking and Deciding’, Baron wrote about decision-making and missing information: when information is perceived as missing a decision-maker is reluctant to make a choice (Baron, 2000). Perhaps this explains the decision-making behaviours observed of the Low Completeness group; the explanations of low-completeness are perceived to have information absent, or ‘missing’, causing the users to be reluctant to accept the suggested diagnosis.

The results discussed in this section could be generalized to other decision support systems, not just those used within a clinical setting. However, with the low number of participants a more confident statement cannot be made regarding completeness and automation bias or reluctance. A larger-scale study conducted across multiple domains is needed to determine if these results are truly generalizable across a wider context.

## **Section 5.2 The Impact of Completeness on Confidence and Trust**

The results presented in 0 and Section 4.4 (pertaining to RQ2 and RQ3, respectively) suggest that explanations of high completeness have a positive impact on a user’s trust in a CDSS and confidence in achieving a diagnosis when aided by it. These results will be discussed in this section together as the implications are quite similar: explanations of high completeness increase the intelligibility of the system and assure the user that an appropriate reasoning process is being used while explanations with low completeness result in incorrect mental models and doubts about the quality of the system’s reasoning.

### **5.2.1 The Impact of Explanations of Low Completeness on Confidence and Trust**

One of the purposes of explanations is to indicate to the user that an appropriate reasoning process has been taken, one that is similar to the user’s own (Hasling et al., 1984). The results of this study suggest that the participants of the Low Completeness group perceived that the CDSS’s reasoning was not only *inappropriate* but also *inferior* to their own. System explanations inform the user’s mental model of how the system works (Kulesza et al., 2012). The participants of the Low Completeness group felt that the explanations did not reflect their own reasoning process. That is, with only the examination results given in the explanations, these participants assumed the CDSS *did not consider* the symptoms or medical history of the patient. Thus, it is reasonable to state that as a result of explanations of low-completeness, the clinical users formed an incorrect mental model of the system’s reasoning process. Numerous research studies on automated systems and context-aware systems have suggested that a user’s trust and/or

confidence is impacted by their mental model (Kulesza et al., 2013), perception of the system's abilities (Detweiler & Broekens, 2009) (Seong & Bisantz, 2008), and perception that the system follows their own reasoning (Alexander, 2006). The results from this study suggest the same to hold true for users of decision support systems.

### **5.2.2 The Impact of Explanations of High Completeness on Confidence and Trust**

Participants in the High Completeness group felt that the explanations showed the system's reasoning; that the patient information listed in the explanation showed that the CDSS was using the same information that the users would to achieve a diagnosis. This is important, as it indicates that *Why...* explanations of high-completeness are not only successful in providing intelligibility but also in assuring the user that the system is using an appropriate reasoning process. Intelligibility and appropriate reasoning, as mentioned before, are two of the major obstacles to adoption and utilization of clinical decision support systems (Alberdi et al., 2001) (Alexander, 2006) (O'Sullivan et al., 2014). This finding supports the recent work of Kulesza et al (2013), who suggest that high completeness increases a user's trust. This finding also supports earlier work done by Lee and Moray (1994), who determined that a user's confidence in their decision-making often increases when given more information during a judgment task. While Kulesza et al and Lee and Moray were focusing on automated systems, neither was pertaining to decision support systems or the medical domain. The results from this study suggest that their findings relate to clinical decision support systems as well; a more complete CDSS explanation leads to a more confident clinical decision-maker.

If a user is more confident in their own abilities, or trusts their own reasoning more than that of the system, the system will not be used (de Vries et al., 2003). Thus, it can be recommended that clinical decision support systems employ explanations of high completeness. However, in order to make a more conclusive determination of the impact of completeness on a clinician's decision-making a larger-scale research study should be performed.

