



## City Research Online

### City, University of London Institutional Repository

---

**Citation:** Goddard, Kate (2012). Automation bias and prescribing decision support – rates, mediators and mitigators. (Unpublished Doctoral thesis, City University London)

This is the unspecified version of the paper.

This version of the publication may differ from the final published version.

---

**Permanent repository link:** <https://openaccess.city.ac.uk/id/eprint/3005/>

**Link to published version:**

**Copyright:** City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

**Reuse:** Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

CITY UNIVERSITY

# Automation Bias and Prescribing Decision Support – Rates, Mediators and Mitigators

---

A thesis submitted in fulfillment of the  
requirements for a PhD in Health Informatics

**Kate Goddard**

**April 2012**

## Contents

List of Figures .....	6
List of Tables .....	7
Acknowledgement .....	8
Declaration .....	9
Glossary .....	10
Abstract .....	12
1. Introduction .....	15
1.1 Background and aims .....	16
1.2 Automation Bias .....	16
1.2.1 Increasing accuracy of judgement .....	19
1.3 Study aims and objectives .....	21
1.4 Purpose and structure of thesis .....	22
1.4.1 Justification for study and methods .....	23
2. Literature Review .....	25
2.1 Introduction .....	25
2.2 Decision Support Systems .....	25
2.2.1 Types of Decision Support Systems .....	26
2.3 Domain for simulation - prescribing .....	27
2.3.1 Examples of prescribing Decision Support Systems .....	30
2.3.2 International differences in prescribing .....	32
2.4 General Practitioner response rates .....	40
2.5 Human judgement, decision making and advice taking .....	41
2.5.1 Heuristic use .....	43
2.5.2 Methods of measuring reliance .....	45
2.6 Factors which affect automation reliance and influence Automation Bias .....	45
2.6.1 Context .....	46
2.6.2 Task / event .....	46
2.6.3 User .....	48
2.6.4 Decision Support System characteristics .....	54
2.6.5 Outcomes .....	57
2.7 Strength of relationships .....	57
2.8 Previous models of reliance .....	59
2.8.1 Empirically tested models .....	59

2.8.2 Theoretical models (generated from literature searches) .....	61
2.8.3 The Theory of Technology Dominance .....	63
2.9 Conceptual model of reliance and Automation Bias.....	64
2.9.1 Model for automation reliance .....	65
2.9.2 Decision making process .....	66
2.9.3 Ontology of Automation Bias .....	67
2.10 Summary .....	86
3. Systematic Review .....	87
3.1 Introduction.....	87
3.2 Review aim and objectives .....	88
3.3 Review methods.....	88
3.3.1 Sources of studies.....	88
3.3.2 Search strategy .....	89
3.3.3 Search terms:.....	89
3.3.4 Databases .....	90
3.4 Eligibility criteria for studies .....	91
3.4.1 Inclusion criteria .....	91
3.4.2 Outcome measures .....	92
3.4.3 Exclusion criteria .....	92
3.4.4 Quality assessment.....	93
3.5 Data extraction .....	96
3.5.1 Draft data extraction form.....	97
3.6 Findings.....	98
3.7 Results.....	99
3.7.1 Automation Bias rate .....	99
3.7.2 Causes of Automation Bias.....	103
3.7.3 Automation Bias avoidance .....	105
3.8 Systematic review conclusion.....	107
3.9 Summary .....	108
4. Discussion and Conclusions of Reviews .....	109
4.1 Introduction.....	109
4.2 Rate and influencers of Automation Bias .....	110
4.3 Automation Bias influencing research factors .....	112
4.4 Diagram hypotheses.....	113

4.5 Study design.....	114
4.6 Summary .....	115
5. Empirical Study .....	116
5.1 Introduction.....	116
5.2 Setting .....	116
5.3 Participants.....	116
5.3.1 Sample size calculations .....	116
5.4 Design .....	119
5.5 Procedure .....	121
5.6 Resources .....	123
5.7 Clinical Decision Support simulation .....	123
5.7.1 Simulator.....	123
5.8 Ethics .....	126
5.9 Scenarios.....	127
5.9.1 Scenario generation.....	127
5.9.2 Scenario validation.....	129
5.9.3 Scenario presentation .....	133
5.10 Pilot study .....	133
5.11 Participant recruitment.....	136
5.11.1 Stages of recruitment .....	138
5.12 Summary .....	139
6. Results .....	140
6.1 Introduction.....	140
6.2 Answer coding .....	140
6.3 Response rate .....	142
6.4 Data preparation.....	143
6.5 Descriptive statistics of participants .....	144
6.6 Results.....	146
6.6.1 Proportion of correct advice.....	146
6.6.2 Influencing factors: .....	149
6.6.3 Qualitative analysis.....	155
6.7 Summary .....	156
7. Overall Discussion.....	157
7.1 Introduction.....	157

7.2 Results summary and discussion.....	157
7.2.1 Reviews and models.....	157
7.2.2 Empirical study .....	157
7.3 Limitations .....	164
7.3.1 Reviews of literature .....	164
7.3.2 Ontology of Automation Bias .....	165
7.3.3 Empirical study .....	165
7.4 Summary .....	170
8. Conclusion .....	172
8.1 Introduction.....	172
8.2 Achievements based on aims and objectives .....	172
8.3 Factors to consider / Recommendations .....	174
8.4 Addition to literature.....	175
8.4.1 Outputs and benefits of the literature reviews.....	176
8.4.2 Benefits of the study overall .....	176
8.5 Future work: Designing Clinical Decision Support for appropriate reliance.....	177
8.5.1 Possible extra information to include.....	177
8.5.2 Background .....	178
8.5.3 Aim .....	180
8.5.4 Methodology .....	180
8.5.5 Results/ Analysis.....	182
8.6 Summary .....	183
Appendix A: Conceptual model of reliance and Automation Bias.....	184
Appendix B: Quality assessment checklist for systematic review.....	185
Appendix C: Final extracted papers from systematic review .....	188
Appendix D: Invitation email .....	232
Appendix E: Information sheet.....	233
Appendix F: Consent form.....	235
Appendix G: Debrief information.....	236
Appendix H: City University ethical approval .....	237
Appendix I: NHS REC ethical approval .....	238
Appendix J: Twenty final prescribing scenarios.....	241
<i>References</i> .....	261

## List of Figures

Figure 1.1 The study methodology as influenced by the Medical Research Council framework for complex interventions.....	24
Figure 2.1 Model of empirical results – strengths of relationships.....	60
Figure 2.2 Framework for automation.....	61
Figure 2.3 Model of reliance.....	62
Figure 2.4 Theory of Technology Dominance.....	63
Figure 2.5 Conceptual model of reliance and Automation Bias.....	65
Figure 2.6 Breakdown of the evaluation process.....	66
Figure 2.7 Taxonomy of basic categories in the DOLCE ontology.....	70
Figure 2.8 The different types of OWL properties .....	76
Figure 2.9 The enduring branch of the ontology as illustrated in OntoGraf.....	81
Figure 2.10 The perdurant branch of the ontology as illustrated in OntoGraf .....	82
Figure 2.11 The quality branch of the ontology as illustrated in OntoGraf.....	83
Figure 3.1 Venn diagram of search concepts.....	87
Figure 3.2 RevMan meta-analysis .....	102
Figure 3.3 Funnel plot to assess publication bias .....	102
Figure 4.1 Diagram of factors to include in study and possible directions of relationships..	112
Figure 5.1 Test procedure for each case .....	122
Figure 5.2 Exponential discriminative snowball sampling.....	137
Figure 6.1 Flow chart of recruitment dropout rate.....	143
Figure 6.2 Age range and gender of participants.....	144
Figure 6.3 Participant-stated frequency of use .....	145
Figure 6.4 Participant-stated general trust in Clinical Decision Support Systems .....	145
Figure 6.5 Number and type of decision scenarios by participant.....	148
Figure 6.6 Comparison of trust ratings by participant .....	150
Figure 6.7 Mean number of decision switches by high versus low trust and confidence.....	152
Figure 6.8 Relationship between years of clinical experience and decision switching .....	153

## List of Tables

Table 2.1 Percentage of hospitals at each stage of Electronic Medical Record adoption .....	31
Table 2.2 Table to show results of the Unified Model of Technology Acceptance .....	59
Table 2.3 Ontology classes, definitions and the DOLCE ontology class type .....	70
Table 2.4 Ontology properties .....	77
Table 2.5 Conceptual model properties .....	79
Table 3.1: Combinations of search concepts in the systematic review.....	91
Table 3.2 Stages of the systematic review and numbers of papers found per search engine ..	95
Table 3.3: Landis and Koch table for interpreting k values.....	96
Table 3.4: Column headings within the systematic review extraction table.....	97
Table 3.5 Profile of papers found in systematic review.....	98
Table 5.1 Correlation of difficulty ratings between 3 raters over 35 scenarios.....	130
Table 5.2 Correlation of difficulty ratings between 3 raters over 20 scenarios.....	131
Table 5.3 Correlation of difficulty ratings between 2 raters over 20 scenarios.....	132
Table 5.4 Correlation of difficulty ratings over all 5 raters over final 20 scenarios.....	132
Table 5.5 Example scenarios with difficulty and sample correct and incorrect answers .....	133
Table 6.1: Response rate (opening and displaying email, and clicking to website) .....	142
Table 6.2 Changes in scenario prescription pre and post advice.....	147
Table 8.1 Output benefits by target group – literature reviews .....	176
Table 8.2 Benefits for different groups from the overall study .....	176

## **Acknowledgement**

I would like to thank my first and second supervisors Professor Abdul Roudsari and Professor Jeremy Wyatt for their invaluable advice, guidance and support through the duration of this PhD research.

I would also like to thank my colleagues, in particular Dr Kamran Golchin for his skills and boundless patience with designing and programming the Decision Support Simulator, and Dr Peter Weller, Dr Sanaa Henni and Gill Smith for their help and encouragement.

Thank you also to my family and friends for their support and understanding throughout the project.

## **Declaration**

I grant powers of discretion to the University Librarian to allow this thesis to be copied in whole or in part without further reference to me. This permission covers only single copies made for study purposes subject to normal conditions of acknowledgement.

## Glossary

Automation Bias : A bias towards accepting computerised information in a heuristic manner. This can lead to systematic errors if the automation is not fully reliable.

Automation Bias error : A systematic error stemming from over-reliance on automated advice which is incorrect. Can be defined as commission error or omission error.

Clinical Decision Support System : Computer software, which is designed to assist physicians and other health professionals with decision making tasks.

Cognitive fit : Correspondence between task, information presentation format and individual cognitive style.

Commission error : Error resulting from when a piece of incorrect advice is erroneously followed.

Decision accuracy : The degree to which a decision conforms to the standard accepted correct response.

Decision Support System : software, which is designed to assist users with decision making tasks.

Ecological validity : The degree to which the behaviours observed and recorded in a study reflect the behaviours that actually occur in natural setting.

ePrescribing : Electronic transmission of prescription information from the prescriber's computer to a pharmacy computer. It replaces a paper prescription that the patient would otherwise carry or fax to the pharmacy.

Heuristic : Cognitive rule-of-thumb, or mental shortcut that allows people to make decisions and solve problems quickly and efficiently with incomplete information.

**Judge Advisor System** : A type of advice structure often studied in advice taking research. The two roles in a JAS are the judge and advisor roles. The judge is the decision maker who evaluates information concerning a particular decision and makes the final judgment on the decision outcome. The advisor is an individual who provides advice, information, or suggestions to the judge.

**Judgement and Decision Making** : A field of psychological research investigating the process of making decisions and judgements.

**Omission error** : Error which occurs when appropriate action is not taken because the user was not prompted by automation, or following false negative advice.

**PICO framework** : Abbreviation for Population, Intervention, Comparison, Outcome. This is a technique used in evidence based medicine to frame and answer a clinical question.

**Satisficing** : A decision-making strategy that aims for a satisfactory or adequate result, rather than the optimal solution.

## Abstract

*Purpose:* Computerised clinical decision support systems (CDSS) are implemented within healthcare settings as a method to improve clinical decision quality, safety and effectiveness, and ultimately patient outcomes. Though CDSSs tend to improve practitioner performance and clinical outcomes, relatively little is known about specific impact of inaccurate CDSS output on clinicians. Although there is high heterogeneity between CDSS types and studies, reviews of the ability of CDSS to prevent medication errors through incorrect decisions have generally been consistently positive, working by improving clinical judgement and decision making. However, it is known that the occasional incorrect advice given may tempt users to reverse a correct decision, and thus introduce new errors. These systematic errors can stem from *Automation Bias (AB)*, an effect which has had little investigation within the healthcare field, where users have a tendency to use automated advice heuristically.

Research is required to assess the rate of AB, identify factors and situations involved in overreliance and propose ways to mitigate risk and refine the appropriate usage of CDSS; this can provide information to promote awareness of the effect, and ensure the maximisation of the impact of benefits gained from the implementation of CDSS.

*Background:* A broader literature review was carried out coupled with a systematic review of studies investigating the impact of automated decision support on user decisions over various clinical and non-clinical domains. This aimed to identify gaps in the literature and build an evidence-based model of reliance on Decision Support Systems (DSS), particularly a bias towards over-using automation. The literature review and systematic review revealed a number of postulates - that CDSS are socio-technical systems, and that factors involved in CDSS misuse can vary from overarching social or cultural factors, individual cognitive variables to more specific technology design issues. However, the systematic review revealed there is a paucity of deliberate empirical evidence for this effect.

The reviews identified the variables involved in automation bias to develop a conceptual model of overreliance, the initial development of an ontology for AB, and ultimately inform an empirical study to investigate persuasive potential factors involved: task difficulty, time pressure, CDSS trust, decision confidence, CDSS experience and clinical experience.

The domain of primary care prescribing was chosen within which to carry out an empirical study, due to the evidence supporting CDSS usefulness in prescribing, and the high rate of prescribing error.

*Empirical Study Methodology:* Twenty simulated prescribing scenarios with associated correct and incorrect answers were developed and validated by prescribing experts. An online Clinical Decision Support Simulator was used to display scenarios to users. NHS General Practitioners (GPs) were contacted via emails through associates of the Centre for Health Informatics, and through a healthcare mailing list company.

Twenty-six GPs participated in the empirical study. The study was designed so each participant viewed and gave prescriptions for 20 prescribing scenarios, 10 coded as “hard” and 10 coded as “medium” prescribing scenarios (N = 520 prescribing cases were answered overall). Scenarios were accompanied by correct advice 70% of the time, and incorrect advice 30% of the time (in equal proportions in either task difficulty condition). Both the order of scenario presentation and the correct/incorrect nature of advice were randomised to prevent order effects.

The planned time pressure condition was dropped due to low response rate.

*Results:* To compare with previous literature which took overall decisions into account, taking individual cases into account (N=520), the pre advice accuracy rate of the clinicians was 50.4%, which improved to 58.3% post advice. The CDSS improved the decision accuracy in 13.1% of prescribing cases. The rate of AB, as measured by decision switches from correct pre advice, to incorrect post advice was 5.2% of all cases at a CDSS accuracy rate of 70% - leading to a net improvement of 8%.

However, the above by-case type of analysis may not enable generalisation of results (but illustrates rates in this specific situation); individual participant differences must be taken into account. By participant (N = 26) when advice was correct, decisions were more likely to be switched to a correct prescription, when advice was incorrect decisions were more likely to be switched to an incorrect prescription.

There was a significant correlation between decision switching and AB error.

By participant, more immediate factors such as trust in the specific CDSS, decision confidence, and task difficulty influenced rate of decision switching. Lower clinical experience was associated with more decision switching (but not higher AB rate). The rate of

AB was somewhat problematic to analyse due to low number of instances – the effect could potentially have been greater. The between subjects effect of time pressure could not be investigated due to low response rate.

Age, DSS experience and trust in CDSS generally were not significantly associated with decision switching.

*Conclusion:* There is a gap in the current literature investigating inappropriate CDSS use, but the general literature supports an interactive multi-factorial aetiology for automation misuse. Automation bias is a consistent effect with various potential direct and indirect causal factors. It may be mitigated by altering advice characteristics to aid clinicians' awareness of advice correctness and support their own informed judgement – this needs further empirical investigation. Users' own clinical judgement must always be maintained, and systems should not be followed unquestioningly.

## 1. Introduction

Information and Communication Technology (ICT) has great potential to improve the medical and healthcare process, both in terms of management and outcomes, by improving the efficiency and effectiveness of services. In many settings, such as primary care, there has been a significant rise in the implementation and use of technology<sup>1</sup> (cited in Coiera, 2006<sup>2</sup>). Electronic Health Records (EHR), Tele-Medicine, and Clinical Decision Support Systems (CDSS) are just some of the interventions which have been introduced to improve the quality, safety and effectiveness of the healthcare service.

Medical decision making is a fundamental component of the healthcare pathway. However, errors in this process have been shown in many studies to be a large component of overall medical errors. For example, in one prospective one month study of 36,200 prescriptions, 1.5% were found to have a prescribing error, a quarter of which were serious. When the serious errors were examined, 58% of the errors originated in the prescribing decision (the remaining 42% in medication order writing)<sup>3</sup>. Flaws in medical decision making stem from a number of potential causes: from underlying human cognitive limitations, slips or gaps in knowledge, or problems with the healthcare workflow<sup>4</sup>. Medication knowledge deficiency is believed to be one of the most common contributing factors in prescribing error<sup>5</sup>.

Technological innovation is a valid way to remedy this, for example by covering for gaps in knowledge, or acting as reminders. Thus CDSS can be a helpful way to improve decision quality. However, in part due to complexities of healthcare, decision support is not infallible, and so complete reliance on its output can be misguided. To optimise the benefits gained it is wise to investigate the nature of new errors the automation introduces through examining what causes inappropriate reliance. This study will focus on overreliance in particular.

It has been found that the use of CDSS can improve clinical decisions, leading to improved patient outcomes and a more effective healthcare service<sup>6,7,8</sup>, despite a low uptake in many settings. However, the implementation of CDSS may incur unexpected outcomes by introducing detrimental, machine-related errors.

This project aims to examine the nature of automation bias (AB) which can lead to machine related error. This cognitive bias has been little studied in clinical settings; the research

ultimately aims to add to the understanding of this kind of systematic error, and so aid looking at ways in which clinician judgement about the accuracy of decision advice can be appropriately calibrated to make optimal use of the technology.

**Overall aim:** To investigate the rate of AB related error, the factors associated with it, and to help inform further studies into enhancing the ability of users of DSS users to detect bad advice to decrease over-reliance / misuse.

**Overall hypothesis:** AB is a frequently occurring and replicable effect. The effect has various potential direct and indirect causal factors, and errors could potentially be mitigated by altering CDSS design and/or advice characteristics to help clinicians be aware of when they receive correct or incorrect advice.

### **1.1 Background and aims**

DSSs have great potential to improve clinical decisions, actions and patient outcomes<sup>6,9</sup>, by providing advice, filtered or enhanced information, or by providing prompts or alerts to the user. However, most studies have emphasised the accuracy of the computer system alone, without placing clinicians in the role of direct users. To the researcher's knowledge there is no comprehensive or systematic review of the overall accuracies of CDSSs, however it is known that occasional incorrect pieces of advice they give they may tempt users to reverse a correct decision they have already made, and thus to introduce new errors<sup>2</sup>. These errors can be a result of *automation bias*<sup>10</sup>, in which users tend to accept computer output without sufficient thought, or the opposite problem of *errors of dismissal*, where helpful computer advice is ignored. Clinicians routinely disable or ignore the alarms or alerts on clinical monitoring devices<sup>11</sup> for a variety of reasons, such as unsuitable content, and interruption to workflow<sup>12</sup>. CDSS may also result in errors where clinicians draw incorrect assessments of the evidence – possibly shaped by cognitive decision biases.