## **Section 5.3 The Impact of Completeness on a Clinical User's Workload**

Why would more complete explanations demand more of a user's time? Of course, more information would require more time to read than less information. But is this temporal demand a concern? The first of the Ten Commandments set forth by Bates et al (2003) for clinical decision support systems suggests it is crucial, stating that speed is the primary goal and the most valued parameter. However, the results of this study do not suggest that the temporal demands outweigh the benefits of a more complete explanation, aligning with a recent finding by Kulesza et al (2013). The difference between the calculated workload of each group is less than 0.10 points (very minimal) and there were no comments suggesting that the explanations required too much time. Indeed, the analysis of the decision-making tasks that led to the highest workload ratings suggest that a more complete explanation or additional explanation types (discussed in 5.4.2) may actually bring additional benefits to the user's workload.

However, if any additional content is added then the impact on the user's workload would need to be re-evaluated to be certain that the benefits are not overshadowed by the cost. Additionally, as Parasuraman and Riley wrote, 'the nature of workload in real work settings can be fundamentally different from workload in most laboratory settings' (Parasuraman & Riley, 1997). In order to determine the true effect of explanation completeness, the users must be studied in a real work setting.

## **Section 5.4 Information Used and Desired**

### **5.4.1 Patient Information Wanted in an Explanation**

The results presented in Section 4.6.1 suggest that explanations with high completeness are more desirable to clinicians. The participants in the Low Completeness group confirm this; they mentioned the Medical History and Symptoms during their decision-making tasks and even stated that the explanations should include this information. However, there does appear to be some patient information that is more desirable, or perhaps more informative, to clinical users than other information.

The transcripts suggest that the Medical History, Symptoms, and Examination Results each play a different role in the diagnostic process. According to several statements, a patient's symptoms are the basis upon which diagnostic theories are developed. The examination results are used to confirm or reject the hypothesized diagnoses. The clinical value of the information included in



a patient's medical history, however, is variable. A heavier emphasis is placed on a patient's emotional state, recent illnesses, bad habits, gender and age than on the toll that the disease is taking on their life. Perhaps this can be explained by the subjective nature of the information, as suggested by one participant. However, these different bits are not used consistently; in some instances they hold importance, whilst in others they do not. Thus, a distinction cannot be made at this point regarding which pieces of Medical History should be included in an explanation and which should not. This is not to suggest that the design of a CDSS should disregard less objective measures of a person's health, but perhaps these findings are indicative that additional thought must go into the explanations provided. Symptoms and examination results are highly appropriate, desired, and frequently used. One can say with a high degree of confidence that the symptoms and examination results should be included in the explanation. However, determining which pieces of the medical history are worth including requires further exploration and is likely dependent upon the diagnosis at hand.

#### **5.4.2 Additional Explanations Desired**

The prototype used in this study included *Certainty* and *Why...* explanations. While Section 4.6.1 suggests that the explanations provided to the High Completeness group are preferable, the results detailed in Section 4.6.2 indicate that these are not the only explanations that are desired. Four additional explanation types emerged from the coding: Certainty Details, Disease Details, More Than a Match, and Differential Diagnosis. To the best of this author's knowledge, these findings are novel in that the four explanation types do not match others that have been previously defined. The following four sub-sections discuss each of these new CDSS explanation types.

##### **5.4.2.1 Certainty Details**

*Certainty* explanations "inform users how (un)certain the application is of the output value produced" (Lim & Dey, 2010) and provide users with a rational basis upon which they may begin determining whether the system's suggestion should be accepted. Providing certainty in the form of a percentage is not uncommon practice and is included as one of the eight explanation types put forth by Lim and Dey in their oft-referenced intelligibility toolkit (Lim & Dey, 2010). Lim and Dey suggest that *Certainty* explanations "report the probability of inference" (Lim & Dey, 2010), which is how the system certainty was presented in this study. However, the transcripts related to Certainty Details suggest that simply providing a percentage of system certainty is not enough.

As mentioned in the results, this finding may be a result of manipulating the diagnoses and certainty percentages. In many of the instances, participants



questioned the system certainty because it surprised them as being quite high when they expected it to be low or vice-versa. Thus, the request for Certainty Details could be explained by the surprise of the participants: as Gregor and Benbasat (1999) wrote, explanations become important 'when the user perceives an anomaly.' The *Why...* explanations provided justification for the suggestion but there was no justification, or explanation, for the certainty.