### **1.2 Automation Bias**

Previous investigations into automation bias have primarily focussed on the aviation and motoring fields. Investigations into the human factors involved in healthcare systems is a relatively more recent field and until a recent paper (Goddard et al, 2011)<sup>13,14</sup>, no systematic reviews have been found on this phenomenon relating to healthcare or in general.

Medical error has been considered the third most frequent cause of death in Britain after cancer and heart disease<sup>15</sup>. More people are killed in preventable hospital error than on the roads<sup>16</sup>. In 2004 the National Patient Safety Agency (NPSA) produced a report that stated 10% of patients admitted to hospitals would suffer a patient safety incident - almost one million people in 2002/3 - and up to half of these could have been prevented. It added that 72,000 of these incidents may have contributed to the death of the patient<sup>17</sup>. As an example, between 2005 and 2010 there were 1,085 reports of incidents to the Medicines and Healthcare products Regulatory Agency (MHRA) involving infusing devices; 21% were attributed to user error (this may have been higher as in 68% of cases, no cause was established)<sup>18</sup>.

Intervention in the form of computerisation is often presented as a solution, however computer systems can introduce new errors, such as, and stemming from, latent errors in design and content. If users are unaware of such errors and the DSS is used inappropriately, this can render the intervention less effective or at worst dangerous. Some critics of DSS argue that if it is necessary to check the advice, then perhaps the DSS is less useful.

CDSS may sometimes be ineffective. A 2011<sup>19</sup> study into patients with vascular conditions found that, versus a control group, the use of a web-based vascular tracking and clinical support application did not significantly improve patient outcomes. Researchers concluded that “clinicians are correct to remain sceptical about the cost effectiveness of [clinical decision support] systems and should continue to demand evidence that they improve patient outcomes”. Reviews often show improved physician performance (as measured by reduction in errors, or compliance with protocols and guidelines<sup>6,20</sup> but little difference to patient outcomes.

In a follow up to a seminal 1993 paper looking at ineffective automation use<sup>21</sup>, Parasuraman and Riley (1997)<sup>22</sup> discussed the sorts of errors which may occur via anecdotal evidence and results from various empirical studies. Three aspects of ineffective human use of automation are discussed: disuse (underuse), misuse (over reliance) and abuse (improper application of support). Disuse and misuse can be seen as two opposite ends of an automation usage spectrum; errors of dismissal have been described as a form of anti-automation bias. The majority of the literature reports disuse (and non-acceptance) of automation, which is a much more extensive field of study.

In a study looking at the enhancement effects of clinical decision making by the use of 2 computerised diagnostic support systems (the Quick Medical Reference (QMR) and Iliad systems), Friedman et al (1999)<sup>23</sup>, noticed that in some cases, clinicians would override their own correct decisions in favour of the erroneous advice from the DSS – in 12% of cases the DSS caused the doctor to put the correct diagnosis on their list, but in 6% of cases their own correct decisions were dropped in favour of the erroneous computer-generated advice – a net gain of 6% overall (with a slightly higher net gain with QMR).

Skitka et al (1999)<sup>24</sup> introduced the term “automation bias” when studying the effects of incorrect computer advice on decisions taken by students (lab aviation simulation). “Automation bias” was defined as “the use of automation as a heuristic replacement for vigilant information seeking and processing”<sup>25</sup>. Students given unreliable advice (in some tasks) made more errors than those not given automated advice. Even when told the tool was not 100% reliable, people still often used the advice even when incorrect or contradicted – they interpreted these errors as “automation bias”. Errors of omitting the correct answer because the DSS failed to advise anything were explained in terms of complacency or reduced vigilance.

Reason (1990)<sup>26</sup> postulates that there are 2 main types of error which stem from different cognitive causes; those that are knowledge based, and those that are skill based.

The Skitka study identified two types of AB error, which can be a result of either of these error types:

- Commission errors - follow bad advice, or following false positive advice
- Omission errors – appropriate action not taken because not prompted by automation, or following false negative advice. Inaction most often involves slips, lapses, and mistakes.

Omission errors comprise omitting the entire task or steps in a task. Commission errors can involve committing a selection error (e.g. wrong selection, misposition, issuing of wrong command), errors of sequence, timing errors (too late or early), or qualitative (too little or too much).

Often the literature has looked solely at overall clinical or DSS accuracy and clinical outcomes without investigating the aetiology and types of errors. More recent papers have started looking at human factors involved in appropriate design and use of automation in general – for example the factor of trust<sup>27</sup>, and other social, cognitive and motivational factors. The research often touches on the concept of AB without explicit definition or investigation. Due to the concept being relatively and new and undefined in terms of a field of study, a number of synonyms have been used in the literature– automation-induced complacency<sup>28</sup>, over-reliance on automation<sup>22</sup> and confirmation bias<sup>29</sup> are some of the various terms used to describe the concept of AB.

Though a relatively new concept for explicit empirical study, AB has plenty of anecdotal evidence. For example a quarter of financial advisors believe that the US stock market plunge in May 2010 was caused by an overreliance on computer systems for financial trading<sup>30</sup>. The Turkish Airlines Flight 1951 crash in 2009 was attributed to overreliance on a faulty radio altimeter<sup>31</sup>. Some pilots, convinced that advanced electronic navigation systems coupled with flight management system computers, or over-reliance on them, are partially responsible for these accidents, have termed these CFIT accidents "computerized flight into terrain". The European Joint Aviation Authorities' Future Aviation Safety Team has identified "crew reliance on automation" as the top potential safety risk in future aircraft<sup>32</sup>. This kind of evidence may also implicate certain factors which may increase the risk of AB (such as high pressure situations).

### **1.2.1 Increasing accuracy of judgement**

The psychology of human-computer interaction provides useful insights into how information systems can be designed to provide optimal behavioural outcomes, particularly the Judgement and Decision making (JDM) literature surrounding advice-taking. The field of health informatics can benefit from the application of these tools and theories to improve usability and usefulness of clinical DSS, which is subject to human cognitive perceptual and attentional filters and cognitive constraints. Of relevance here, research into advice-taking and the Judge-Advisor System (JAS) paradigm<sup>1</sup> systematically investigates how advice affects the decision-making process.

---

<sup>1</sup> A JAS is a group in which one member has the role of decision-maker and other members act as advisors. The experimental paradigm often involves recording pre and post advice information.

Being socio-technical systems, final decisions involving DSS are an outcome of the user, the technology and the context i.e. task and clinical setting. Numerous factors have been investigated in the psychology and decision making literature into the factors affecting advice taking and post advice accuracy. For example, types of cognitive processing of advice may affect final user decisions – cognitive dual processing theories posit a split between more effortful logical processing and quicker, rule-of-thumb based processing. Bottom up factors, such as designing sophisticated automation that suggests an uncertain course of action seems to encourage people to accept the imperfect advice, even though information to decide independently on a better course of action is available<sup>33</sup>. Rather than using their own knowledge or more effortful processing, clinicians may view and use automation advice as a heuristic for the correct response even when said advice is erroneous. Appropriate cognitive engineering has been posited as a method of overcoming this, by enhancing performance of cognitive tasks by means of a number of intervention types, methods such as user-centred design of human-machine interaction, and/or work redesign to manage cognitive workload and increase human reliability.

Main user-focussed factors that increase advice utilisation are posited to include judge's self-confidence, trust in the source of advice, and judge or advisor expertise level (e.g. Azen and Budescu, 2003<sup>34</sup>). Accountability is also a factor which affects adherence to DSS advice and cross verification behaviour, increasing both<sup>35</sup>. Research has shown that judges' post-advice decision accuracy is related to the weight the judge gives to each advisor's recommendation<sup>36</sup>. The technology design can also affect decision through a number of features which affect usability and effective use of the system such as the interface design, the decision support rules, and training. When judges and advisors have more decision-relevant information, they are on average more accurate - judges become more capable of discriminating between good and bad advice (weighing the former more highly). The setting, particularly in some clinical situations, is one fundamental factor in the "goodness-of-fit" or "cognitive fit" of a DSS: factors such as time constraints and user distraction in a busy environment where multi-tasking may be required are potential situations where incorrect use of a DSS may occur.

There are a number of ways that CDSS developers might seek to help users recognise when the CDSS advice is likely to be wrong and so avoid these types of errors of dismissal and

automation bias. These methods involve giving the user extra information in addition to the advice, for example:

- Giving an explanation of the advice e.g. how the prescriptive information was inferred (decision relevant information). If the user detects that the explanation is suspect or there is a missing/inaccurate component to the knowledge base, then they are more likely to realise that the advice is wrong too.
- Giving the user a calculated probability estimate of the system's certainty about its advice may also help discriminate good and bad advice, e.g. "The most likely diagnoses are acute appendicitis (60%) and non specific abdominal pain (30%)". These may be drawn from simple base rates.

A key question for developers, users, patients and those who purchase decision support systems is which of these methods is most likely to lead to users correctly adhering to correct advice and ignoring incorrect advice. Some of the extra information is likely to affect users in avoid faulty advice; others to encourage trust in "correct" advice.

To answer this question requires a carefully designed empirical study. While real-life study (set in a primary care environment) would give reliable results, it does not allow manipulation of factors under study (further reasoning for the methods used is discussed in Chapter 4) , as laboratory-based before-after studies of the impact of advice on simulated cases<sup>37</sup> can reveal which factors have the greatest potential to impact clinical decisions. To allow realistic sample sizes, such a study needs to be carried out using a simulated CDSS which issues incorrect advice more frequently than would usually be the case.

### **1.3 Study aims and objectives**

The study aim is to improve the safety, usability, clinical acceptance and effectiveness of CDSS by helping users to detect when to adhere to or ignore CDSS output, by:

- Investigating and testing the ability of users to detect bad advice
- Investigating the risk factors leading to overreliance on automation via a literature and systematic review
- Investigating the influence of potential risk factors in an empirical study
- Proposing follow on studies to investigate the effect of interventions to avoid AB e.g. the addition of confidence information, or source for advice etc.

- Formulating recommendations for CDSS developers about how to make the output from the CDSS more transparent to users

#### **1.4 Purpose and structure of this thesis**

Chapters 1 and 2 of this thesis provide definitions for DSS, and a description of the gaps in the literature that will be addressed, including the problem of medication errors (and the pros and cons of DSS interventions). A review of the prescribing domain outlines the high variability and inconsistency of prescribing decisions. In addition potential difficulties recruiting for research in this area are discussed in terms of General Practitioner (GP) response rates. This is followed by an examination of the factors postulated in the literature to affect reliance on automation for judgment and decision making which are worked into a conceptual broad model of the causes of AB. This then leads to the development of a draft skeleton ontology of AB.

Chapter 3 describes a more targeted systematic review which involved a comprehensive search of the literature to examine the existence of empirical evidence for overreliance on automation.

Chapter 4 summarises the results of the literature reviews and the gaps in knowledge. A testable model is described to clarify the hypotheses elicited by the reviews relating to factors which affect the tendency to over-rely on automated advice. The chapter also discusses the design of the study and other designs which were considered and rejected.

Chapter 5 describes the methodology to empirically examine AB in relation to Primary Care Prescribing. Chapter 6 outlines the results of the empirical study. The pilot study and results are also described.

Chapter 7 contains the overall discussion which relates results back to the principle issues in the literature, and describes the challenges and limitations of the study.

Chapter 8 reflects back on the extent to which study objectives have been achieved, and clarifies the contribution to knowledge. Future work is suggested to investigate CDSS factors which could mitigate AB and encourage appropriate reliance.

### 1.4.1 Justification for study and methods

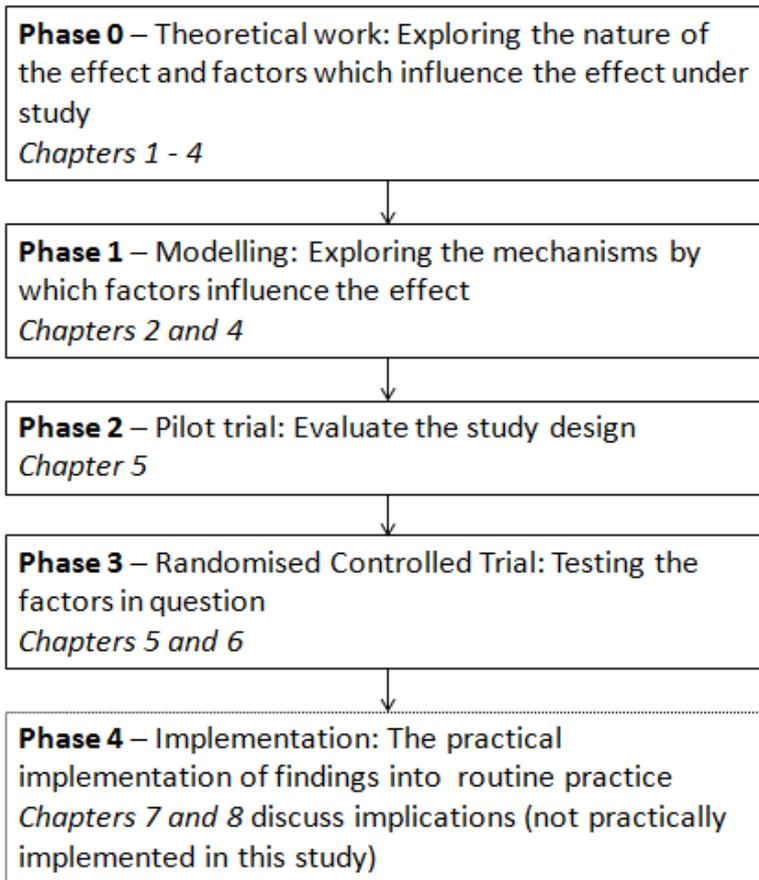
a) Why study automation bias?

- Medication errors are a major risk in healthcare, and prescribing errors in particular are an area for improvement
- There has been a slow but increasing uptake of Healthcare ICT<sup>38,39</sup>. DSS are getting more important in NHS, for example, with the uptake of ePrescribing (with additional decision support) – in primary care ePrescribing is now the norm and the electronic transmission of prescriptions from GP to high street pharmacies is being implemented through the Electronic Prescription Service (EPS)<sup>40</sup>
- Increasing evidence of DSS failures<sup>41</sup>, but a lack of targeted empirical evidence for rates or causal factors
- Opportunity to inform and improve interface design and implementation factors

The project will be carried out in 3 broad stages. First, a literature review and will provide a conceptual diagram of the causes and possible consequences of AB, stage two involves using the model to identify effect modifiers and carry out a systematic review and, and stage three involves the design and implementation of empirical studies using insights from stages one and two.

b) Why carry out the project this way?

The overall methodology of this study was informed by the Medical Research Council (MRC) framework for complex interventions (see Campbell, 2007<sup>42</sup>). The MRC framework outlines first carrying out theoretical research into the factors affecting the study, generating a model of how the system under investigation works. This should be followed by a pilot study to optimise the trial measures, followed by the final study. Figure 1.1 illustrates the stages of the framework with the Chapters in this thesis which cover these stages. Campbell (2007) notes that the stages can be approached serially, or in parallel by combining stages 0-2 into a larger, more iterative activity to develop understanding of the problem.



**Figure 1.1** The study methodology as influenced by the Medical Research Council framework for complex interventions

## **2. Literature Review**

### **2.1 Introduction**

This review chapter describes DSS and their uses. The problems faced in the field of prescribing, such as high variability and high error rate, are also discussed alongside the positive and negative impact of DSS intervention; the increasing use of technology in healthcare highlights this may be a particularly important field of study. The potential difficulties recruiting GPs for this research area are discussed.

Overreliance on automated advice and the factors which may influence the rate are outlined and possible relationships are drawn. These are illustrated by a conceptual model of factors in the literature.

A possible application of the conceptual model is demonstrated in the development of a draft ontology in the Protégé open-source tool.

### **2.2 Decision Support Systems**

DSS are ICT systems that support decision making processes. They generically contain knowledge and reasoning components which use the contextual input to generate advice. Simple CDSSs provide narrative information requiring further processing and analysis before clinical decisions are made. Recently, CDSSs have become increasingly sophisticated by matching patient information with computerised knowledge and using algorithms to generate patient-specific assessments or treatment recommendations.

CDSS have been defined as automated tools that generate advice about a patient using two or more of their data items<sup>43</sup>. They provide computerised provision of assessments, advice or prompts specific to the problem and informed by a knowledge base on the basis of individual problem features (data). For example, the NHS Clinical Assessment System was based on hundreds of guidelines and used by nurses to triage over 7 million callers per year at 22 NHS Direct call centres<sup>44</sup>. Many types of DSS exist and provide relevant information to advise this study, thus shall be included within the scope of the initial research.

DSS have been found to be of benefit in certain domains. For example, a recent<sup>45</sup> retrospective medical imaging study showed found that the implementation of a CDSS built into ordering systems for selected high-volume imaging procedures resulted in reduction of inappropriate medical imaging after intervention: 23.4% lower for low back pain lumbar MRI, 23.2% lower for headache head MRI, and 26.8% lower for sinusitis sinus CT.

### 2.2.1 Types of Decision Support Systems

By the action required by the user:

*Passive:* In these types of systems, the user must make an explicit request to the system for information i.e. a description of the patient's case which the CDSS returns advice for. Two types of system fall under this category:

- consultant systems which use patient state information to provide advice
- critiquing systems which use patient information and therapeutic or investigative plan to make a critique of the physician's proposals.

*Semi-active:* These systems are invoked automatically and act as "watchdog" type systems (automatic reminders or alarms), providing information, knowledge and or/ procedural rules when triggered.

*Active:* These systems can provide advice tailored to specific patients and can work totally automatically, without the input of the user and can provide alerts for additional examinations, continuous therapeutic examinations (e.g. a pacemaker) or surgical assistance.

By the type of inference process:

From the literature, inference mechanisms used in CDSSs include rule based (use chains of Boolean "if-then" rules to reach a conclusion), Bayesian (use probabilities to predict diagnoses), heuristic (include statistical measures such as the Support Vector Machine<sup>2</sup>), neural networks (mainly used during DSS development; black box modelling technique that models relationships by learning from historical data), genetic algorithms (uses iterations to extract the best solutions) and case-based (sometimes medical knowledge is difficult to be modelled; medical experts can use concrete examples to express knowledge. In this situation, the case-based reasoning (CBR) approach is used in CDSSs)<sup>46</sup>.

---

<sup>2</sup> Support Vector Machine: a concept in statistics and computer science for a set of related supervised learning methods which analyse data and recognise patterns. Used for classification and regression.

### **2.3 Domain for simulation - prescribing**

CDSSs support a range of prescribing practice activities including drug treatment selection, and checking allergies and drug interactions. Additionally CDSS can be applied to other aspects of the prescribing process, such as monitoring and stopping therapies.

Prescribing is one of the most common tasks in daily general practice. It also has much potential for error. Evidence exists which indicate prescribing decisions are often suboptimal<sup>47</sup>, and this has been found in hospital and primary care settings<sup>48</sup>. Medication errors are the third most prevalent types of patient safety errors in England<sup>49</sup> and prescribing error is the biggest cause of medication error<sup>50,51</sup>.