However, it may also be that these requests are caused by the lack of a differential diagnosis being present. Indeed, when the system certainty was low the participants wanted to know what else the system thought it might be. Perhaps by providing a Differential Diagnosis explanation, the Certainty Details will no longer be a concern. This topic requires additional exploration before determining whether Certainty Details are a necessary feature of a CDSS and, if so, what the explanation should include.

#### 5.4.2.2 Disease Details

In the seven decision-making instances where additional disease detail was sought the participants asked for a disease description, examples of other common symptoms, and a list of suggested examinations to confirm the diagnosis.

*"I don't know whether it is possible to say 'Meniere's disease, usually a person...' [\*pulls out reference book\*] Okay, like... Having a little summary. Like, if its Meniere's, to say 'Features such as hearing loss and tinnitus suggest Meniere's disease or Acoustic Neuroma. Something like that. Just a quick summary.'" A03*

Given that the participants of this study were not specially trained in diagnosing balance diseases, it seems likely that they would be unfamiliar with some of the diseases suggested by the CDSS. The additional information that was requested would enable these participants to determine if the patient's pathophysiology matches the suggested disease. This method of information seeking and usage is common amongst novice clinicians (Larkin et al., 1980). Given the results of this study, it is conceivable that this use of information also applies to clinicians who are novices in a particular medical domain. Expert clinicians can make quick diagnostic hypotheses based on prior personal experience but a novice must revert to reasoning from their biomedical knowledge, resulting in 'a slower and often less accurate processing of the information' (Alberdi et al., 2001). Thus, a CDSS designed to aid novice or non-specialist users could perhaps reduce the Mental Demand and improve the decisions made by including Disease Details.

Using the comments made by the participants as a guideline, a Disease Details explanation would include a brief explanation of the affected physiology, a list common symptoms, suggested examinations and expected results that would confirm the diagnosis, and suggested treatments.

However, given that this information was not frequently requested during the study, and given that users will inevitably learn about the diseases over time, it can be recommended that Disease Details are provided on demand. Of course, the exact information that should be included must be confirmed.

#### 5.4.2.3 More Than a Match

A *Why...* explanation is defined by Lim and Dey as one that informs users “why the application derived its output value from the current (or previous) input values. For rule-based systems, this returns the conditions (rules) that were true such that the output was selected” (Lim & Dey, 2010). In the eight More Than a Match instances, participants requested the CDSS to provide information beyond the scope of a *Why...* explanation. The listed matches to the suggested disease were felt to be insufficient, as if the system was merely regurgitating the information that the participant had input, rather than explaining the significance of the information. Thus, the More Than a Match requests are for the system to provide information that it never received. Indeed, this is a request for the system to make inferences about the symptoms and test results that were entered by the user. This is, in a way, similar to the NEOMYCIN work done by Clancy, who sought to create educational explanations for clinicians seeking advice (Clancy, 1985).

Perhaps this too is an effect of the limited expertise that these general-knowledge clinicians have on the specialized domain of balance diseases. Again, psychological research on medical reasoning shows that less experienced clinicians rely on their “knowledge of underlying pathophysiology and anatomy,” or “biomedical knowledge,” when trying to reach a diagnosis (Alberdi et al., 2001). Indeed, many of the decision-making tasks included instances where the participants drew biomedical inferences from the provided patient data. For example:

*“I guess with the Romberg we've ruled out anything cerebellar, and you can see with the Gait Test that his balance is diminished because he's only taking these small steps. Smooth Pursuit says nothing with... whatever that reflex is called.” B01*

Based upon the comments made by participants, a More Than a Match explanation should first indicate the patient information that matches the disease profile, and then explain how that information is clinically significant and indicative of the suggested diagnosis. In fact, a More Than a Match explanation could have quite a positive effect on clinical users and their practice, as it would make ‘associations between pieces of information that clinicians might miss’ (Bates et al., 2003). Of course, an individual with no medical training cannot construct such an explanation and, as suggested by O'Sullivan et al (2014), collaboration between HCI and medical professionals would be necessary.