Sayers et al (2009)<sup>52</sup> found 12.4% prescriptions contained one or more errors and 6.2% drug items contained one or more errors. Of the errors the majority were minor (72.9%), a smaller number (24.7%) were major nuisance errors, and 2.4% were potentially serious errors.

Gandhi et al (2005)<sup>53</sup> found 7.6% of outpatient prescriptions contained errors. More recently the General Medical Council (GMC) commissioned a major report<sup>54</sup> studying prescribing errors specifically by foundational trainees in inpatient hospitals in terms of rates and causes. Investigation methods included literature reviews, empirical evidence or prescribing errors in hospitals and qualitative exploration of the perceived causes of prescribing error. The research checked 124,260 medication orders across 16 hospitals for errors. Errors were detected by pharmacists during routine pharmacy on set data collections days, coding for error type and severity, these errors were then discussed at validation meetings and then recorded in a database. Twenty-eight prescribing errors were defined, most common being “omission on admission”, “overdose” and “underdose”. Error severities were coded as minor, significant, serious or potentially lethal. The report found that 11077 of these medication orders contained errors – a mean error rate of 8.9%. Almost 2% of the errors were classified as potentially lethal (with 5% serious, 53% significant, and 40% minor). Errors were made by all grades of doctor with the highest error rate (10.3%) found with FY2 doctors. The study reported that pharmacists had to intercept most of the errors, which meant that very few errors eventually harmed patients. However pharmacists were responsible for detecting and reporting the errors, and there was no further analysis on the errors the pharmacists themselves may have missed; it is possible that the error rate was higher. Data collection forms included a section for reporting actual patient harm, however, this was rarely completed and therefore analysis was not feasible.

As a follow up study, recent research into prescribing errors in Primary Care funded by the GMC has indicated that prescribing errors persist with GP e-prescribing<sup>55</sup>. A prospective study over 15 UK General Practices spanning 18 months examined 6,048 unique prescription items for 1,777 patients. Prescribing or monitoring errors were detected for one in eight patients, involving around one in 20 of all prescription items. The vast majority of the errors were of mild to moderate severity, with one in 550 items being associated with a severe error. The research indicated that pick lists of drugs, which arrange the options in alphabetical order, could easily lead to clinicians prescribing the wrong drug – these were cited as the most dangerous issues. A wide range of underlying causes of error were identified relating to the prescriber (pertinent to this study a factor mentioned was “(over)-reliance of decision support systems for alerts of drug interactions and contraindications” (pg 108), the team, the working environment, and the task. Defences against error were also identified, including strategies employed by individual prescribers and primary care teams, and making best use of health information technology (as part of the report’s literature review, decision support was cited as part of a complex intervention (as prescribing errors stem from multifactorial causes) to reduce the risk of prescribing error). As part of the same collaboration group, a recent article in the *Lancet*<sup>56</sup> investigated the effect of a pharmacist-led information technology intervention on medication errors in a cluster randomised, controlled trial over 72 General Practices. The practices were allocated to either computer-generated simple feedback (control) or a pharmacist-led complex information technology intervention (PINCER), composed of feedback, educational outreach, and dedicated support. The cost per error avoided was also estimated by incremental cost-effectiveness analysis. The study concluded that the PINCER intervention was an effective method for reducing a range of medication errors in general practices with computerised clinical records. The intervention was also found to have a 95% probability of being cost effective if the decision-makers ceiling willingness to pay reached £75 per error avoided at 6 months. Resulting from this research, the GMC has called for smarter software to help GPs reduce prescribing errors. This research aimed to corroborate and supplement these findings; CDSS can benefit prescribing by reducing medication error, but designers, implementers and clinicians have to be aware that new errors may arise due to over-reliance of advice, and that this effect can be influenced by a number of factors suggested in the literature.

The domain of using DSS for prescribing in Primary Care will be used for this study as it is a common field of error, and also has much scope for a wide range of potential simulated patient scenarios. It is an example of multi attribute decision making, resulting in a reasonably high rate of suboptimal choices.

It has also been shown that CDSS are mostly consistently effective in the area of prescribing. Pearson (2009)<sup>57</sup> carried out a systematic review to evaluate the impact of CDSSs on prescribing practice. In a review of 56 papers (38 addressing initiating, 23 monitoring and three stopping therapy); 88.5% of studies resulted in at least one positive outcome as a result of CDSS intervention, and 44.1% of studies led to  $\geq 50\%$  statistically significant outcomes. Due to heterogeneity in study methodology, comparison groups, setting, intervention targets, and outcomes, the authors reported on the impact of CDSS on measures relating to prescribing which were deemed indicative or a surrogate/proxy for ultimate patient outcomes, such as laboratory or monitoring tests relevant for the safe and appropriate use of particular medicines. Outcomes were then compared between control and intervention groups; whether the intervention favoured the CDSS or the comparison group, and whether this was a statistically significant result. Durieux et al. (2008)<sup>58</sup> carried out a Cochrane review on computerised advice for drug dosage, and found significant benefits, including reduced risk of toxic dose (rate ratio of 0.45) and reduced length of hospital stay (standardised mean difference -0.35 days). However, some studies that have examined the impact of CDSS on prescribing have reported no change in error rates<sup>59</sup>, or adverse drug events<sup>60</sup>. Some of these differences can be explained due to lack of standardised outcome measures for measuring errors.

Other domains, for example, that of diagnosis, also appear to have evidence for effectiveness, albeit weaker evidence than prescribing. A systematic review into the effect of CDSS on practitioner performance and patient outcome by Garg (2005)<sup>6</sup> found 10 trials which evaluated diagnostic systems. All studies measured practitioner performance; the CDSS was beneficial (statistically significant positive effect) in 4 studies (40%). Of the 5 trials assessing patient outcomes, none reported improvement. In the same review, 29 studies were found assessing prescribing systems - single-drug dosing improved practitioner performance in 15 of 24 studies (62%), and 2 of the 18 systems assessing patient outcomes reported an improvement.

In terms of attitudes towards automation relating to prescribing, negative attitudes towards e-prescribing systems are comparatively rare, but in an edition of JAMIA published in 2005 focusing on e-prescribing, Miller et al (2005)<sup>61</sup> urge a degree of caution:

"Clinicians should be wary of developing a false sense of security and unrealistic expectations based on use of e-prescribing applications alone, when more complex systems may be required."

### **2.3.1 Examples of prescribing Clinical Decision Support Systems**

CDSSs have almost 40 years of history, from first generation examples such as MYCIN developed in the early-mid seventies to aid infectious disease diagnosis and second to more recent CDSS such as Isabel and the web-based DXplain. Electronic prescribing (ePrescribing) with varying degrees of decision support is increasingly the norm in primary care within the UK<sup>62,63</sup> and is being increasingly used to improve patient safety, improve quality of care, and improve efficiency of healthcare delivery (e.g. cost savings)<sup>64,65</sup>. The UK has been rolling out the Connecting for Health Electronic Prescribing System (EPS), with the first system having gone live in Leeds Calverly Medical Centre in 2009. The EPS is developed to enable prescribers to send prescriptions electronically to a dispenser of the patient's choice. ePrescribing systems most often use decision support systems to provide medical information at the point of prescribing.

The most commonly used GP software system in the UK is the Egton Medical Information System (EMIS). The system has an optional integrated decision support module (Odyssey) which aids clinical assessments. EMIS Web for GPs (an integrated record system) is a recent development which incorporates an integrated prescribing CDSS; recent figures stated that 360 practices had implemented the system<sup>3</sup>. A number of standalone CDSS specifically tailored for the domain of prescribing exist. For example ScriptSwitch<sup>66</sup> is a point of care tool for Primary Care use, and operates by providing a recommended prescription, if a match is found (with the actions "Accept", "Edit Original", and "Prescribe Original"). It has been implemented in 6,500 GP practices across 138 NHS Primary Care Trusts (of 10,112 in the UK in 2010<sup>67</sup>).

The Prodigy CDSS is a guideline based tool for the support of chronic disease management. After diagnosis is made Prodigy provides medical advice and therapeutic suggestions.

---

<sup>3</sup> <http://www.ehi.co.uk/news/industry/7462/emis-web-reaches-360-practices>

Despite the general view that the presence of ePrescribing with elements of decision support is now commonplace in primary care, no recent overall UK-specific quantitative data were found in the literature to outline the implementation and uptake of Prescribing Decision Support in GP Practice. This lack of information is further complicated with the finding that provision of CDSS does not automatically imply uptake<sup>68</sup>.

In the US and Canada, healthcare IT and Electronic Medical Record (EMR) uptake is measured by the Healthcare Information and Management Systems Society (HIMSS) according to the EMR Adoption Model (EMRAM). Table 2.1 outlines the stages of EMRAM and the percentage of hospitals which have obtained them as of early 2012 in the US and Canada. The EMRAM is a tool that is used to evaluate the impact of the American Recovery and Reinvestment Act of 2009 (ARRA) funding on EMR adoption for 5 years. CDSS adoption is higher in the US than Canada. The majority of hospitals have installed CDSS for error checking.

**Table 2.1 Percentage of hospitals at each stage of Electronic Medical Record adoption**

Stage	Capabilities	US 2012 (N = 5318)	Canada 2012 (N = 639)
7	Complete EMR; Continuity of Care Document transactions to share data; Data warehousing; Data continuity with Emergency Department, ambulatory and outpatient care	1.2%	0.0%
6	Physician documentation (structured templates), <b>full CDSS</b> (variance and compliance alerts), full Radiology Picture Archiving and Communications System	5.2%	0.5%
5	Closed loop medical administration	8.4%	0.3%
4	Computerised Physician Order Entry, <b>CDSS (clinical protocols)</b>	13.2%	2.5%
3	Nursing/clinical documentation (flow sheets), <b>CDSS (error checking)</b> , Picture Archiving and Communications System	43.9%	36.2%
2	Clinical data repository, controlled medical vocabulary, <b>CDSS</b> , document imaging, Health Information Exchange capable	12.1%	21.9%
1	All ancillaries – laboratory, radiology, pharmacy - installed	5.5%	15.2%

0	All three ancillaries (laboratory, radiology, pharmacy) not installed	8.4%	23.5%
---	---	------	-------

### 2.3.2 International differences in prescribing

Medicine is not a globally standardised domain, variations can occur inter- and intra-nationally, over time and between prescribers. For example there are differences in evidence weighting, prescribing systems, drug names and there is evidence to show this is independent of patient characteristics.

This heterogeneity can render recruiting prescribers from different countries for studies unreliable. For promoting more generalisable results a homogenous participant sample will be aimed at, for example in terms of prescriber type and geopolitical location.

Several factors contribute to geographical differences in prescribing: from physicians' attitudes (which can involve issues such as diagnostic uncertainty, and time or market pressure), to socio-cultural and economic determinants (e.g. the demographic and morbidity/mortality profiles of an area or socioeconomic status), and the existing healthcare systems, which influence drug regulation and the national pharmaceutical market structure<sup>69</sup>.

#### 2.3.2.1 Geographical differences

Globally there are marked differences in prescribing, in terms of whether a drug is prescribed, and what is prescribed. Comparing international patterns of prescribing is very difficult as data sources are few, of uncertain accuracy, and often incomplete and thus of dubious comparability.

As a solution to this, Jolleys et al (1996)<sup>70</sup> used an alternative to direct prescribing information, defining pharmaceutical sales information as an indicator for comparison of Hormone Replacement Therapy (HRT) usage in the USA and Europe. Similar morbidity rates could be expected over the countries under study thus should not have been a confounding factor. The results showed a wide variation in the percentage of the eligible female population in each country calculated to have been taking HRT, from <1%–20%. HRT usage fell into three groupings: USA being the greatest user with UK and Scandinavian countries in the middle group and continental Europe having the lowest usage. The authors of

this study postulated that reasons for the discrepancy could be due to health beliefs and prescribers' and women's attitudes to HRT, and also access to healthcare.

Concordantly, in terms of the weak effect of population morbidity on prescribing patterns, in an observational study spanning 13 European countries, Butler et al (2009)<sup>71</sup> also found that variation in primary care clinical presentation did not explain high variation in antibiotic prescribing for acute cough. Antibiotic prescribing ranged from 20 – 90% (53% on average); they also found that the classes of antibiotic varied greatly. For example, amoxicillin was the most commonly prescribed antibiotic overall, but this ranged from 3% of antibiotics prescribed in Norway to 83% in England. When factoring out clinical presentation and demographics, the differences in antibiotic prescribing remained (from Norway odds ratio of 0.18, to Slovakia odd ratio of 11.2, at the 95% confidence interval). Similarly, additional prescribing variation was not related to clinically important differences in recovery; once clinical presentation was taken into account, this persisted – the rate of recovery was similar for patients whether or not they were prescribed antibiotics.

Fretheim et al (2005)<sup>72</sup> described prescribing patterns of antihypertensive drugs in ten countries (Canada, France, Germany, UK, US and the Nordic countries) via questionnaires and interviews with academics, drug regulatory agencies and MDs of drug companies. High variation was again found in prescribing patterns; thiazides accounted for 25% of consumption in the UK, contrasted with 6% in Norway. Conversely alpha blocking agents account for 8% in Norway, which is twice as high a rate as in any other countries in the study.

Similarly Stolk et al (2006)<sup>73</sup> also studied variation in antihypertensive drug utilization and guideline preferences between six European countries (Denmark, Finland, Germany, Norway, Sweden, the Netherlands). They compared the utilisation per class of hypertensive drugs in each country by class. They also analysed guideline preferences in relation to actual use. Classified Per class, relative standard deviations (RSD) across countries were computed. Hypertension guidelines were requested from national medical associations. They found that antihypertensive use patterns varied widely across the countries in absolute and relative terms. They found that total antihypertensive utilization varied considerably, ranging from 152.4 (Netherlands) to 246.9 (Germany) – Defined Daily Dose<sup>4</sup> (DDD)/1000 persons/day.

---

<sup>4</sup> The WHO's definition is: "The DDD is the assumed average maintenance dose per day for a drug used for its main indication in adults."

Relative Standard Deviation was highest for Thiazide Diuretics (TD) (106.2%) and alpha blockers (AB) (93.6%). Where guidelines advocated TDs (Norway and Netherlands), TD utilization was below (Norway) or just above (Netherlands) median TD use. In addition, they concluded that the guidelines seemed disconnected from clinical practice in some countries, and none of the guidelines discussed current utilization.

Lawson and Jick (1976)<sup>74</sup> compared prescribing habits for patients hospitalized in medical wards of university hospitals in America and Scotland. American patients received nearly twice as many drugs both during and prior to hospitalisation than did comparable Scots. These differences persisted for both more specific (e.g. diabetes, hypertension, anaemia) and symptomatic (e.g. anxiety, pain, dehydration) therapies, and this was despite the comparatively higher cost to US patients in terms of adverse events and financial costs.

Differences also appear at local level, and over time. For example, Handelsman (2004)<sup>75</sup> found that testosterone prescribing in Australia over 11 years there were two periods (1993–1994 and 1998–1999) of prominent upsurge followed by declines in the national total prescribing of testosterone. This was seen in spite of a lack of new evidence to justify the surges; the authors postulate this was as a result of promotional activity to prescribe testosterone for older men, rather than overcoming the under-diagnosis of androgen deficiency related to pituitary or testicular disease in younger men. The Australian Pharmaceutical Benefits Scheme (PBS) introduced specific restrictions for androgen prescribing which targeted androgen prescribing for older men without changing standard medical treatment for men with classical androgen deficiency because of underlying testicular or pituitary disease; curtailments in prescribing were partial and temporary, and the authors suggest it may have even encouraged more private (non-PBS) prescriptions for testosterone, despite the significant financial disincentive to patients. The huge driving force of commercial and populist pressure was only partly mitigated by the regulatory barriers, and thus the authors recommended that professional and community education is necessary to improve appropriate diagnosis, and discourage unproven treatments.

Prescribing patterns may also change cyclically over time, for example, annual seasonality affects the types of diseases people develop. McClean et al (2011)<sup>76</sup> investigated antimicrobial prescribing in nursing homes across Europe, using point prevalence studies in April and November. Overall the mean prevalence of antimicrobial prescribing was 6.5% in April and 5.0% in November. The most commonly prescribed antimicrobials were

methenamine, trimethoprim and co-amoxiclav (17.5%, 11.4% and 11.1% respectively) in April, and co-amoxiclav, nitrofurantoin and methenamine (12.2%, 12.2% and 11.5%) in November. There was large variation in overall mean antimicrobial prescribing in the selected nursing homes from each of the contributing countries, ranging from 1.4% in Germany and Latvia to 19.4% in Northern Ireland in April, and 1.2% in Latvia to 13.4% in Finland in November. The findings suggest that there is considerable variation in antimicrobial prescribing in nursing homes across and within European countries.

As already mentioned prescribing differences are shown irrespective of differences in morbidity. International differences in prescribing for chronic heart failure (CHF) have also been shown repeatedly<sup>77</sup>. For example Sturm (2007)<sup>78</sup> noted that there are marked differences across Europe for prescribing therapeutic drugs for (prescribing for ACE-inhibitors ranges from 48-76%, for beta-blockers even lower<sup>79</sup>). Sturm stated that most of the research into differences had focussed on the patient and comorbid conditions, and that the role of the healthcare setting and culture remained less clear; this was the aspect under scrutiny in their study. In a survey to primary care practices from 14 European countries into CHF, the influence of country (factoring out patient characteristics) was assessed using multinomial logistic regression. They found that country of residence clearly influenced prescribed drug volume and choice of drug regimes. Countries determined the number of drugs used and the likelihood of individual drug regimes. There was also much variation of prescribing of guideline-recommended drug regimes ranging from 28.1% in Turkey to 61.8% in Hungary.

In an overview of international differences in antibiotic prescribing Clavenna et al (2011)<sup>80</sup> found 15 studies which compared prescription prevalence and/or prescription rate. A total of eight countries were involved in the studies: Italy, Canada, the USA, The Netherlands, Denmark, the UK, Sweden and Croatia. Canada and Italy had the highest paediatric antibiotic prescription rates, with northern EU countries (The Netherlands and UK) having significantly lower rates. The prevalence in Italy was found to be nearly fourfold higher than in the UK (52% vs 14%, respectively), and the prescription rate was fourfold higher than in Denmark and The Netherlands (1.3 vs 0.3 prescriptions/person/year). More locally, within Italy, the authors carried out a comparison between four different Italian regions over 2005-2007. Prevalence at the regional level ranged from 41% in Lazio to 54% in Umbria. At the level on local health units, of 148 units, the rate ranged from 32% to 60%. The place of residence was

identified as a key determinant of receiving an antibiotic prescription, independent of gender and age.

### **2.3.2.2 Primary healthcare structures, policies and guidelines**

The primary healthcare structure whereby doctors are the coordinators of resources and the gateway to specialists emerged as a mainstay of national health systems after the Second World War, with the implementation of the UK's National Health Service. Many more economically developed countries (with a notable exception being the USA) adapted various elements of this basic structure involving a hierarchy of levels of care: self care, primary, secondary and tertiary. Before its implementation dispensing was a relatively minor source of income; after dispensing became the dominant source of income<sup>81</sup>. This pattern of increasing dominance of dispensing prescriptions has increased over time and has extended to other countries. In contrast, the US does not have comprehensive coverage of medical insurance (this tends to be the domain of insurance companies and employer-based systems) – patients are generally able to self-refer to specialist care and have no requirements to register with primary care clinics<sup>82</sup>. There is an increased emphasis on income and work with dispensing medicines in the US<sup>83</sup>, partly due to not having a national official class of “pharmacy” medicines, and having a less extensive range on non-prescription ingredients than other countries.

Within different healthcare systems, various protocols and guidelines also influence prescribing. There are no international standards for prescribing.

With heart failure, the prescribing guidelines could be a relevant factor. Sturm et al (2005)<sup>84</sup> investigated reasons for major international differences in CHF treatment (as described in section 2.3.2). The authors suggested variation in national guideline recommendations being a relevant factor and thus explored the variation of heart failure 14 national guidelines in Europe, which were compared to heart failure treatment guidelines of the European Society of Cardiology. Relationships between recommendations in prescribing were investigated by comparing national prescribing patterns to the selected guidelines from that country.

The guidelines themselves varied in terms of length, evidence ratings, and the amount of literature included (two countries had no guidelines until 2000). They found that relationships between recommendation and prescribing for selected recommendations was inconsistent amongst countries; thus differences in guideline recommendations were not sufficient to explain variation of prescribing among countries and other factors must be considered.

Reggi et al (2003)<sup>85</sup> documented the variability of prescribing information concerning the indications, side effects and cautions of selected drugs over 26 countries, using the British National Formulary (BNF) as the reference text. Comparison of the different written materials showed substantial disagreement between materials available to prescribers and patients in different countries. There were even significant disagreements within one country when different brand names of the same drug were compared.

The authors explain that the discord in the literature surrounding the drugs analysed was likely to be because evidence availability and weighting is different in different countries, which may in turn have a negative effect of misleading caregivers and patients. The authors suggest using widespread approach involving national regulatory authorities to improve and standardise guidelines both at the national and international level, by further training and education and independent (non commercially funded) research and collaboration and information interchange respectively.

Different weighting of the evidence in relation to actions was cited as a major factor in international differences. For example Australia places a far greater emphasis on the evidence for interventions in skin screening than other countries such as the UK; Helfand et al (2001)<sup>86</sup> reviewed the evidence on stages of cancer found in screening versus usual practice and found conflicting results from ecological studies in Australia and the UK that evaluated the thickness of melanomas after public information campaigns.

Van Duijn et al (2005)<sup>87</sup> explored possible reasons for differences in antibiotic use in outpatients in the Belgium, UK and the Netherlands (with high, moderate and low antibiotic use respectively); factors such as a quality assurance policy programme for GPs on respiratory tract infections, financial dependence on patients, and demographics were explored. Demographics and health care system characteristics were associated with differences in outpatient antibiotic use. Patients' views about respiratory tract infections and antibiotics also were shown to be moderately associated with antibiotic use. It was suggested that being more directly dependent on patients for income increases risk factors for higher levels of outpatient antibiotic use prescribed by GPs, while a higher degree of peer influence might be a possible moderating factor in prescribing<sup>88</sup>.

There are numerous healthcare-structure related reasons for international prescribing differences. The training status, for example, of GPs was found to be the characteristic most

associated with prescribing variation<sup>89</sup>, in terms of rates, classes of antibiotic and performance indicators of antibiotic prescribing.

The Fretheim et al (2005)<sup>72</sup> study suggested international prescribing differences were a product of reimbursement policies, traditions, opinion leaders with conflicts of interests, domestic pharmaceutical production, and clinical practice guidelines. Differences between Norwegian and UK prescribing were explained as “Norwegian physicians are early adopters of new interventions while the British are more conservative; there are many clinical trials conducted in Norway involving many general practitioners; there is higher cost-awareness among physicians in the UK, in part due to fund holding; and there are publicly funded pharmaceutical advisors in the UK.” The authors note that the two most compelling factors which could influence prescribing are the promotion of cheaper drugs by UK pharmaceutical advisors, and promotion of more expensive drug in Norway via “seeding trials”.

Patten et al (2005)<sup>90</sup> carried out a Cochrane review to examine international dosage differences in antidepressant clinical trials. The US maximum and mean dosages of fluoxetine and comparison drugs were almost twice as high as the European prescribed dosages. The authors suggested this was due to the conduct of clinical trials of tolerability and efficacy affecting prescriber behaviour. They also discuss the direction of causation; whether the dosage differences reflect a different style of practice in the US, or contribute to the perpetuation of different practice styles

### **2.3.2.3 Human factors involved**

The differences in healthcare structures can have reciprocal effects on prescriber behaviour (as a feedback loop part of a dynamic system). Physiologically, humans are born similar the world over. Differences in lifestyle can affect disease incidence and prevalence. Do differences in prescribing therefore come from patient side or physician side?

Socio-economic factors as well as medical needs can affect attitudes. According to some international studies, drug prescriptions are influenced by the type of prescriber. For example, inappropriate use of antibiotics was observed among more primary care physicians than family paediatricians in three international studies<sup>91,92,93</sup>. De Las Cuevas et al (2002)<sup>94</sup> assessed intensity and sources of variations between prescribers for antidepressants in Tenerife. Prescribing by GPs mirrored that of psychiatrists; however private doctors (mainly psychiatrists) were found to have a higher use of new and uncommon antidepressants. Psychiatrists acknowledged the pressures of promotion by the pharmaceutical industry and half recognised a personal relationship with some ‘company representatives’. Economic and

social factors were acknowledged to play a major part in this variation, in addition to differences in morbidity.

Attitudes alone do not determine prescribing behaviour e.g. also crucial is the level of perceived need for the drug, but attitude can temper them. In terms of patient and physician attitudes, which may be a determinant of prescribing behaviours, Peyrot et al (2005)<sup>95</sup> found international differences in insulin-prescribing attitudes and behaviours. U.S. physicians were significantly more disposed to delay insulin therapy than physicians in all other countries, except for India and Japan. They also added that factors, such as level of perceived need affect prescribing behaviour. If the need is perceived as greater in the U.S. than in other countries, U.S. physicians might be more likely to prescribe insulin even if they have a higher threshold for making that choice. The level of perceived need might itself be a function of attitudes or it could be a result of actual differences in need, e.g., higher BMI, worse glycemic control, patient unwillingness to change lifestyles, etc.

Physician roles were investigated by Castelo-Branco and Ferrer (2006)<sup>96</sup> in terms of HRT prescribing - they prospectively compared the frequency of prescribing between gynaecologists and GPs. They found that only 10% of gynaecologists and 19.4% of GPs had never prescribed HRT. Reasons for not prescribing were fears of adverse effects and cancer in the GP group, and adverse effects and social alarm in the gynaecologist group. HRT prescribing in Spain is a controversial issue with adverse effects and the fear of cancer negatively influencing people's attitudes, whereas climacteric complaints, quality of life and the prevention of osteoporosis are positive influencers.

Phytoestrogens were most commonly used overall; however, GPs were more willing to use peripheral drugs such as antidepressants and benzodiazepines than gynaecologists. HRT prescriptions were used significantly more frequently for symptomatic women by gynaecologists. The main reasons for prescribing HRT were climacteric complaints and the prevention of osteoporosis for GPs and, climacteric complaints and improvement in life quality for gynaecologists. Seventy-eight percent of gynaecologists prescribing hormones referred a high degree of satisfaction with HRT, whereas only 50% of GPs expressed a similar attitude.

### **2.3.2.4 Conclusion**

There is a high variation in prescribing over a number of levels, which may be irrespective of patient – focussed factors (such as morbidity). To mitigate this factor confounding the results, the study sample will be taken from only UK NHS GPs.

### **2.4 General Practitioner response rates**

GP response rates to surveys are thought to have been falling for many years<sup>97</sup>, and rates are known to vary greatly, with many mediating factors, including volume of requests, questionnaire length, insufficient background information, the perceived value and salience of the research, and financial incentive<sup>98</sup>.

The estimated rate of 5-8% is low compared to some online studies which specifically looked at GP response rates to studies. For example, Bonevski (2011)<sup>99</sup> trialled recruitment strategies for Australian GP participants, involving the use of a general practice authority (local division of general practice) endorsement cover letter and consequent telephone follow up calls of non-responders. They took a subsample of 1666 GPs from the Australasian Medical Publishing Company (AMPCo) database and sent all study materials, alongside incentives of a teabag (to “take a break from their busy day”) and a chance to win a \$500 holiday voucher. GPs were asked to fill in a 15 minute survey into vitamin D. The response rates obtained in a trial of standard research group letterhead invitations (25.8%) versus general practice cover letter (32.5%) were not statistically significantly different; nor were the response rates obtained in the trial of a telephone reminder call. However they achieved an overall response rate of 30.3% (500 respondents). When asked about their preferred mode of survey administration 81.1% of respondents nominated mailed survey; 17.1% stated online survey; 1.7% nominated face-to-face survey; and telephone survey was the least preferred method (0.2%).

Bonevski et al (2010) stated that it was difficult to find GP research studies with good response rates (defined as 70% or above). They quoted an ongoing Australian benchmark study of general practice, which repeatedly obtained response rates of under 30%<sup>100</sup>. Obviously low response rates may cast implications on the generalisability of a study, however it does not follow that research with lower responses is not valid, if taken with caveats<sup>101</sup>.

Morris et al (2000)<sup>102</sup> sent a 10 minute postal questionnaire about minor ailment consultations to 759 GPs in 8 English Health Authorities (HAs). The average response rate was 54.5%, ranging from 35% to 72%, with the lowest rates found in London HAs. The most important stated factors in returning the questionnaires being time spent, originating institution, questionnaire design, sending or reminders, and general tone of request.

## **2.5 Human judgement and decision making and advice taking**

In the psychological JDM literature, as with automation research, the emphasis with respect reliance on advice has been on advice discounting and automation disuse e.g. egocentric bias; a robust finding whereby judges may have biased preference for their own opinions because they believe them to be superior to those of others<sup>103</sup>, or anchoring bias; people start with an implicitly suggested reference point, an "anchor", and make adjustments to it to reach their final decision<sup>104</sup>. Automation disuse has a number of posited causal factors including unsuitable advice, excessive alert frequency, and interruption of prescribing workflow<sup>68</sup>. Overreliance on advice is a much less investigated field which nevertheless has some empirical backing<sup>105,106</sup>.

This research will be informed in part by advice taking literature, where humans are the advisors to help inform the processes which lead to overreliance. However differences have been shown between computerised and human advice. In general, automation is perceived to be more credible than humans, and there appears to exist a 'bias toward automation'<sup>107,108</sup>, which means there is higher initial trust in automation, but conversely can render automation errors more salient to judges. This may lead to a rapid decrease in trust if automation generates errors, leading to a breakdown in dependence. However there is variation in this finding; Ostermann et al (2004)<sup>109</sup> found a small increase in trust with talking head interface (text vs text and speech). Trust in human advisors may be more complex, as it is more tied up and influenced by more conscious and subjective values such as motivation. Hedlund et al (1998)<sup>110</sup> also found that face-to-face interactions differed from those interacting via a computer. Advisors in face-to-face interactions gave more accurate recommendations and gathered more of the task-relevant available information, whereas computer-mediation helped judges effectively weigh the quality of advisor recommendations. The latter finding may be explained by the implication that computer mediation reduces judges' reliance on cues extraneous to the accuracy of recommendations. Other studies found no difference in human versus automation advice dependence; Madhavan and Wiegmann (2005)<sup>111</sup> found that the

belief that an advisor was either human or automated did not globally influence dependence strategies. Despite these potential differences, and bearing them in mind, the general human advice taking literature with respect to JDM can provide explanation on the human processes involved in advice taking. This thesis aims to look at how advice from an automated source is used and will be compared to literature where general advice is over utilised.

For the study to be about AB in DSS, the simulated system must be *presented* to the user as a computerised DSS / advisory system rather than, for example, advice from a remote person, an expert, a librarian doing a search etc. The following section explores types of systems, and thus potential formats for simulation.

Advice taking has been linked to topics of persuasion and attitude change in psychology literature<sup>112</sup>. For example, advice taking increases as the distance between judges' initial opinions and advisor recommendation decreases (Yaniv, 2004b<sup>113</sup>). This effect was particularly strong for more knowledgeable judges. Additionally, Harries, et al (2004)<sup>114</sup> found that judges discounted advisors whose recommendations were very different from those of other advisors (i.e., judges discount outlying advice). Thus it can be inferred that advice (even if incorrect) that is closer to a judge's original opinion is more likely to be accepted. In relation, advice taking is linked to the literature on belief updating. Thus, advice taking could operate according to Hogarth and Einhorn's (1992)<sup>115</sup> belief adjustment model. Over-reliance on the advice could be a function of how the information is encoded (i.e., relative to the pre-advice opinion or a constant), task length (the number of pieces of advice received), the complexity of the information, and how the information is processed (i.e. advice is given in a step-by-step, or they can process only after all advice has been received).

Dual process theories posit that judgment and thus decisions can be based on either logical Bayesian processing, or faster, rule-of-thumb (heuristic) based processing. For example by using a "confidence heuristic"<sup>116</sup> advisors use their own confidence levels to infer their ability, expertise, task-related knowledge, or accuracy on a given task.

Many similar heuristics exist, many are internal and self generated; automation bias may be a case of using automation as the heuristic source for the accurate answer particularly if people are less confident of their own opinion (counteracts variance of trust in self or trust in human sources).

Mosier and Skitka (1996)<sup>117</sup> proposed that insufficient cognitive processing was the cause of over reliance on automation. Rather than carry out more effortful logical processing of information, people often use effort-saving strategies called heuristics. They also coined the term automation bias to refer to “the tendency to use automated cues as a heuristic replacement for vigilant information seeking and processing”.

### **2.5.1 Heuristic use**

Decision-making involves cognitive processes of selecting a course of action from amongst multiple alternatives. Doing so rationally in an everyday, real-time context, people face constraints of incomplete information, and time and processing capacity limitations. Thus people often have to make inductive inferences about unknown aspects of our environment. An increasingly explored counter has been the notion of bounded rationality, which basically expounds previous views by incorporating notions of limited search and stopping rules. Two main related theories have been derived from this: a model proposing satisficing heuristics for searching through a sequence of available alternatives<sup>118</sup>, which has been followed by ‘fast and frugal’ heuristics<sup>119</sup> which use little information and computation to make decisions. Heuristics are cognitive rules of thumb that help simplify decisions. Satisficing involves the user settling for a satisfactory solution rather than the best; the user sets up a goal and searches through alternatives until one is found that reaches this level. This contrasts with the more traditional models of decision-making which were based on assumptions of unbounded rationality, whereby humans were intrinsically rational beings who adhered to normative (behaving as predicted by the rules of logic, statistics and probability) and descriptive ideals of how humans should reason. For example, people should not be influenced by ‘mood, context or mode of presentation’ (Shafir and LeBoeuf, 2002<sup>120</sup>). According to this theory, standard statistical processes are the tools by which inference and decision-making are carried out; multiple regression for example, is used to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable. Hammond (1990)<sup>121</sup> used this as a model of inductive inference in multiple-cue learning, and Bayes’s theorem (a result in probability theory) is a model of human reasoning and memory<sup>122</sup>. Intuitively, ecologically and empirically there is evidence for various effects to the contrary. Examples include the mere exposure effect, which occurs when repeated exposure to a stimulus increases the positive affect associated with it<sup>123</sup> (exploited in advertising) and asymmetric dominance (which violates logical assumptions of invariance).

Aside from limitations in human processing, rationality may also be bounded by individual differences such as age, experience, knowledge and attitude.

A number of studies have shown increasing environmental pressure increases satisficing and heuristic use, and non-compensatory decision strategies<sup>124</sup>. Rather than evaluate *all* options of a choice set on all the appropriately weighted criteria, people tend to rely on strategies where a high score on one criterion cannot compensate for a low score on another. Usual reasons for this include time pressure, the ability to achieve optimisation and sometimes the recognition that optimising the decision making strategy to find the best solution may not be worth the marginal cost to obtain it (this is also dependent on the risk involved in making suboptimal decisions). Also the presence of too many alternatives, so called “tyranny of choice”, may attenuate decision making strategies by providing an overload of information exceeding the cognitive and environmental resources<sup>125</sup>. When making complex decisions, such as those with multiple alternatives, or when under time pressure, people simplify tasks and ignore a lot of information. One sort of simplification is switching from compensatory (comparing and weighing multiple cues) to non-compensatory decision strategies. Non-compensatory decision categories rely on the use of heuristics (i.e. fewer cues and fewer resources used to judge and decide).

Agosto (2002)<sup>126</sup> investigated bounded rationality and satisficing in “young people’s” web based decision making in terms of how time constraints, information overload and personal preferences affected the satisficing behaviour. Major satisficing behaviours found when faced with these limitations were reduction (filtering out information) and termination (early search stopping). It is reasonable to assume that increasing environmental pressure would also increase use on external decision aids as people attempt to compensate for an increase in environmental demand putting pressure of cognitive resources. It has been suggested that proficient decision makers can be people that rarely make decisions in an overly analytical or rational way (i.e. by compensatory decision strategies)<sup>127</sup>, instead choosing a course of action using knowledge and experience (i.e. “know” what to do, rather than figuring out what to do), making decisions quickly and largely automatically<sup>128</sup>.

While it can be argued that heuristics are an adaptive method of saving cognitive and time resources and are useful when advice is reliable, their use may also lead to systematic biases which may lead to error. Graber<sup>129</sup> found that cognitive factors contributed to diagnostic error

in 74% of cases. The most common cognitive problems involved faulty synthesis. Premature closure, i.e. termination, was the single most common cause. Reducing the biases should decrease inappropriate automation use. Beck et al (2002)<sup>130</sup> found that use could be better calibrated by providing participants multiple forms of feedback of the aid's performance.

### **2.5.2 Methods of measuring reliance**

Over the literature review a diverse set of methods for measuring reliability on automation were found in addition to few researchers providing strict definitions of the appropriateness of reliance. This renders any meta analysis of experimental findings more difficult. Wang et al (2008)<sup>131</sup> carried out a literature search and organised methods found into four main perspectives; automated task performance, user consistency with automation, behavioural indicators (e.g. cross verification, attention allocation) and response bias (appropriateness of reliance on binary 'signal' and 'noise' feedback). The reliance measures most pertinent to this study are those of task performance and consistency indicator.

#### Performance indicator:

This uses the difference between performance (in this case accuracy, or error rate) of the automated task when receiving correct feedback and when receiving incorrect feedback. This will be the primary method of reliance measurement used, as the study aims to look at responses to incorrect advice.

#### Consistency indicators:

1. Reliance is indicated by the percentage of opportunities that the users follow automation feedback
2. The correlation between users' decision and automation feedback

For this research, the consistency indicators are more direct measures of AB i.e. using negative consultations as the primary outcome measure of AB. Performance measures will be used as more inferential methods of testing for experimental effects.

### **2.6 Factors which affect automation reliance and influence Automation Bias**

A broad literature search was carried out to explore hypothetical and tested factors which influence AB. The search criteria were kept fairly loose to incorporate different types of

human-automation interaction with the assumption that this could provide illumination to the more specific area of decision support and human over-reliance. Factors which may affect the calibration or reliance on DSS may be categorised into 4 broad areas: That of the context (organisation, culture), the task/event, the user, and the DSS/automation itself.