#### 5.4.2.4 Differential Diagnosis

A differential diagnosis was requested in one quarter of the decision-making tasks. A differential diagnosis is used as an alternative hypothesis that a clinician can rule out using the patient's symptoms and examination results. Thus, it makes sense that a differential was requested in one quarter of the decision-making tasks.

*“So, once again if we have that typical picture we need all the signs pointing towards or against. These exams all say normal, so they'd point towards other things, but... Going back for example to if we had that text book case and these are the positives - that gives me three out of how over many criteria - it would also be helpful to sort of say 'The second differential is xyz and these are the things which it fulfills. But the Hallpike result is normal therefore it seems against it'.” B02*

These results suggest that a differential diagnosis is a desirable feature to include in any CDSS designed to aid a user in achieving a diagnosis. Including a Differential Diagnosis explanation may also improve trust and confidence, as it would provide further indication that the system's reasoning is similar to the user's.

However, the way in which this information should be presented is not clear at this point in time. Would clinicians want the differential to be presented in the same way as the suggested diagnosis? Should the information be presented side-by-side, for better comparison? Would using Lim and Dey's (2010) *Why Not...* explanations work as effectively as Differential Diagnosis might? As with the Certainty Details, More Than a Match, and Disease Details, the means and effect of including a Differential Diagnosis requires additional research.

### Section 5.5 Future Work

Whilst this author believes the results to be credible, generalizable, and novel, it is recommended that a larger scale study be performed to validate the findings. However, there are also implications that can be drawn from the findings that can inform or inspire future work on clinical decision support systems.

The findings from this research suggest that *Why...* type explanations with high completeness are preferable to clinicians using a clinical decision support system. The high completeness improves trust, confidence, and has little impact on workload. Unfortunately, they may not be successful in preventing automation bias. This statement naturally leads one to wonder what the appropriate level of completeness might be, and perhaps even consider planning a study similar to the one described in Kulesza et al's aptly named 'Too Little, Too Much, or Just Right?' paper (2013). In consideration

that the participants frequently mentioned the information included in the highly complete explanations and that additional explanations were requested, finding the pivot point of completeness may not be the best approach in this domain. Rather, this author believes that the key may lie in the new explanation types that emerged from the study results.

Four explanation types were identified from the results, ones that are specific to clinical decision support systems. To remind the reader, these are: Certainty Details, Disease Details, More Than a Match, and Differential Diagnosis. Lim and Dey (2011) recently reported that some explanation types are more effective in providing intelligibility. Reflecting upon this, perhaps one of these four explanation types would be more effective in preventing automation bias and providing intelligibility than the *Why...* explanations used in this research. Additional research is required to define these explanation types, understand the instances in which they would hold the most value, and measure the effects they will have on the user's trust, workload, confidence, and decision-making.

Furthermore, this author proposes that the HCI community should seek to understand the reasoning strategies used by clinicians when faced with different problems. Reasoning strategies amongst clinically trained individuals are known to vary based upon domain and expertise (Alberdi et al., 2001) (Arocha et al., 2005) (Gregor & Benbasat, 1999). Structuring explanations specifically to reflect the different reasoning strategies may be the solution to the problems researched here (Arocha et al., 2005) (O'Sullivan et al., 2014).

## Chapter 6 **Evaluation, Reflections, and Conclusion**

This research project explored the impact of explanation completeness on clinical users of CDSSs, an area of knowledge that this author found as lacking within the HCI community. The following sections summarise the work that has been done and include the author's reflections on what was learnt and what could be improved.

### **Section 6.1 Review of Objectives**

Section 1.2 introduces the five objectives set for this research project. Each of the five objectives have been attained over the course of the project.