### **2.6.1 Context**

The context of DSS use is possibly the most diffuse and difficult area to quantify as an area of potential causes for automation bias because, intuitively, effects are widespread and indirect. Context comprises cultural, organisational and environmental factors.

The culture of an organisation can affect reliance on automation<sup>132</sup>; with culture defined as a set of social norms and expectations that reflect shared learning and life experiences. Cultural differences associated with power distance (e.g., dependence on authority), uncertainty avoidance, and individualist and collectivist attitudes can influence the development and role of trust<sup>133</sup>.

Environmental factors such as heating, lighting, ventilation (and extraneous distractions) can add to environmental stress.

Organisational factors involve indirect issues such as job design, politics and the organisation of work; it also encompasses more direct training and support factors, in this case more specifically for implementation of new technology.

More quantifiable is the impact of training and support<sup>134</sup> on automation usage and accuracy. For example, Masalonis (2003)<sup>135</sup> described how training enhanced the appropriateness of trust in the context of situation-specific reliability of decision aids for air traffic controllers. More recently, Bahner(2008)<sup>136</sup> investigated how exposing participants to rare automation failures as part of a preventative training intervention. It was found that training reduced complacent behaviours (in this case, a lack of cross verification of advice); conclusions from this were that exposure to automation failures could sensitise users as to when advice given was incorrect.

### **2.6.2 Task / event**

The type of task / event that the DSS/automation is intended to support comprises issues such as the number of different variables the user has to attend to or monitor, and task-related environmental constraints such as task complexity, time pressure, risks involved.

In an aviation study, Mosier et al (2007)<sup>137</sup> examined the impact of operational variables on diagnosis and decision-making processes, focusing on information search. Time pressure, a common operational variable, had a strong negative effect on information search and diagnosis accuracy, and the presence of incongruent information heightened these negative effects. Skitka et al (2000a)<sup>35</sup> found that time pressure decreased cross verification behaviours. Diagnosis confidence was unrelated to accuracy and was negatively related to amount of information accessed. This effect was also similar to one found in a study on decision reliance with auditors and an auditing decision aid – there was a positive correlation between time pressure and decision aid reliance<sup>138</sup>. The psychological research shows that humans tend to change decision strategies based on the amount of information demands<sup>139</sup> and time available<sup>140</sup>.

Gomaa (2008)<sup>141</sup> found evidence that risk (in this case, litigation risk) increased decision aid reliance - based on an experiment involving 118 audit practitioners, it was found that auditors relied more on decision aid advice when either litigation risk or internal control risk was high. When both risks were simultaneously high, the litigation risk was found to amplify awareness of legal defensibility, which increased decision aid reliance, even as confidence in the quality of their judgements deteriorated. Numerous studies (much of them from the healthcare field) have shown a mismatch between actual and perceived situational risk<sup>142,143</sup> (this can be mediated by personal factors such as self efficacy<sup>144</sup>) and perceived risk has been shown to impact the person's risk taking behaviour<sup>145</sup>.

Task difficulty has been found to increase reliance on decision aids<sup>138,146</sup>; as task difficulty increases to reach the user's cognitive capacity, aid from external resources is increasingly, and potentially erroneously relied on. Information load has been shown to affect decision performance by stimulating the decision-maker to alter decision making strategies (Newell and Simon, 1972)<sup>147</sup>. Jacoby et al. (1974)<sup>148</sup> demonstrated that decision-makers tended to make worse decisions, yet were more satisfied, and more confident about their decisions, as the information load increased. Chinburapa et al (1993)<sup>149</sup> found that increasing task complexity caused physicians to shift from using compensatory to noncompensatory decision-making processes (i.e. more effortful processing, to more heuristic type processing). Task load has been shown to interact with trust to influence automation reliance<sup>150</sup>, a positive relationship between automation trust and automation use exists and there is a suggestion that task load has a negative effect on the positive relationship between automation trust and

automation use. Participants with a higher task load exhibited over-reliance on their automated information systems to assist them in their decision-making. The researchers suggested that “such an over-reliance can lead to vulnerabilities of deception and suggests the need for automated deception detection capabilities”<sup>151</sup>. Berner (1999)<sup>152</sup> - Physicians' diagnostic performance was significantly better ( $p < 0.01$ ) on the easier cases and the cases for which Quick Medical Reference (QMR) could provide higher-quality information. Physicians' diagnostic performance can be strongly influenced by the quality of information the system produces and the type of cases on which the system is used.

Repetitive tasks have been shown to elicit habituation effects which may increase the risk of automation bias errors by the reduction of vigilance and error detection<sup>153</sup>. This may be particularly relevant in monitoring situations.

Different situations require different types of decision support in terms of their level of activeness e.g. monitoring versus diagnostic situations. The types of DSS used fit the nature of the situation and desired outcome and may involve qualitatively and quantitatively different processes.

### **2.6.3 User**

The socio-technical system formed between a decision aid and the user is crucially dependent on the human factors involved. There are a number of theories which postulate how human factors are factors in automation reliance. Riley (1989)<sup>154</sup> suggested that reliance, trust and confidence act as the primary cognitive mediators for human-computer interaction. In their Framework for Automation Use, Dzindolet et al (2010)<sup>155</sup> group human factors involved in reliance into social, motivational and cognitive factors. The Theory of Technology Dominance (TTD)<sup>156</sup> was developed in the late 90s and posits that DSS and task experience, and task complexity and cognitive fit are important factors when investigating reliance on intelligent decision aids. This section of the review aims to look at numerous different user-centred factors cited in the literature.

Individual differences have been shown to produce differences in reliance on automation. Ho et al (2005)<sup>157</sup> found that when using a medicine management system older users were more likely to trust in the aid and were less confident in their performance, but they did not calibrate trust differently than younger adults. It was also found that older adults were more

reliant on the decision aid and committed more automation related errors through over reliance. A signal detection analysis indicated that older adults were less sensitive to automation failures. Riley (1994a)<sup>158</sup> (cited in Parasuraman, (1997)<sup>22</sup>) also found that individual differences in patterns of automation use were commonplace, particularly between those who cited fatigue as an influence, and those who cited other factors. Singh et al (1993)<sup>159</sup> found that people who tended to have inflated estimates of an automated aid's reliability were more likely to trust and rely (and over rely) on the automation. Probst et al. (2009)<sup>160</sup> studied individual differences related to willingness to use a computer based DSS. They found that physicians generally believed DSS to be beneficial in general and also specialised medicine (of 59 physicians, of various domains). They preferred to use DSS as information systems rather than as tools for diagnosis. Additionally, confidence in one's own diagnostic ability (in slight disagreement with the bulk of confidence and decision-making literature), computer use, Internet use and attitude toward statistics did not play a major role in physicians' willingness to use CDSS.

Huber (1983)<sup>161</sup> stated that cognitive style should be considered in the design of DSS. Chakraborty et al (2008)<sup>162</sup> found that cognitive style had significant direct effects on perceived usefulness, perceived ease of use, and subjective norms of automation. Both perceived usefulness and subjective norms affected actual technology usage significantly. People with innovative cognitive styles are more likely to perceive a new technology as useful and easy to use than are those with adaptive cognitive styles. There may also be a difference in appropriateness of reliance depending on whether the user has a primarily compensatory or non-compensatory style of decision making. Compensatory strategist may be more likely to take account of all the information available, whereas non-compensatory strategist is more likely to adhere fewer cues on which to base their decision. Bergman and Fors (2005)<sup>163</sup> found a correlation between positivity towards DSS and learning style, and also between learning style and computer skill. Results indicated that the use of CDSS did not guarantee correct diagnosis (i.e. use was suboptimal) and that learning style might influence the results. "Cognitive fit" theory<sup>164</sup> proposes that the correspondence between task and information presentation format leads to superior task performance for individual users; a number of studies have shown that tailoring the way information is presented to individual cognitive styles does enhance performance e.g. manipulating information visualisation to support decision making tasks<sup>165,166</sup>. Operationally, cognitive fit is not measured per se, but rather manipulated in experimental studies that employ the construct.

Many studies have highlighted the impact of clinical task experience on appropriateness of reliance – studies generally imply that the more task experience a user has, the less likely they are to rely on automation, with overreliance tending to be more prevalent in less experienced groups<sup>23</sup>. Berner (1999)<sup>152</sup> assessed a group of clinicians working a set of difficult cases and using the QMR DSS, and suggested that the extent of benefit gained by different users varied with their level of experience.

Dreiseitl and Binder (2005)<sup>167</sup> observed that in 24% of the cases in which the physicians' diagnoses did not match those of the decision support system, the physicians changed their diagnoses. There was a slight but significant negative correlation between susceptibility to change and experience level of the physicians. Physicians were significantly less likely to follow the decision system's recommendations when they were confident of their initial diagnoses. In a simulated harvesting task<sup>168</sup>, it was found that domain experience had a major impact on behavioural reliance with those with more experience being less likely to rely on the automation.

Lee et al (2004)<sup>169</sup> performed an empirical investigation into the effect of users' DSS expertise on their problem-solving strategies. The results indicated that individuals who had only recently learned to use the DSS were confused or restricted by the set of functions provided by the system and did not plan well for their use of the DSS. Those who had previous knowledge of the system exhibited more focused and efficient problem-solving behaviour, suggesting that problem-solving strategies depended significantly on the user's level of system expertise.

Immediate experience of DSS use may also affect reliance. The sequencing of errors is not often reported in studies investigating the effect of reliability on automation reliance, despite there being much investigation into this effect e.g., Bliss et al. (1995)<sup>170</sup>; Parasuraman, Molloy & Singh (1993)<sup>21</sup>; Vries, Midden & Bouwhuis (2003)<sup>171</sup>; Wiegmann et al. (2002)<sup>172</sup>, amongst others. In general, studies tend to report overall levels of reliability; the sequence at which the errors arise is rarely reported or even mentioned.

However, there is evidence that the placing of errors over time can effect reliance and overall reported trust at the end of a session. Wickens and Xu (2002)<sup>173</sup> suggest that humans interacting with an initially reliable system would have a different perception of the first automation error than humans interacting a system that is less reliable later on in time. It is argued that the first automation failure can result in a more pronounced drop of trust and reliance on the automation than subsequent failures.

There is mixed evidence for the existence of a “first failure effect”. Molloy and Parasuraman (1996)<sup>174</sup> for example found that reliance on monitoring of the automated engine failure monitoring system dropped on detection of error in a flight simulation task. However, Wickens, Helleberg and Xu (2002)<sup>175</sup> found that reliance did not drop significantly as a result of first error.

Overall, evidence suggests, the distribution of errors, therefore, may be a component in reliance. Sanchez (2006)<sup>176</sup>, for example, found that participants who were exposed to automation error at the beginning or end of a series of cases relied more on automation than participants who were consistently shown error. This, it was suggested, implied that when automation frequently and randomly generates errors, humans’ reliance is more likely to remain lower than if the automation behaves reliably for an extended period of time.

In a study of how accountability affected errors, Skitka et al (2000)<sup>35</sup> explored the extent to which omission errors can be reduced under conditions of social accountability. Results indicated that making participants accountable for either their overall performance or their decision accuracy led to lower rates of automation bias. Omission errors were found to be the result of cognitive vigilance decrements, whereas commission errors proved to be the result of a combination of a failure to take into account information and a belief in the superior judgement of automated aids. This corroborated earlier findings by Mosier et al (1996)<sup>177</sup> who found that participants who perceived themselves “accountable” for their strategies of interaction with the automation were significantly more likely to verify its correct functioning, and committed significantly fewer automation-related errors than those who did not report this perception.

Trust in automation has perhaps been one of the topics with the most investigation in terms of investigating properly calibrated reliance, with the assumption that the higher the level of trust placed in automation, the more the user is likely to rely on it<sup>171</sup>. If too much trust is placed on an unreliable system, automation bias may occur. It has been found to affect reliance in many domains such as car navigation systems<sup>178</sup> and aviation automation<sup>179</sup>. Trust in automation is often calibrated according to the user’s perception of advisor competence. Muir (1994)<sup>180</sup> found that trust was significantly reduced by any sign of incompetence in the automation, even one which had no effect on overall system performance. Generally, operators’ trust altered very little with experience. Distrust in one function of an automatic component spread to reduce trust in another function of the same

component, but not to other components in the same system, or to other systems. There was a high positive correlation between operators' trust in and use of the automation; operators used automation they trusted and rejected distrusted automation, preferring manual control. There was an inverse relationship between trust and monitoring of the automation.

Somewhat in contrast to above, results from a different study implied that trust does vary over time, and also with the type of malfunction<sup>181</sup>, with the dynamics of trust in and use of automation depending on the occurrence patterns of malfunctions. If continuous malfunctions occur, operator trust reduces significantly, and eventually the operator does not rely on the automation, even under circumstances that are easy for the automation to handle. The longer the continuity, the longer this effect lasts. In contrast, discrete malfunctioning is found to not cause a significant decline in the operator's level of trust.

Trust has generally been tested as specific to the DSS being used in a study. It may be that there is a dissociation between specific general levels of trust in automation. General levels may be more influenced by individual differences, for example whether someone is more technophobic generally, and cognitive style.

Trust is also shown to affect, and be a product of a number of other factors such as complacency, situational awareness and mental workload<sup>182</sup> amongst other factors.

A number of variables have been investigated empirically as mediators in the role of trust on reliance, such as intrusiveness of the automation advice<sup>170</sup> (advice is more trusted and utilised when it is asked for, as in the JDM literature<sup>183</sup>), decision aid reliability<sup>184</sup> and user awareness of reliability / errors<sup>185</sup>. Mood also may affect decision making, with acute positive affect having been found to improve the decision analysis process (it was argued by making physicians more compensatory in their judgement<sup>186</sup>).

Working parallel to trust in automation is the confidence the user has in their own task-related abilities and is linked with “egocentric bias”.

The literature shows that less confident judges seek greater amounts of advice<sup>187</sup>. Post-advice confidence is sensitive to increasing advisor accuracy<sup>188</sup>. Confidence levels are also higher when there is a greater amount of information on which advisors can base their recommendations and when judges receive recommendations from numerous advisors<sup>189</sup>. Some researchers have also found that judges can be overconfident in their own judgment and decisions<sup>190</sup>. Overconfidence literature indicates that the prevalence of overconfidence depends on the type of task used, with overconfidence being more likely in judgment than in

choice tasks<sup>191</sup>. Westbrook (2005)<sup>192</sup> found that the information obtained from an online evidence system influenced clinicians' confidence in their answers to the clinical scenarios, and that many clinicians placed confidence in information that led them to incorrect answers.

Lee and Moray (1992)<sup>193</sup> identified self confidence as an important factor a trade off with trust in the automation, when trust in the automation exceeded self-confidence, the automation was more likely to be used and over-relied on. Trust and confidence have emerged as the critical factors in investigations into human-automation mismatches in the context of machining<sup>194</sup>.

In relation, self efficacy is often defined in the psychological literature as being a situation specific example of confidence<sup>195,196</sup>.

Mental fatigue has been shown to decrease attention and reaction times and also increase both commission and omission errors in tasks<sup>197</sup>, and increases with time spent on the tasks. There is also a potential dissociation in the effects of mental fatigue on goal-directed (top-down) and stimulus-driven (bottom-up) attention: mental fatigue results in a reduction in goal-directed attention, leaving subjects performing in a more stimulus-driven fashion. Decision fatigue is a fairly new term which is posited to be the result of pressure on a finite store of mental energy for exerting self-control. Once mentally depleted, people become reluctant to make trade-offs, which involve a particularly advanced and taxing form of decision making i.e. shifts occur from compensatory to non compensatory decision strategies<sup>198</sup>. Research shows the decisions judges make are highly influenced by the length of time since a work break: "the percentage of favorable rulings drops gradually from ≈65% to nearly zero within each decision session and returns abruptly to ≈65% after a break."<sup>199</sup>

Wu et al. (2008)<sup>200</sup> tested an extended technology acceptance model (original by Ventakesh and Davis, 2000<sup>201</sup>), finding that management support was positively linked perceived usefulness, ease of use and subjective norm (the degree to which an individual believes that people think she/he should use the system). Subjective norm was positively related to trust, and perceived ease of use is positively related to perceived usefulness. All four factors of perceived usefulness, perceived ease of use, subjective norm and trust were positively linked to technology acceptance, trust having the strongest relationship.

It is likely that these factors, as well as accepting technology in general, will be linked to the extent to which advice from technology is utilised.

Studies have shown dissociation (but strong positive relationship) between intention to use technology and behavioural use of technology<sup>22,201</sup>, and dissociation between stated and observed usage (Yeh and Wickens, 1988<sup>202</sup>).

The concept of user “complacency” has not been well operationally defined in the literature, but appears to be linked to behaviours of deficient cross-verification<sup>117</sup>, unawareness of dangers of failures, lower effort to engage<sup>22</sup>, or trade-offs in high workload situations (i.e. conservation of cognitive resources)<sup>203,204</sup>. Complacency can include a loss of situational awareness<sup>205,206</sup>, and a higher risk that automation failure detection will be delayed<sup>21</sup>.

Monitoring behaviours are also related to the user’s perceived reliability of the automation<sup>207</sup>. Singh (1993)<sup>28</sup> suggested that though the potential for complacency is an independent factor affecting overreliance, it does relate to components of trust and confidence. Bahner et al (2008)<sup>136</sup> posited that complacency is reflected in inappropriate checking and monitoring of automated functions. It was found that lack of verification behaviour was associated with more automation bias, but was mitigated by training.

The cognitive overhead involved may influence the use of automation. This comprises the ease of use of the system and the required effort to engage<sup>208</sup>. The more difficult to use and the more effort required to engage in automation is related to higher disuse.

#### **2.6.4 Decision Support System characteristics e.g. Interface**

The characteristics of the DSS have been found to affect user reliance and performance. Madhavan and Wiegmann (2007)<sup>209</sup> hypothesised that the visible behaviour of a decision aid affects its perceived reliability. They stated that the salience of the advice, the ease of the task at hand, and the types of errors (whether it was omission or commission) would affect this perception. Madhavan et al (2007)<sup>210</sup> found that trust was degraded more quickly when the decision aid made errors on easier tasks.

Studies suggest that people tend to apply social norm factors (as they would in human-human interactions) to human-automation interactions. Factors such as source and authority<sup>209,211</sup>, affect the perceived credibility of a system. System authority or expertise in particular, affects how people assess its reliability and utilise its advice. Automation bias may exist due to people’s tendency to assume automated aids to be experts. Madhavan and Wiegmann

(2006)<sup>211</sup> provided information to users about the “expert” or “novice” status of the aid, and the aids in both groups were rated as more reliable than the human source in both cases, more so the “expert” system. Participants agreed more with an automated ‘novice’ than a human ‘novice’ suggesting a bias toward automation.

The location of advice with respect to the non-automated information or raw data may also influence people’s vulnerability to automation bias<sup>212,213</sup>. In a series of laboratory studies, Jamieson et al (2007)<sup>214</sup> examined the effects of system reliability information and interface features on human trust in, and reliance on, individual combat identification systems. It was found that providing reliability information led to more appropriate reliance on that feedback. It was also found that the method of displaying reliability information affected the participants’ sensitivity - the display format (integrated vs. separated) affected the participants’ reliance on the system. When reliability information was integrated with the feedback, and thus easily accessible, participants relied on the aid more appropriately. When the feedback reliability information was integrated with the feedback itself the participant could more easily access the information while determining the results of the inquiry feedback.

The display format and content of the advice given can affect the use of automation. Because it is unlikely that users will be able to see and understand the workings of the automation, the perception of the information it gives may be mediated by how it is displayed.

Appropriateness of trust and reliance may depend somewhat on the content and format of the display. Internet based interactions have been the focus of much of this research (e.g. technology credibility research<sup>215</sup>). In many of these studies, the perceived credibility depends on superficial features and is not directly linked to the system’s true capability. Visual design factors of the interface e.g. colours and a balanced layout, can also induce trust<sup>216</sup>. Karvonen and Parkkinen (2001)<sup>217</sup> found that trusted websites tended to be text based, use empty space as a structural element and have strictly structured grouping. Trust and reliance increased when information was displayed in a way that provided clear, deliberate and concrete details that were consistent and clearly organised.

A speech interface study showed that people were more trusting of a system that used synthetic speech consistently, as compared with one that used a combination of synthetic and human speech<sup>218</sup>. Consistent presentation style is likely to lead to a greater perceived

reliability of a system. Berner et al (2003)<sup>219</sup> found that people tended to be mostly swayed by the most prominent advised diagnoses. It was found that physicians were strongly anchored by their initial diagnoses prior to using the CDSS. This corroborated Teich et al (2000)<sup>220</sup> who found that physicians were more receptive to advice that did not require a change in initial plans, but that changes in diagnoses after using the CDSS related to presence or absence of the correct diagnosis in the top 10 diagnoses displayed by the CDSS.

Models in the persuasion literature may inform the way information is presented and its credibility, as already outlined, Hogarth and Einhorn's<sup>115</sup> belief adjustment model suggests over-reliance could be a function of how the information is encoded, task length, the complexity of the information, and how the information is processed.

A key method of categorising the degree of input from the user or the degree of autonomy of the machine is the Levels of Automation categorisation. Various levels of automation can be introduced in decision support systems, from fully automated where the operator is completely left out of the decision process to minimal levels of automation where the automation only presents the relevant data<sup>10,221</sup>. It has been suggested that close to fully automated systems can induce a state of "complacency" by taking too much control away from the user along with a sense of accountability. Adaptive automation has been proposed as a method of avoiding this, whereby the user or system can initiate changes in the level of automation<sup>222</sup>. Recently, systems have been developed that follow the "neuroergonomics" approach and even use psychophysiological measures to trigger changes in the state of automation<sup>223</sup>.

The degree of success of a CDSS may inform as to which factors affect appropriate reliance on automation. Kawamoto et al (2005)<sup>20</sup> carried out a systematic review into features which led to DSS success; four features were found to be independent predictors of improved clinical practice: automatic provision of decision support as part of clinician workflow, provision of recommendations rather than just assessments, provision of decision support at the time and location of decision making, and computer based decision support as opposed to manual. In addition to this a system providing the user with a rationale as to why the system might err increases trust and reliance in a system - after observing the automated aid make errors, participants distrusted even reliable aids. This was only mitigated if an explanation

was provided as to why the aid might err. Knowing why the aid might err increased trust in the decision aid and increased reliance, even when the trust was unwarranted<sup>185</sup>.

Other differences which may affect the way a case is processed is the mode of the advice. There is evidence that response mode can affect the decision making process. For instance Billings and Scherer<sup>224</sup> found that choice tasks gave rise to more non-compensatory decision strategies than judgment tasks.

Psychological concepts within the psychophysics domain may have a bearing on the salience of the advice and thus the amount of attention it receives. For example the Weber-Fechner effect<sup>225</sup> (or similarly Stevens' power law<sup>226</sup>), which proposes that the magnitude of a physical stimulus is positively related to its perceived intensity. This theory is corroborated by the Berner et al. (2003)<sup>219</sup> findings outlined above. This implies that the user's response is also in proportion to the magnitude of the stimulus. It may be that increasing the advice salience, DSS advice usage will be affected.

### **2.6.5 Outcomes**

Though not directly related to the factors being looked at in this study (i.e. those precipitating AB), outcomes could act as a form of feedback into the DSS-user system to affect reliance. Garg (2005)<sup>6</sup> carried out a review of controlled trials to assess the effect of CDSS on practitioner performance and patient outcomes. These effects will feed back into the healthcare institution.

### **2.7 Strength of relationships**

The literature review suggests a number of factors which may be involved in tempering reliance on automated decision support. The strongest evidence (in terms of number and quality of studies, and the directness of the relationship) does revolve around certain concepts. Attitudinal concepts such as trust and self-confidence have shown strong positive relationships between automation misuse and disuse respectively. In terms of non attitudinal user characteristics, experience, both in terms of task-related and DSS experience have a strong evidence base for relationships to reliance and tendency to misuse. This corroborates Azen and Budescu findings<sup>34</sup>. Task factors such as task difficulty/complexity and environmental constraints such as time pressure can increase satisficing behaviour and

heuristic use leading to automation bias. Finally the cognitive “fit” between the DSS and the user shown to have strong underlying effects on judgement and decision making behaviour.

In addition to more direct relationships, there are also many interrelationships between different factors which may not directly impact reliance but are part of the system of influence. For example, physician accuracy is found to be related negatively to task difficulty (this is likely to be mediated by factors such as task experience)<sup>227</sup>. Fatigue and attention have also shown to be negatively related, Boksem et al (2005)<sup>197</sup>. These and many other relationships may indirectly affect reliance behaviour, or have mediating effects.

The next section attempts to use pre existing theories and literature to create a conceptual model of overreliance, and resulting from that a pilot ontology and testable model for AB.

## 2.8 Previous models of reliance

The literature review found a number of models (mostly theoretical) for *reliance*, general user *acceptance* or *intention to use*. Examples of these are outlined below.

### 2.8.1 Empirically tested models

Ventakesh et al. (2003)<sup>228</sup> reviewed and compared the user acceptance literature to generate an overall psychological meta-model of reliance. The unified model was empirically validated (using questionnaires with statements and Likert scales to validate the model) using hierarchical regression techniques.

This model looked at direct effects of Performance Expectancy, Effort Expectancy, Social Influence on Behavioural Intention, which was mediated by Gender, Age, Experience, and Voluntariness of Use.

The results of which are shown in table 2.2 below:

**Table 2.2** Table to show results of the Unified Model of Technology Acceptance (taken from Ventakesh et al. 2003)

Dependent Variables	Independent Variables	Moderators	Explanation
Behavioral Intention	Performance expectancy	Gender, Age	Effect stronger for men and younger workers
Behavioral Intention	Effort expectancy	Gender, Age, Experience	Effect stronger for women, older workers, and those with limited experience
Behavioral Intention	Social influence	Gender, Age, Voluntariness, Experience	Effect stronger for women, older workers, under conditions of mandatory use, and with limited experience
Behavioral Intention	Facilitating conditions	None	Nonsignificant due to the effect being captured by effort expectancy
Usage	Facilitating conditions	Age, Experience	Effect stronger for older workers with increasing experience
Behavioral Intention	Computer self-efficacy	None	Nonsignificant due to the effect being captured by effort expectancy
Behavioral Intention	Computer anxiety	None	Nonsignificant due to the effect being captured by effort expectancy
Behavioral Intention	Attitude toward using tech.	None	Nonsignificant to the effect being captured by process expectancy and effort expectancy
Usage	Behavioral intention	None	Direct effect

· Workman (2005)<sup>229</sup> used theory of planned behaviour<sup>5</sup> to formulate hypotheses about the use, disuse, and misuse of an expert system decision support. It was found that DSS use was negatively related to errors, whereas misuse of DSS was positively related to errors. More positive attitudes and social influences led to increased DSS use, while perceptions of control had no apparent effect. The interaction of social influences and attitudes had a significant non-linear relationship with DSS misuse.

· Wu et al. (2008)<sup>200</sup> looked more specifically at healthcare, and professionals' intention to use an adverse event reporting system. The tested model is shown below (fig 2.1). The results indicated that perceived usefulness, perceived ease of use, subjective norm, and trust had a significant effect on a professional's intention to use.

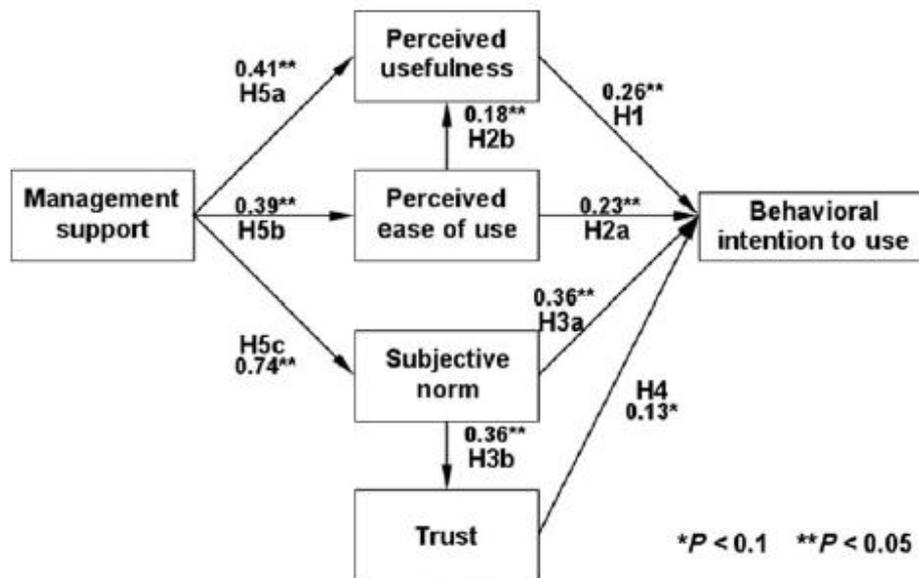


Figure 2.1 Model of empirical results – strengths of relationships (taken from Wu et al. 2008)

<sup>5</sup> A strong theory from the persuasion literature linking personal attitude, subjective norms, and perceived behavioural control to intentions and behaviour. Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In J. Kuhl & J. Beckmann (Eds.), *Action control: From cognition to behavior*. Berlin, Heidelberg, New York: Springer-Verlag.

## 2.8.2 Theoretical models (generated from literature searches)

· Dzindolet (2011)<sup>230</sup> created a “Framework of Automation Use”, which predicts automation reliance decisions are determined by cognitive, social, and motivational factors, with AB being part of the cognitive processes affecting reliance. The model, derived from literature reviews, is shown below (fig 2.2):

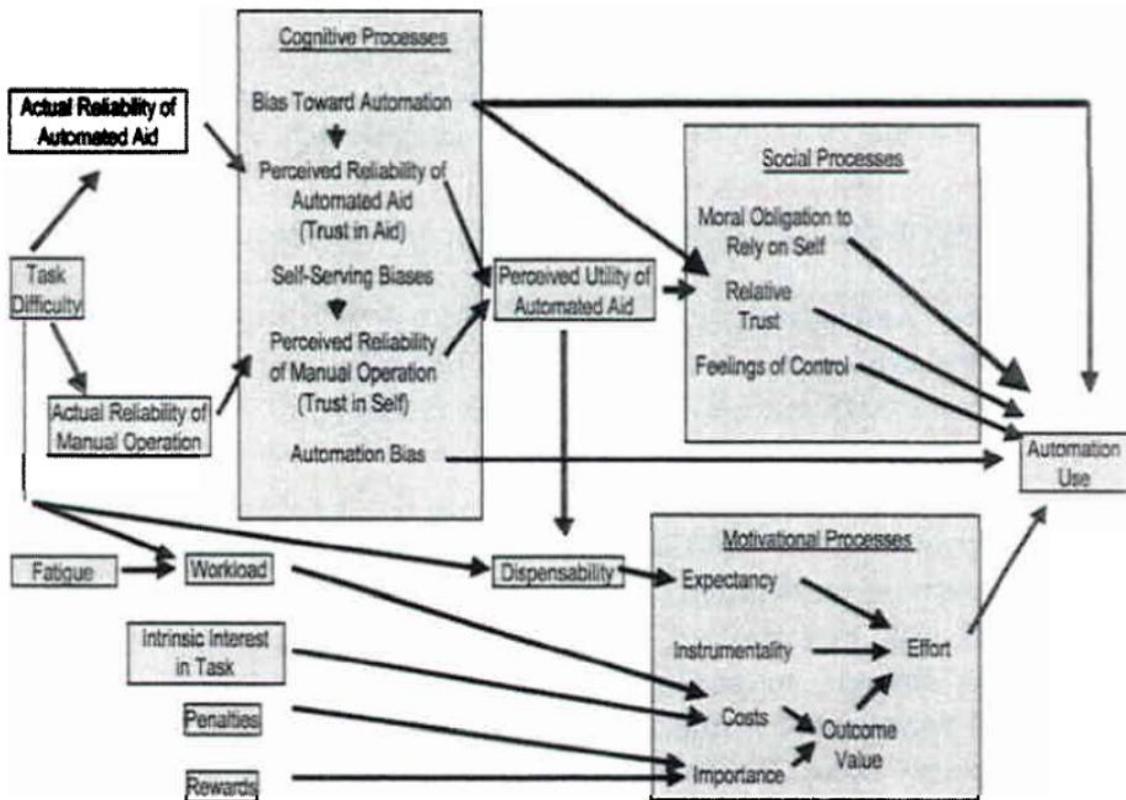


Figure 2.2 Framework for Automation Use (taken from Dzindolet et al. 2011)

· Alberdi et al. (2009)<sup>231</sup> specifically carried out a literature review into AB to assess the causes of omission errors in alerting systems. They cited 15 hierarchical causal factors (including time pressure, self trust, trust in tool, and cognitive overload), and 6 potential triggers of the causal factors (including unexpected tool behaviour, uncertainty/difficulty of the demand, and no other source of information), which lead to increased omission error rate. This model could potentially be transformed into an ontology for AB.

Parasuraman and Riley (1997)<sup>22</sup> generated a model of reliance, see fig 2.3 (solid lines had empirical support, dotted lines are hypothesised relationships), which was primarily derived from their work in the aviation field. Factors including monitoring of automation included workload, automation reliability, and saliency of indicators.

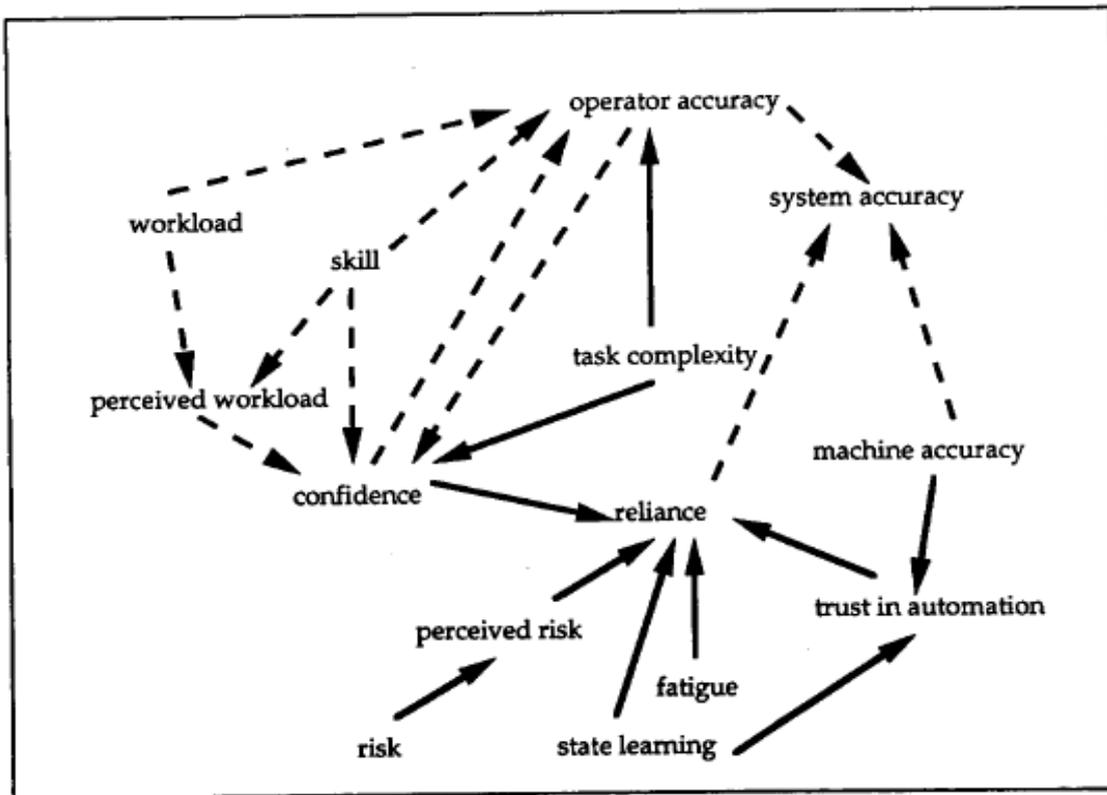
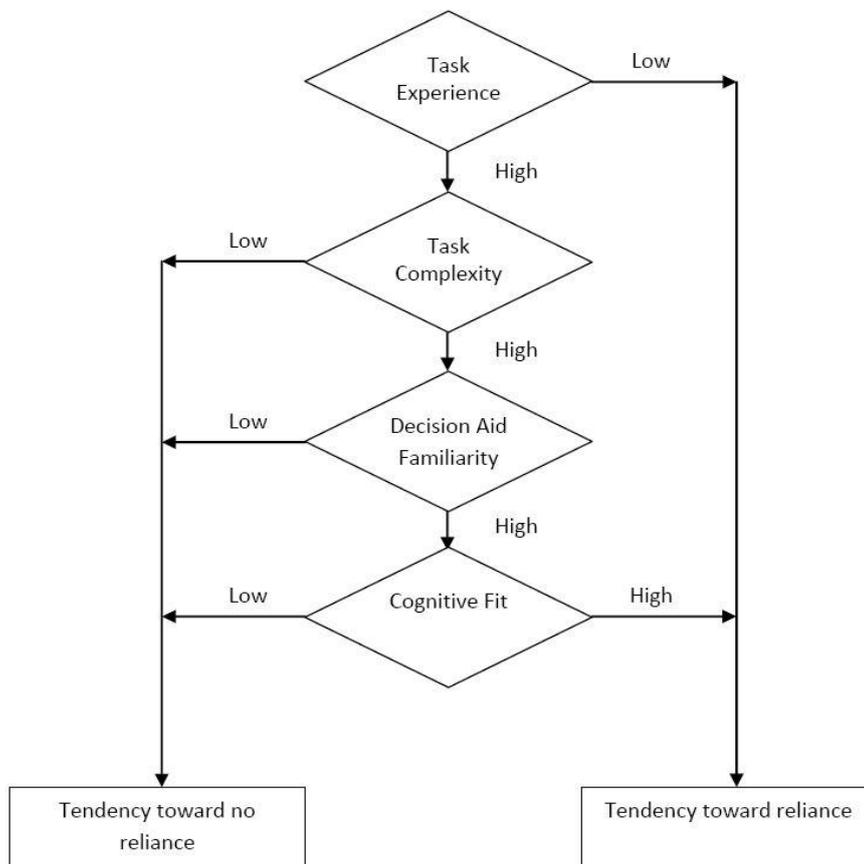


Figure 2.3 Model of reliance (taken from Parasuraman and Riley, 1997)

### 2.8.3 The Theory of Technology Dominance

(Arnold and Sutton, 1998)<sup>156</sup>

One of the simplest models to predict reliance, the TTD (fig 2.4) predicts factors which lead to susceptibility to dominance by technology. This theory posits that reliance, and its appropriate use is a function of task experience, decision aid experience, task complexity and cognitive fit.