Objectives 1, 2, and 3 were reached through a review of the related literature. The result of Objective 1, which called for a clarification of the role of decision support systems within the medical domain, is provided in Section 2.3, in which clinical decision support systems are defined along with examples of their various purposes. Following this is Section 2.4, which includes the results of Objective 2: the issue of CDSS reliability and a user's bias towards automation. Objective 3 was also achieved in this chapter, as Section 2.5 outlines the means in which existing literature suggests system explanations might provide intelligibility and mitigate the potential for automation bias.

Objective 4 required the design of an experimental study that would answer the five research questions. The study that was designed is described in detail in Chapter 3.

Finally, Objective 5, which required evaluating the results of the study in order to answer the five research questions, is concluded in Chapter 4 and Chapter 5, where the results related to each research question are presented and discussed.

### **Section 6.2 Literature examined**

Chapter 2 provides the critical context of the research focus, explaining the background and current state of clinical decision support systems as well as the issues that exist in present day. The potential solution, system explanations, is introduced and, as are the positive and negative effects that explanation completeness could cause. In order to provide the reader with a full picture, a wide array of literature was analysed. Literature on intelligent, automated, and decision support systems was reviewed, as were the effects of these on decision-making, trust, confidence, and workload. Intelligibility and explanations were also explored. Of course, within the timeline for this

project it was impossible to review all of the existing literature related to the research, but the key publications were covered. As such, this chapter shapes the importance and necessity for the research reported in this document.

## **Section 6.3 Planning and Method**

Overall, this project ran smoothly and with no major deviations to the plan. The schedule was useful and largely adhered to, but the timeline of this research was dependent on the activities completed on the EMBalance project. For example, the Wizard-of-Oz prototype used in the study could not be completed until the prototype designed for the EMBalance project was concluded, meaning that recruitment and testing also had to wait. However, the project plan was designed with the anticipation of such obstacles and, thus, there was no major disruption to the progression of the research.

There was, however, one lesson learnt by this author in regards to planning, and that is in regards to the scheduling of study sessions. The lesson is this: a researcher of clinically-trained individuals must be *very* flexible and prepared. The need for flexibility comes as a result of the hours with which this population works; several of the study sessions needed to be conducted very early in the morning or very late in the evening. An inability to meet during these times would result in fewer participants. The need to be prepared is also a result of this; there were several instances in which a session was scheduled to occur less than 24 hours after recruitment. Thus, study materials needed to be available every day, in the event that a session could be conducted.

Whilst this author stands behind the chosen methods to approach this research project, there are a few aspects of the study that could be improved upon. Changes that were made to the goals and method since the original proposal are discussed below, followed by reflections on the case studies, recruitment, and balanced groups.

### **6.3.1 Changes to Goals and Methods**

The original proposal written for this research project shows a greater focus on trust and appropriate use than the project reported here. Indeed, the scope of the project shifted over the course of the first few months, as a review of related literature prompted additional questions to be asked. Given that the topic of trust was still explored whilst also exploring various other facets of decision support system use, this author feels that expanding the scope of the research has resulted in a greater contribution to academic knowledge. As mentioned, the proposal is included in Appendix A.

### 6.3.2 Case study validation

The case studies, as stated in Section 3.4.2.1, were created from medical literature and diagnostic flow diagrams. A high level of care, attention, and effort was taken to ensure that these case studies were accurate. However, one cannot be completely certain of their accuracy without validation. Had a neuro-otologist or other expert of balance disorders reviewed and validated the case studies prior to the start of the research sessions, the accuracy would have been certain. Having not done so, any future work that builds off of this research should consider that these case studies have not been validated.

### 6.3.3 Troubles with Recruitment

It has been acknowledged since the beginning of this research project that the target population would be very difficult to recruit. However, the difficulty was, in fact, still underestimated.

The recruitment process began with what was considered, at the time, to be an extensive outreach. After a considerable amount of time passed with barely any responses, it became clear that what was once considered extensive was actually quite modest, causing this researcher to reach out through additional mediums. Even with all of the efforts described in 3.3.2, a surprisingly low number of interested parties emerged. Additionally, nearly half of the received responses led nowhere, despite numerous attempts on behalf of the researcher. Even attempts to access the network of the individuals who participated led to nothing.

Regarding the recruitment of clinically-trained individuals, there are three major lessons. First, one should begin the recruitment process boldly, sending the advertisement through what seems to be an excessive amount of contact channels. The most successful resources seemed to be on-line forums, general practitioner mailing lists, and strong words of encouragement from lecturers. Second, the process cannot begin *early*. That is, one cannot begin the recruitment process before they are able to conduct the research. As described above, these target individuals tend to schedule things on a last-minute basis, making it nearly impossible to plan more than two days in advance. Third, the incentive for participation should be deeply considered. In the beginning, it was hoped that no monetary incentive would be necessary and that the promise of food and drink would be sufficient. This appeared to have been wrong, as no responses were received until a modest monetary incentive replaced food and drink on the recruitment advertisements. What is puzzling is that several participants actually stated, at the completion of the study session, that the money was not an incentive and that they would be quite happy with food and drink. Perhaps a more creative approach should be taken to incentivize these individuals to participate in future research.



#### **6.3.4 Balanced groups**

The use of balanced groups has been mentioned numerous times throughout this report, but a few additional statements must be made on the topic.

First, it is rather obvious that the attempt to balance the groups evenly was unsuccessful. With four participants in one group and three in the other, there is no way to even begin to imply that this was balanced. However, the troubles of recruitment have already been discussed.

What does need to be mentioned is the measure upon which balance would be created: a four-point self-rating of expertise. This seemed appropriate at the time, but in hindsight was too subjective. All of the participants rated themselves either a two or a three. However, some of the more experienced participants gave themselves a rating of two, while some of the very inexperienced participants gave themselves a three. Perhaps balancing the groups based on years of experience would have worked better, or even balancing based upon confidence in diagnosing patients with balance diseases.

#### **6.3.5 Data Analysis**

A combination of quantitative and qualitative data was gathered and analysed, helping the results of this research be more robust. However, there are a few aspects of the analytical process which could be improved upon. First, the amount of quantitative data is, admittedly, small and perhaps not entirely suited for statistical analysis. It is also worth admitting that statistics are rather foreign to this author and having successfully conducted various tests for this research was a considerable achievement. Additionally, regarding the qualitative data, it would have been preferable for a second individual to have independently coded the data and then test for agreement to ensure validity. This was not possible, given the time and scope of the project.

### **Section 6.4 Conclusions**

This report has illustrated the issues that surround clinical decision support systems and proposed a potential solution. A Wizard-of-Oz prototype was designed, as well as eight case studies, both of which were used to conduct an experimental research study focused on tackling these issues through the use of explanations. The findings of this research suggest that explanations with high completeness are effective at providing intelligibility and positively influencing a user's trust and confidence. However, the findings also suggest that explanations of high completeness may cause automation bias. Additionally, four new explanation types were identified. Based upon these findings, avenues for future work have been proposed.



Overall, this research project was conducted and completed successfully. All research questions were answered and all objectives were met. The results of this research are both valid and novel. Whilst there are points along the process that could be improved upon, any further work that builds off of this research can benefit from the lessons described in this chapter.