**Figure 2.4: Diagram/Schematic of The Theory of Technology Dominance**

Task experience in this model has a negative relationship with reliance. In this theory intelligent decision aids aggravate bias in novices' decision-making but mitigate bias in experts' decision-making processes. This has been backed by other studies<sup>232</sup>.

It predicts a positive relationship between reliance and task complexity, decision aid familiarity and cognitive fit, suggestions which have also been supported<sup>233</sup>.

## **2.9 Conceptual Model of Reliance and Automation Bias**

Using the meta-analytic approach<sup>234</sup>, a broad, conceptual model of factors and some interrelationships (which is formative and not definitive), and decision making processes which are potentially involved in producing the automation bias effect was generated (figs 2.5 and 2.6). This will inform the preliminary development of an ontology of overreliance, and a testable model for AB including some of the most compelling (and feasible to test within the same study) influencing factors in the following sections of the report. A larger version of the conceptual model in fig 2.5 is available in Appendix A.

NB. The literature review and conceptual model are to be submitted to the Health Services Research journal.

## 2.9.1 Model for automation reliance

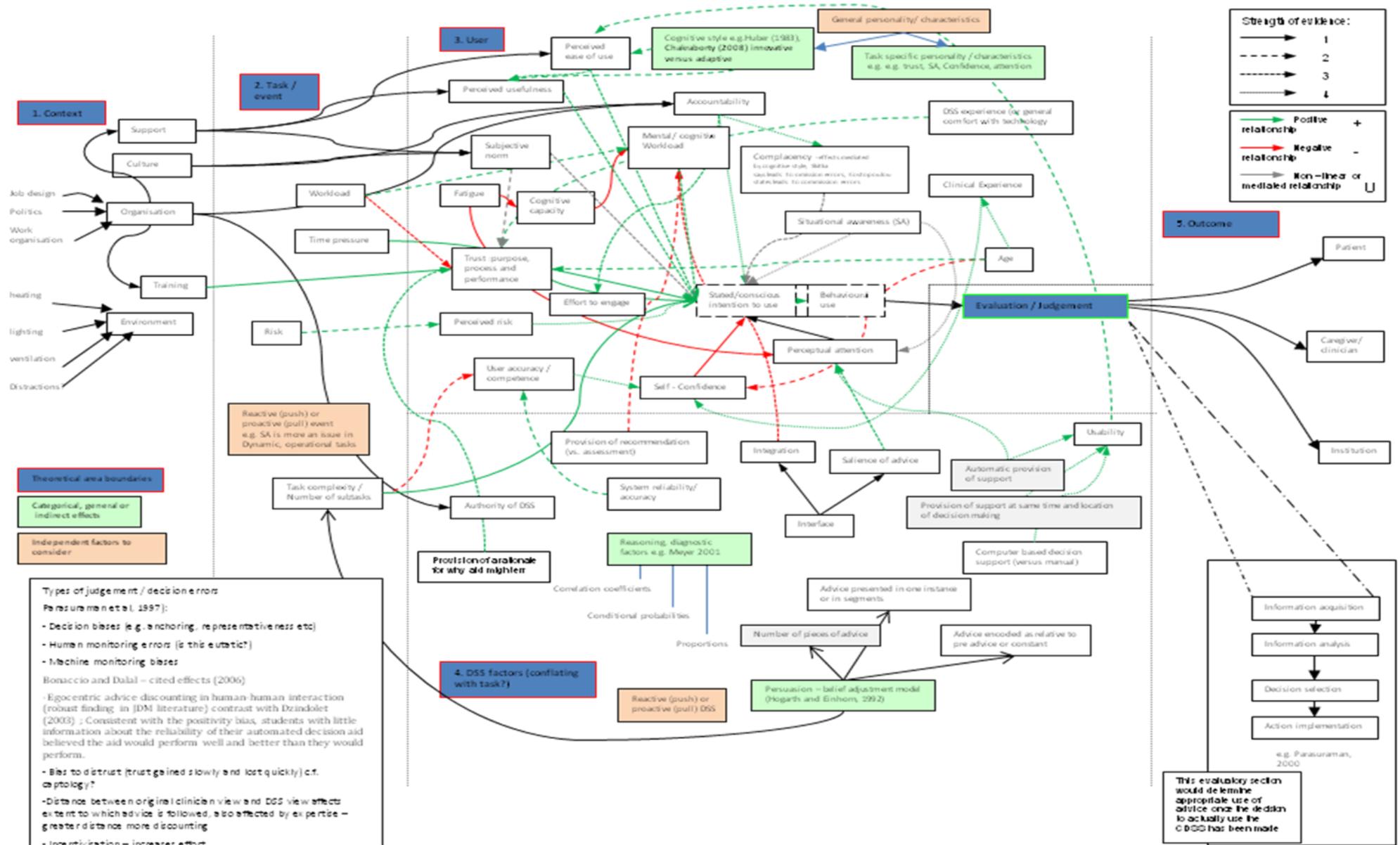


Figure 2.5: Conceptual model of factors and relationships which may lead to reliance and over-reliance in particular

## 2.9.2 Decision making process

Process involved with making a decision to follow advice (see Parasuraman, 2000, and persuasion literature) / Advice acceptance and appropriate response

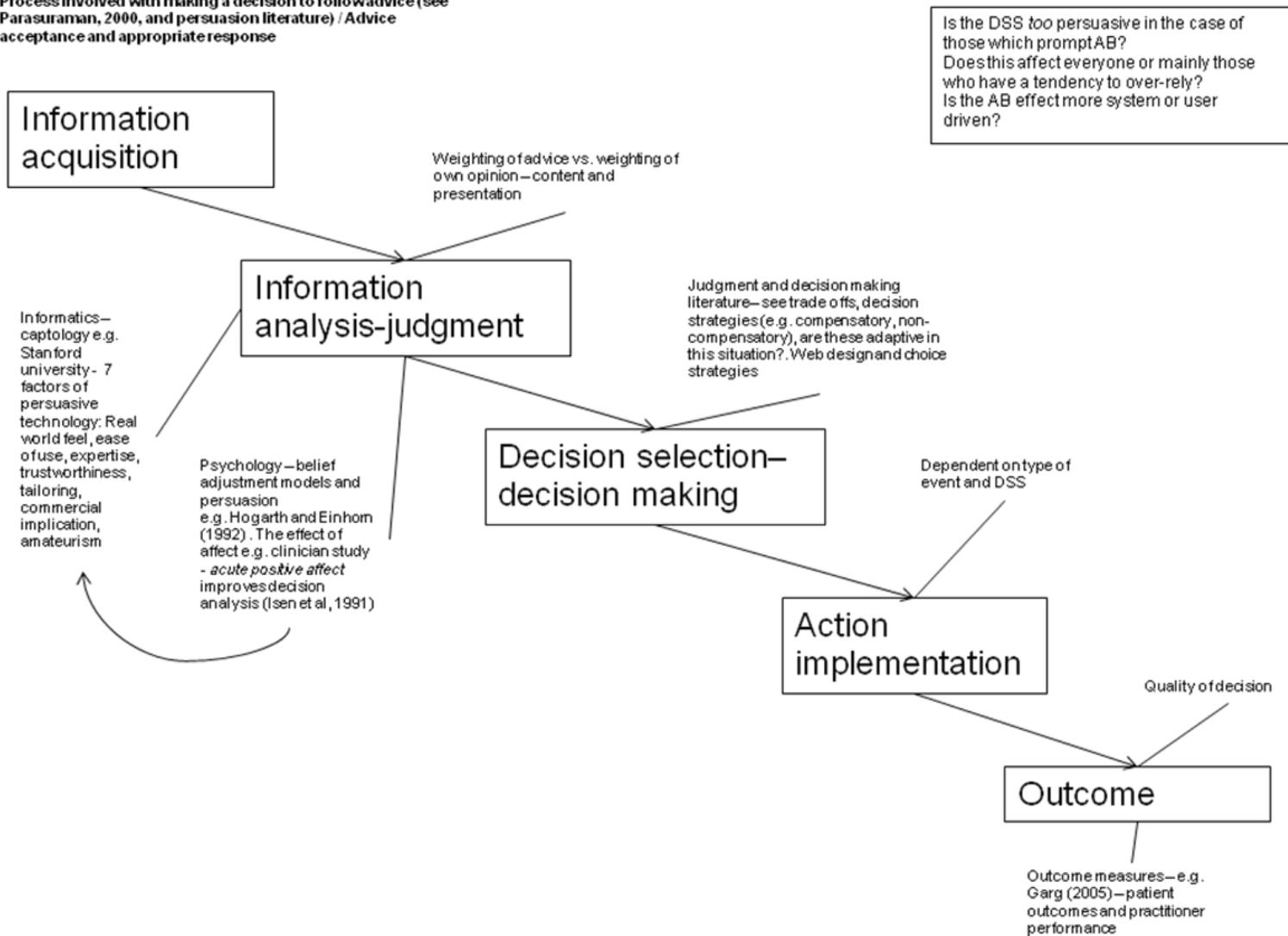


Figure 2.6: Breakdown of the evaluation process and possible factors involved

### 2.9.3 Ontology of Automation Bias

- Why produce an ontology?

Ontologies are formal representations of knowledge structures. They can be utilised to clarify and share a common understanding of semantics and vocabulary, and domain knowledge between people and software agents<sup>235</sup>. In this research it is a potential further application of the conceptual model presented in fig 2.5.

The modelling and creation of an explicit framework and their semantic relationships enables the concept to be represented in a machine readable format and better standardises and defines the currently inadequate definition of AB. In the current context, this may be useful in the domain of decision support, as this information can be incorporated into computer applications and systems and can help to avoid replication and promote interoperability. The term “ontology” covers a range of things including controlled vocabularies e.g. MeSH, hierarchies e.g. gene ontologies, and description logic formalisms e.g. SNOMED-CT. Many medical disciplines have developed general and specialty specific ontologies that domain experts can use to share and annotate information in their fields<sup>236</sup>.

Developing an ontology of AB looks at the concept at a slightly different angle from the conceptual model (fig 2.5). The conceptual model gives information of the *effects* of relevant factors (with some level of evidence base) on each other which concludes with an occurrence (or not) of over-reliance on automated advice; ontology is the study of existence and the broader hierarchical relationships between these factors.

The ontology can be used to illustrate and perhaps predict *instances* of AB or where it is more likely to arise. In doing so it can be used within computer programming to predict situations or instances where AB is more likely; people may need to be given reminders or warnings at point of prescribing to ensure they are aware of the potential for error.

Although there is no consensus over the standard way to develop ontologies, most approaches do have in common certain development elements. Most include 1) a literature review to a) define the scope of the ontology, b) review elements and processes relating to AB, and c) review previous related ontologies, 2) creating an conceptual model of the concept, 3) identifying an upper ontology, and 4) implementing ontology in a formal representation, for example as in Bright et al (2012)<sup>237</sup>.

**Aim:** To formalise instances where AB might be a risk.

### **2.9.3.1 Stage 1: Literature review**

The literature review can be seen in Chapter 2, sections 2.1 – 2.8.

- a) The domain and scope of this project will be fairly broad. Many academic domains have contributed to the literature surrounding the calibration of reliance on automation (examples include, aviation, motoring, and in healthcare CAD studies). Also given that the study of automation bias is fairly new to the healthcare domain, it is wise to maintain flexible criteria in terms of papers to include as relevant when developing a conceptual model of the AB phenomenon.
- b) The elements (classes) and relationships (processes) were investigated in the literature and systematic reviews in Chapters 2 and 3.
- c) A review revealed that to the investigator's knowledge no ontologies exist specifically concerning AB. Most ontologies which exist, particularly in the healthcare world concern concrete entities in biomedicine. Overreliance is a rather more abstract concept. There is however work into ontologies of more abstract psychological concepts, which is perhaps a better match. For example, Lopez et al (2008)<sup>238</sup> produced an ontology to describe emergent emotions and their detection and expression systems, and, as is relevant to this study, they also took into account the effect of context specific factors. There are also a number of decision ontologies which have been fed into applications such as decision support system in health informatics e.g. Nykanen (2003)<sup>239</sup>.

### **2.9.3.2 Stage 2: Conceptual mapping**

The conceptual model (fig 2.5) is used for this section of the process.

The different part of this model can inform the main building blocks of the ontology, and the relationships between them.

### **2.9.3.3 Stage 3: Apply Upper Ontology**

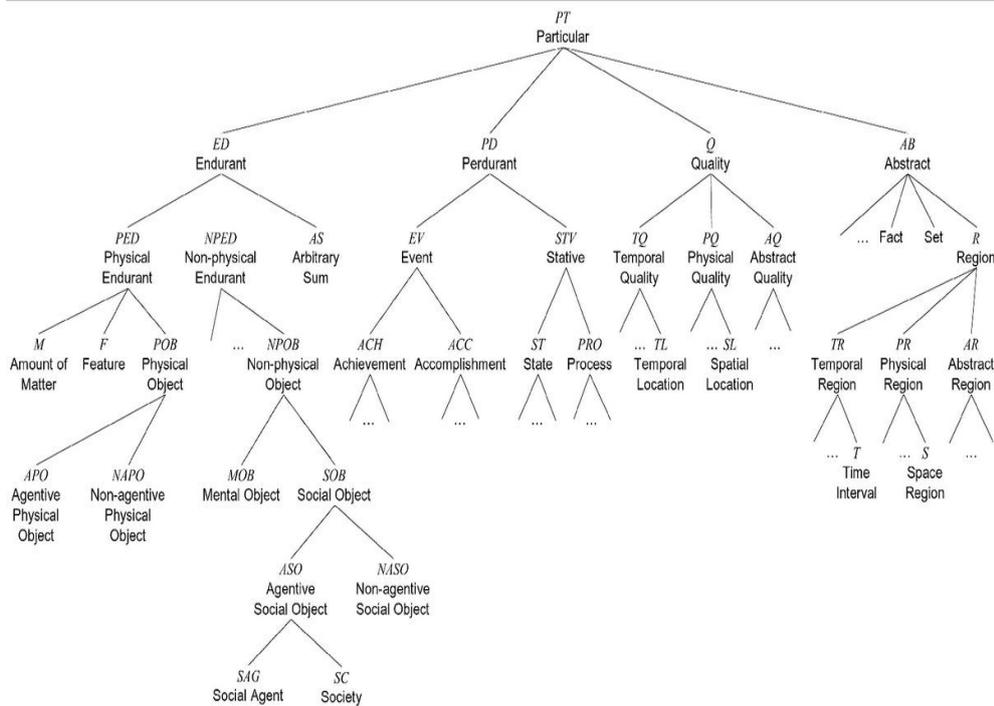
Upper ontologies are generic ontologies about objects or processes; they help clarify and standardise the fundamental semantics of the concept. An upper ontology will be applied, broadly to improve standardisation and generalisation and in doing so promoting interoperability. In developing the ontology this way, it is informed by both bottom-up and top-down processes.

There are a number of commonly used upper ontologies, such as Basic Formal Ontology (BFO), WordNet (designed as a semantic network using psycholinguistic principles) and Cyc

(foundation ontology and several domain-specific ontologies). The DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) upper ontology will be used for this task. DOLCE has a clear cognitive bias and is particularly devoted to the treatment of social entities, such as e.g. organizations, collectives, plans, norms, and information objects. This appears particularly appropriate in light of the Human-Computer Interaction element of this study within an institutional environment. DOLCE enables modelling for situations that may trigger AB.

For this ontology it is possible to supplement the ontology with the crude relationships outlined in the conceptual model, for example, Lopez et al (2008)<sup>238</sup> used self generated OWL properties within the framework of a DOLCE upper ontology e.g. “stores”, “hasInput” and “describes”.

Fig 2.7 outlines the most basic categories of the DOLCE framework. To describe a Particular (or Thing in Protégé OWL), there are three common terms in upper-level formal ontologies: Endurants, Perdurants and Qualities (Abstract, as in fig 2.7, is also seen as a separate higher level term in DOLCE). Endurants are entities which are perceived as complete concepts in time, these can include material objects, or more abstract concepts, such as a country border or a society. Endurants are “is-a” concepts. Perdurants are often what we know as processes, or procedures; they are “happens” concepts. They are related by participation; “an endurant “lives” in time by participating in a perdurant. For example, a person, which is an endurant, may participate in a discussion, which is a perdurant”. Quality describes properties or tropes; these cannot exist in isolation. Examples include colours and sizes. DOLCE also includes the category of Abstract which involves mathematical entities; facts (logical propositions), sets (mathematical sets), and regions (temporal and spatial, time points of intervals or subsets of space).



**Figure 2.7: Taxonomy of basic categories in DOLCE (from Masolo et al, 2003)<sup>240</sup>**

To organise these according to the DOLCE Upper ontology, the groups of instances/ concepts (termed “classes”) from the conceptual model of AB (fig 2.5) are integrated into the categories in the DOLCE ontology. It should be noted that categories are not entirely independent e.g. time pressure can exacerbate workload. There were 64 classes (table 2.3) and 23 example properties (table 2.4; 6 non-DOLCE properties are shown in table 2.5)<sup>6</sup>.