## Chapter 7 References

- Alberdi, E., Ayton, P., Povyakalo, A. & Strigini, L., 2005. Automation Bias and System Design: A case study in medical application. In *People and Systems - Who Are We Designing For. The IEE and MOD HFI DTC Symposium*. London, 2005. IET.
- Alberdi, E. et al., 2001. Expertise and the interpretation of computerized physiological data: implications for the design of computerized monitoring in neonatal intensive care. *International Journal of Human-Computer Studies*, 55(3), pp.191-216.
- Alberdi, E., Povyakalo, A., Strigini, L. & Ayton, P., 2004. Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. *Academic Radiology*, 11(8), pp.909-18.
- Alberdi, E., Stringini, L., Povyakalo, A.A. & Ayton, P., 2009. Why are people's decisions sometimes worse with computer support? In *Computer Safety, Reliability, and Security*, 2009. Springer-Verlag.
- Alexander, G.L., 2006. Issues of Trust and Ethics in Computerized Clinical Decision Support Systems. *Nursing Administration Quarterly*, 30(1), pp.21-29.
- Arocha, J.F., Wang, D. & Patel, V.L., 2005. Identifying reasoning strategies in medical decision making: A methodological guide. *Journal of Biomedical Informatics*, 38, pp.154-71.
- Bamiou, D.-E. et al., 2014. *Comprehensive Analysis On Balance Disorders Guidelines*. Clinical Guidelines. Unpublished European Commissioned Report.
- Baron, J., 2000. *Thinking and Deciding*. 3rd ed. Cambridge: Cambridge University Press.
- Bates, D.W. et al., 2003. Ten Commandments for effective clinical decision support: Making the practice of evidence-based medicine a reality. *Journal of the American Medical Informatics Association*, 10(6), pp.523-30.
- Braun, V. & Clarke, V., 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology*, 3(2), pp.77-101.
- Choi, B.C.K. et al., 1998. cing variability in treatment decision-making: effectiveness of educating clinicians about uncertainty. *Medical Education*, 32(1), pp.105-11.
- Clancy, W., 1985. *Acquiring, Represeneting, and Evaluating a Competence Model of Diagnositc Strategy*. Technical. Stanford: Stanford University Stanford University.
- de Vries, P., Midden, C. & Bouwhuis, D., 2003. The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies*, 58, pp.719-35.

- Detweiler, C. & Broekens, J., 2009. Trust in Online Technology: Towards Practical Guidelines Based on Experimentally Verified Theory. In *Human Computer Interaction. Ambient, Ubiquitous and Intelligent Interaction*. 1st ed. Berlin: Springer-Verlag. pp.605-14.
- Dzindolet, M.T. et al., 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(1), pp.697-718.
- Glass, A., McGuinness, D.L. & Wolverton, M., 2008. Toward Establishing Trust in Adaptive Agents. In *IUI'08*. Maspalomas, Gran Canaria, 2008. ACM.
- Gonul, M.S., Onkal, D. & Lawrence, M., 2006. The effects of structural characteristics of explanations on use of a DSS. *Decision Support Systems*, 42(1), pp.1481-93.
- Grabowski, M. & Sanborn, S.D., 2003. Human performance and embedded intelligent technology in safety-critical systems. *International Journal of Human-Computer Studies*, 58(1), pp.637-70.
- Gregor, S. & Benbasat, I., 1999. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly*, 23(4), pp.497-530.
- Hasling, D.W., Clancey, W.J. & Rennels, G., 1984. Strategic explanations for a diagnostic consultation system. *International Journal of Man-Machine Studies*, 20(1), pp.3-19.
- Kanstrup, A.M., Christiansen, M.B. & Nohr, C., 2010. Four principles for user interface design of computerised clinical decision support systems. *Patient Safety Informatics*, 166, pp.65-73.
- Kawamoto, K., Houlihan, C.A., Balas, E.A. & Lobach, D.F., 2005. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ*, 330, pp.1-8.
- Kong, G., Xu, D.-L. & Yang, J.-B., 2008. Clinical Decision Support Systems: A review on knowledge representation and inference under uncertainties. *International Journal of Computational Intelligence Systems*, 1(2), pp.159-67.
- Kulesza, T., Stumpf, S., Burnett, M. & Kwan, I., 2012. Tell Me More? Effects of Mental Model Soundness on Personalizing an Intelligent Agent. In *CHI 2012*. Austin, 2012. ACM.
- Kulesza, T. et al., 2013. Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models. In *IEEE Symposium on Visual Languages and Human-Centric Computing.*, 2013. IEEE.
- Larkin, J.H., McDermott, J., Simon, D.P. & Simon, H.A., 1980. Models of Competence in Solving Physics Problems. *Cognitive Science*, 4(1), pp.317-45.
- Lee, J.D. & See, K.A., 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1), pp.50-80.
- Lim, B.Y. & Dey, A.K., 2010. Toolkit to Support Intelligibility in Context-Aware Applications. In *UbiComp 2010*. Copenhagen, 2010. ACM.

- Lim, B.Y. & Dey, A.K., 2011. Investigating Intelligibility for Uncertain Context-Aware Applications. In *UbiComp'11*. Beijing, 2011. ACM.
- Lim, B.Y., Dey, A.K. & Avrahami, D., 2009. Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems. In *CHI 2009 - Studying Intelligent Systems*. Boston, 2009. AM.
- Madhavan, P. & Wiegmann, D.A., 2007. Similarities and differences between human-human and human-automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, 8(4), pp.277-301.
- NASA Human Performance Research Group, 1987. *Task Load Index (TLX) v. 1.0 Paper and Pencil Package*. Moffett Field: NASA Ames Research Center NASA.
- NIH National Institute on Deafness and Other Communication Disorders (NIDCD), 2010. *Vestibular Schwannoma (Acoustic Neuroma) and Neurofibromatosis*. [Online] Available at: [http://www.nidcd.nih.gov/health/hearing/pages/acoustic\\_neuroma.aspx](http://www.nidcd.nih.gov/health/hearing/pages/acoustic_neuroma.aspx) [Accessed 27 September 2014]. NIH Pub. No. 99-580.
- Osherooff, J. et al., 2005. *Improving outcomes with clinical decision support: an implementer's guide*. 2nd ed. Chicago: Healthcare Information and Management Systems Society (HIMSS).
- O'Sullivan, D., Fraccaro, P., Carson, E. & Weller, P., 2014. Decision time for clinical decision support systems. *Clinical Medicine*, 14(4), pp.338-41.
- Parasuraman, R. & Riley, V., 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39(2), pp.230-53.
- Povyakalo, A.A., Alberdi, E., Strigini, L. & Ayton, P., 2013. How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography.. *Medical Decision Making*, 33(1), pp.98-107.
- Price, M., 2012. *Preventing Misuse and Disuse of Automated Systems: Effects of System Confidence Display on Trust and Decision Performance*. Clemson: All Dissertations.
- Pu, P. & Chen, L., 2007. Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems*, 20, pp.542-56.
- Purchase, H.C., 2012. *Experimental Human-Computer Interaction: A practical guide with visual examples*. 1st ed. New York, NY: Cambridge University Press.
- Sambasivan, M., Esmaeilzadeh, P., Kumar, N. & Nezakati, H., 2012. Intention to adopt clinical decision support systems in a developing country: effect of Physician's perceived professional autonomy, involvement, and belief: a cross-sectional study. *BMC Medical Informatics & Decision Making*, 12(142), pp.1-8.
- Seong, Y. & Bisanz, A.M., 2008. The impact of cognitive feedback on judgment performance and trust with decision aids. *International Journal of Industrial Ergonomics*, 38, pp.608-25.

- Shortliffe, E.H., 1976. *Computer-Based Medical Consultations: MYCIN*. 1st ed. Amsterdam: Elsevier Scientific Publishing Company.
- Shupert, C.L. & Kulick, B., 2013. *Labrynthitis and Vestibular Neuritis*. [Online] Vestibular Disorders Association (VEDA) Available at: <http://vestibular.org/labyrinthitis-and-vestibular-neuritis> [Accessed 28 September 2014].
- Tintarev, N. & Masthoff, J., 2007. Effective explanations of recommendations: User-Centered Design. In *RecSys'07*. Minneapolis, 2007. ACM.
- Vestibular Disorders Association, 2014. *Vestibular Migraine*. [Online] Vestibular Disorders Association (VEDA) Available at: <http://vestibular.org/migraine-associated-vertigo-mav> [Accessed 27 September 2014].
- 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), pp.352-67.