**Table 2.3 Ontology classes, definitions and DOLCE class type**

<b>Class</b>	<b>Definition</b>	<b>DOLCE Class type</b>
Context	The general environment a prescribing decision is being made in	Perdurant; STV
Task/Event	The specific situation a prescribing decision is being made in	Perdurant; ST
User	The agent of decision making, in this case the prescribing clinician	Endurant; SA
CDSS	The automated decision support system (in this case, to aid prescribing)	Endurant; NAPO
Outcome	The patient-focussed results of the decision	Perdurant; ST
Support	The plan for maintenance of users of the system, inclusive of training plans and real time problem	Endurant; NASO

<sup>6</sup> Class type definitions were taken from the W3C website : [http://www.w3.org/2001/sw/BestPractices/WNET/DLP3941\\_daml.html](http://www.w3.org/2001/sw/BestPractices/WNET/DLP3941_daml.html)

	solving	
Culture	The broad social and organisational environment within which decisions are made	Endurant; NASO
Organisation	The hierarchy and protocols involved in running an institution	Endurant; NASO
Training	The guided experience people are given relating to the CDSS	Endurant; NASO
Workload	Amount of work expected during a particular time period putting pressure on cognitive resources	Perdurant; STV
Time pressure	The limitations of a deadline to complete an action / actions	Perdurant; ST
Risk	The potential for harm / loss / danger	Quality; AQ
Type of task	Pull / push (ontology individuals), reactive versus proactive	Perdurant; ST
Task complexity	How many components within a task, component of task difficulty	Perdurant; ST
Perceived usefulness	The perception of the benefit (in this case, of taking automated advice)	Endurant; MO
Perceived ease of use	The perceived usability of a system	Endurant; MO
Subjective norm	"the person's perception that most people who are important to him or her think he should or should not perform the behavior in question" <sup>241</sup>	Endurant; MO
Fatigue	Weariness resulting from exertion, can relate to repetitive tasks	Perdurant; ST > cognitive state
Cognitive capacity	The amount of information that a person can retain and process at any particular time	Quality; AQ
Trust	In this case, the belief that an automated DSS provides reliable information Purpose, process and performance?	Endurant; MO
Perceived risk	The perception of the probability for harm / loss / danger	Endurant; MO
User accuracy / competence	The ability of a user to carry out actions successfully and efficiently	Quality; AQ

	Accuracy as a subset of competence	
Effort to engage	The driven attempt to use DSS, particularly if novel Driven by motivation	Quality; AQ
Self confidence	Belief in the abilities of oneself	Endurant; MO
Accountability	The personal responsibility and answerability attached to decisions	Quality; AQ
Mental/ cognitive workload	The pressure specifically felt on cognitive resources	Quality; AQ
Cognitive style	The way individuals think, perceive and remember information Innovative versus adaptive	Quality; AQ
Complacency	Not been well operationally defined in the literature; linked to deficient cross-verification, lower effort to engage, loss of situational awareness	Endurant; MO
Situational awareness	The user's perception and understanding of environmental elements in time and/or space	Endurant; MO
General personality	The combination of a person's characteristics or qualities Linked to cognitive style	Quality; AQ
Task specific personality	Produces the attitudes towards DSS involved tasks Subset of general personality	Quality; AQ
Perceptual attention	The mental focus linked to the perception of external stimuli	Perdurant; ST > cognitive state
Clinical experience	The length of time working in clinical environment	Quality; TQ
CDSS experience	The amount of exposure someone has to CDSS	Quality; TQ
Age	Length of time person has existed	Quality; TQ
Conscious /stated intention to use	The aim to use DSS	Perdurant; ACC
Behavioural use	The actual/observable use of DSS	Perdurant; ST > Decision state
Judgment	The rationale behind a decision	Perdurant; ST

Information acquisition	The gaining of information to base a decision on Subset of judgement	Perdurant; ACC
Information analysis	Examination of information Subset of judgement	Perdurant; PRO
Decision	The conclusion reached after deliberation Subset of judgement	Perdurant; ACC > Decision activity
Authority of CDSS	The importance placed on the use of the DSS by the organisation	Quality; AQ
Provision of decision rationale	The DSS providing the reasoning sources and/or processes behind a decision/ piece of advice	Endurant; NASO > information object
Mode of advice	How advice is framed e.g. Recommendation or assessment	Endurant; NASO > information object >formal expression
System reliability	Dependability. In this case the DSS likelihood to provide accurate advice A subset is Accuracy	Quality; AQ
Interface	The visible portion of the DSS	Endurant; NAPO
Advice	The information given by CDSS	Endurant; NASO > information object
Advice integration	How well the advice is integrated into both the interface and workflow Subset of interface	Endurant; NASO > description > system design
Advice salience	How perceptually noticeable the advice is Subset of interface	Endurant; NASO > description > system design
Persuasiveness of advice	How compelling the advice is	Quality; AQ
Number of pieces of advice	Number of units of advice, can also be components of one overall piece of advice Subset of persuasiveness	Endurant; NASO > description > information encoding system
Advice presented in stages or in one instance	Subset of persuasiveness	Endurant; NASO > description > information encoding system

Advice coded relative to pre advice opinion or constant	Whether the advice takes account of pre advice opinion (more critique), or not (independent recommendation) Possibly a subset of persuasiveness	Endurant; NASO > description > information encoding system
Location	Provision of advice at the time and place of decision-making	Quality; PQ
Usability	International Organisation for Standardisation (ISO) defines usability as "The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use."	Quality; AQ
Reasoning - diagnostic factors	How the advice is generated	Endurant; NASO > description > information encoding system > combinatorial system
Correlation coefficients	An example of how a diagnostic property is expressed e.g. in warning DSS Subset of reasoning diagnostic factors	Endurant; NASO > description > information encoding system
Conditional probabilities	An example of how a diagnostic property is expressed e.g. in warning DSS Subset of reasoning diagnostic factors	Endurant; NASO > description > information encoding system
Proportions	An example of how a diagnostic property is expressed e.g. in warning DSS Subset of reasoning diagnostic factors	Endurant; NASO > description > information encoding system
Automatic provision	Characteristic of advice – advice provided without prompting vs. prompting	Endurant; NASO > description > system design
Format of advice	E.g. computer support versus manual support	Endurant; NASO > description > system design

Patient	Subset of outcome	Endurant; NASO > SOB > socially constructed person
Clinician	Subset of outcome	Endurant; NASO > SOB > agentive social object
Clinical institution	Broader definition of organisation (organisation is more focussed on structure and process) Subset of outcome	Endurant; SOB > institution

#### 2.9.3.4 Stage 4: Basic Ontology

It is not within the scope of this PhD to develop a complex ontology of AB. Here the basic initial set up is described in terms of structure according to the DOLCE upper ontology described above, and future steps suggested. It is generated in Protégé, and thus represented in Web Ontology Language (OWL). The Web Ontology Language has more facilities for expressing meaning and semantics than other ontology languages such as XML and RDF (see Box 1 below). Thus, OWL goes beyond these languages in its ability to represent machine interpretable content, especially on the Web.

##### Box 1: Why Protégé OWL?

The advantages of the Protégé tool include that it can be used by domain experts, and has better scalability than similar modelling languages such as UML (Unified Modeling Language). It enables rapid prototyping of models and provides reasoning support at edit-time. Its open architecture allows adaptability – a programmer can integrate plug-ins, which can appear as separate tabs, specific user interface components (widgets), or perform any other task on the current model. It also benefits from a support community (beneficial for beginners).

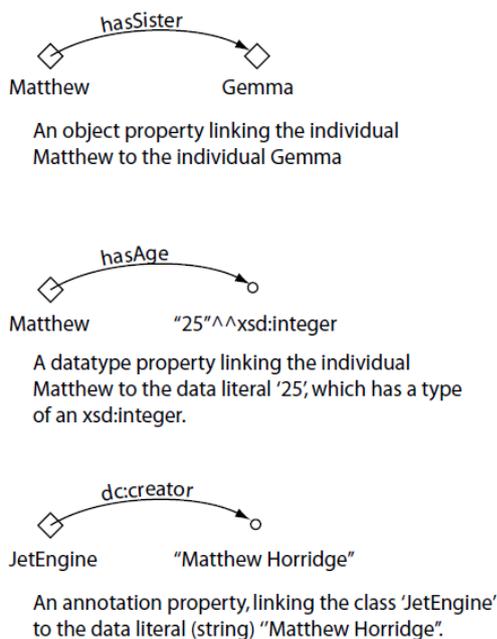
Protégé OWL will be used over other languages such as UML, as comparatively it has:

- Explicit, sharable modeling artifacts
- Open architecture of Semantic Web
- OWL has rich semantics
  - closer to domain than UML
  - built-in reasoning support (DL, SWRL)
- A single language across metalevels

There are 3 OWL basic components – class, individuals and properties. Classes are basic building blocks of an ontology, organised into a hierarchy (which can be carried out according

to pre-existing upper ontology, or from scratch). The word *concept* can often be used for class and they represent a range of individual instances.

Properties in OWL represent relationships. Properties can be categorized as object properties, which relate individuals to other individuals, and datatype properties, which relate individuals to datatype values, such as integers, floats, and strings. OWL includes annotation properties can be used to add information (metadata — data about data) to classes, individuals and object/datatype properties. Examples of these shown in fig 2.8 below:



**Figure 2.8 The Different types of OWL Properties (taken from [http://owl.cs.manchester.ac.uk/tutorials/protegeowltutorial/resources/ProtegeOWLTutorialP4\\_v1\\_1.pdf](http://owl.cs.manchester.ac.uk/tutorials/protegeowltutorial/resources/ProtegeOWLTutorialP4_v1_1.pdf))**

There are 7 types of property: functional (there can be at most one individual that is related to the individual via the property), inverse functional (the inverse property is functional), transitive (a property links A to B and B to C, then if one can infer that it links A to C), symmetric (a property linking A to B, can be inferred to link B to A), antisymmetric (if a I related to individual b then individual b cannot be related to individual a via the same property), reflexive (the property must relate individual a to itself), and irreflexive (a property that relates an individual a to individual b, where individual a and individual b are not the same).

The DOLCE upper ontology has a number of properties which can be applied here. There are 299 properties in DOLCE which are organised into 4 superproperties:

1. Immediate-relation - A relation that does not involve mediators; a non composite relation.

2. Immediate-relation-i [inverse] – Inverse of Immediate-relation.
3. Mediated-relation – A relation that is composed of other relations.
4. Mediated-relation-i [inverse] – Inverse of Mediated-relation

Table 2.4 below has suggestions for DOLCE properties to be applied to an AB ontology for potential future development. The first column lists some DOLCE predefined properties which may be used for the conceptual model factors, the second column defines its associated property hierarchy, and the third column gives an example of the type of relationship the property could describe. DOLCE also recognises that there are overlaps in classes and properties.

**Table 2.4 Ontology properties**

<b>DOLCE Property examples</b>	<b>Property hierarchy</b>	<b>Example relationship</b>
<u>Specific-constant-dependent</u> : Constant dependence between two individuals	Immediate-relation> specific-constant- dependent	Use of CDSS depends on the trust in the CDSS
<u>Specific-constant-constituent</u> : An entity constituting a setting	Immediate-relation> specific-constant- constituent	The culture within a clinical institution
<u>Acts-for</u> : An agentive object can act on behalf of a bestowing power e.g. an employee acts for an organisation that deposes their role	Immediate-relation> acts-for	The clinician acts for the clinical organisation
<u>Attitude-towards</u> : Used to state attitudes, attention, or even subjection that an agent can have towards an action or process.	Immediate-relation> modal-target> attitude-towards	Perceptual attention a user can have towards and interface
<u>Conceives</u> : An agent can form a conception via a mental state and event.	Immediate-relation> conceives	A clinician conceives mental objects/states
<u>Main-goal</u> : Relation between a plan and an end goal	Immediate-relation> part> proper-part> main-goal	Judgement aiming for correct clinical decision
<u>Place</u> : The location of a physical endurant	Mediated-relation> generic-location> approximate-location> place	Interface within the CDSS

<u>Realised-by</u> : An information object is realised about a particular	Immediate-relation> references> realised-by	Advice is taken by clinician
<u>Strong-connection</u> : A connection between 2 entities	Mediated-relation> strong connection	Broad relationship e.g. clinical experience and competence, judgement and decision
<u>Uses</u> : An endurant uses another endurant within a perdurant	Mediated-relation> co-participates-with> uses	The clinician uses the CDSS
<u>Component</u> : A proper part with a role/function in a system/context	Immediate-relation> proper-part> part> component	Personality and cognitive style as a component of the clinician
<u>Characterises</u> : A role that describes a social object	Immediate-relation> characterises	The type of advice that's given (e.g. recommendation or assessment)
<u>Functional-participant</u> : A perdurant is participated in by an object	Immediate-relation> functional-participant	A clinician participates in the judgement process
<u>Participant</u> : Relation between perdurant and endurant	Immediate-relation> participant	A clinician can experience fatigue e.g. the clinician is tired
<u>Setting-for</u> : The relation between situation and the entities within it	Immediate-relation> specific-constant-constituent	A task can be the setting for task complexity
<u>Regulates</u> : Descriptions for the social world; the factors which subjectively dictate how a situation looks	Immediate-relation-i> satisfied-by> regulates	High complexity and workload can regulate a difficult task
<u>Deputes</u> : Figures that can give roles to an endurant entity	Immediate-relation> deputes	Organisation deputes "authority" or "pedigree" of CDSS
<u>Result-of</u> : One perdurant resulting from another	Mediated-relation-i> temporal-relation-i> follows> result-of	Decision as result of judgement
<u>Adopts</u> : An actual desire to perform (or not) the expected action.	Immediate-relation> conceives> adopts	The clinician can adopt the intention to use the CDSS
<u>Generic-location</u> : Very broad definition here an individual is in relation to another individual e.g. may be mental schema, exact or approximate location.	Mediated-relation> generic location	The location of the CDSS (e.g. point of diagnosis/ prescribing, remote etc.)
<u>Inherent-in</u> : The quality of an entity	Immediate-relation> inherent-in	Personality is inherent in a clinician
<u>Functional-participant</u> : An endurant participates in a perdurant within a specific	Immediate-relation> participant-> functional-participant	A clinician acting to make a clinical decision

description		
<u>Interprets</u> : An object that expresses a conception as an information object	Immediate-relation> interprets	CDSS interpreting the situation to generate advice

Depending on the extent and detail of the ontology, the DOLCE properties may be used, or idiosyncratic properties can be generated. Lopez et al (2008)<sup>222</sup> used the DOLCE upper ontology to structure an ontology for describing emotions. For the properties they generated independent OWL object properties such as *hasInput*, *hasOutput*, and *triggers*.

To broaden the ontology to include the crude effects in the conceptual model, the ontology properties listed above in table 2.4 can be supplemented by idiosyncratic properties from this ontology (table 2.5 below):

**Table 2.5 Conceptual model properties**

<b>Conceptual model effect</b>	<b>Description</b>	<b>DOLCE superproperty</b>
correlatesWith <ul style="list-style-type: none"> <li>- positive</li> <li>- negative</li> </ul>	To denote the direction of a relationship – whether - an increase in one class is related to an increase in the other - an increase in one class is related to a decrease in the other	Mediated-effect Immediate-effect (dependent on evidence and influence of other factors)
isDependentOn	To indicate a causal relationship of two classes class	Immediate effect
hasNonLinearRelationshipWith	To denote that whilst there is a relationship, it is not a direct dependency	Mediated-effect Immediate-effect (dependent on evidence and influence of other factors)
isMediatorOf	A factor that affects a relationship between two entities	Mediated-effect
isSubsetOf	To indicate the composite parts of a class	Immediate-effect

Classes can be populated with *individuals*. Individuals represent instances of a class. In this ontology, instances would apply to specific situations where automation bias may have occurred. For example the age, and clinical experience of a particular clinician would be 2 specific numeric instances within the ontology.

Individuals and classes can be similar – the definition depending on the degree of granularity of the ontology (i.e. individuals are seen as atomic, but can still be further broken down into components, depending on scope).

### ***Automation Bias ontology visualisation***

In this section a basic class taxonomy of factors which may lead to AB has been described. The more complex task of attributing individuals and properties to the ontology is a potential matter for future work.

The OWL versions of DOLCE (DOLCE-lite.owl from the DLP3971.zip downloaded from <http://www.loa.istc.cnr.it/DOLCE.html>), have been loaded into Protégé 4.2 build 249 [accessed September 2011]. The ontology was saved in the Manchester syntax. This syntax was used as though it borrows ideas from the OWL Abstract Syntax, it is much less verbose, meaning that it is quicker to write and easier to read and it is this perceived as more user-friendly. Other syntaxes, such as the more common XML/RDF would also have been appropriate to use.

Protégé generates axioms automatically (however closure axioms and restrictions can be input) - which provide explicit logical assertions about the **classes, individuals and properties**). A piece of software called a *reasoner* can infer other facts which are implicitly contained in the ontology, for example if an individual Bob is in class Student, and the class Student is a subclass of the class Person, a reasoner will infer that Bob is a Person.

There are many type of axioms such as Class Declaration (defines a class), Individual Declaration (defining individuals), Class Assertion (an individual belongs to a class), and Property Declaration (defines either a data property to link an individual to data, or object property to link to an individual).

Ontology development presents challenges, particularly to new developers, in that it has to be represented in a formal language. The Protege visual development tool is used as it is generic and flexible and visually represents the ontology in a machine readable format.

The ontology hierarchy according to the DOLCE framework is illustrated according to the OntoGraf programme in the 3 figures below (a figure per the 3 superclasses of enduring, perdurant and quality). OntoGraf is used to visualise the structure of the ontology:

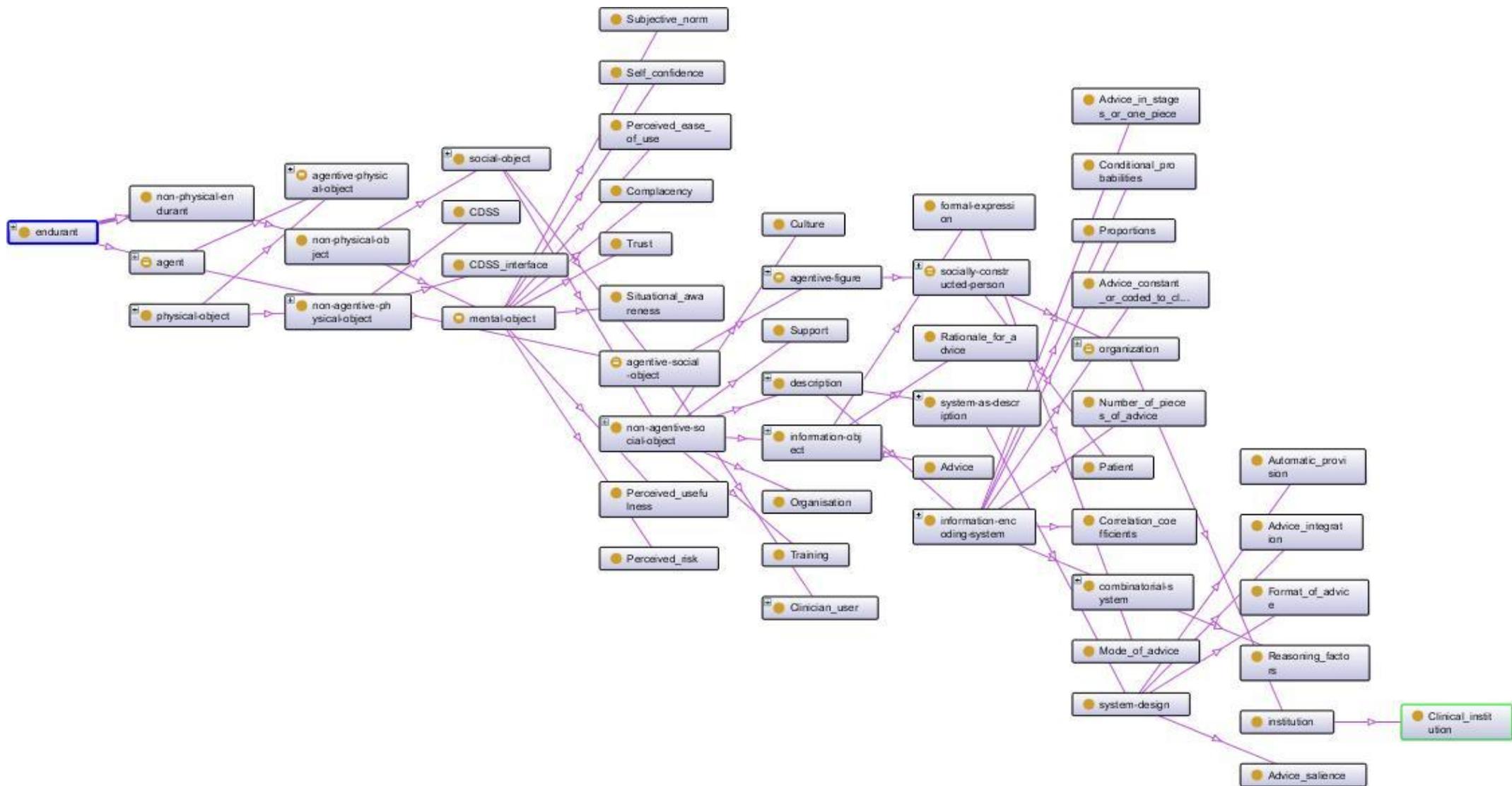


Figure 2.9 The enduring branch of the ontology as illustrated in OntoGraf

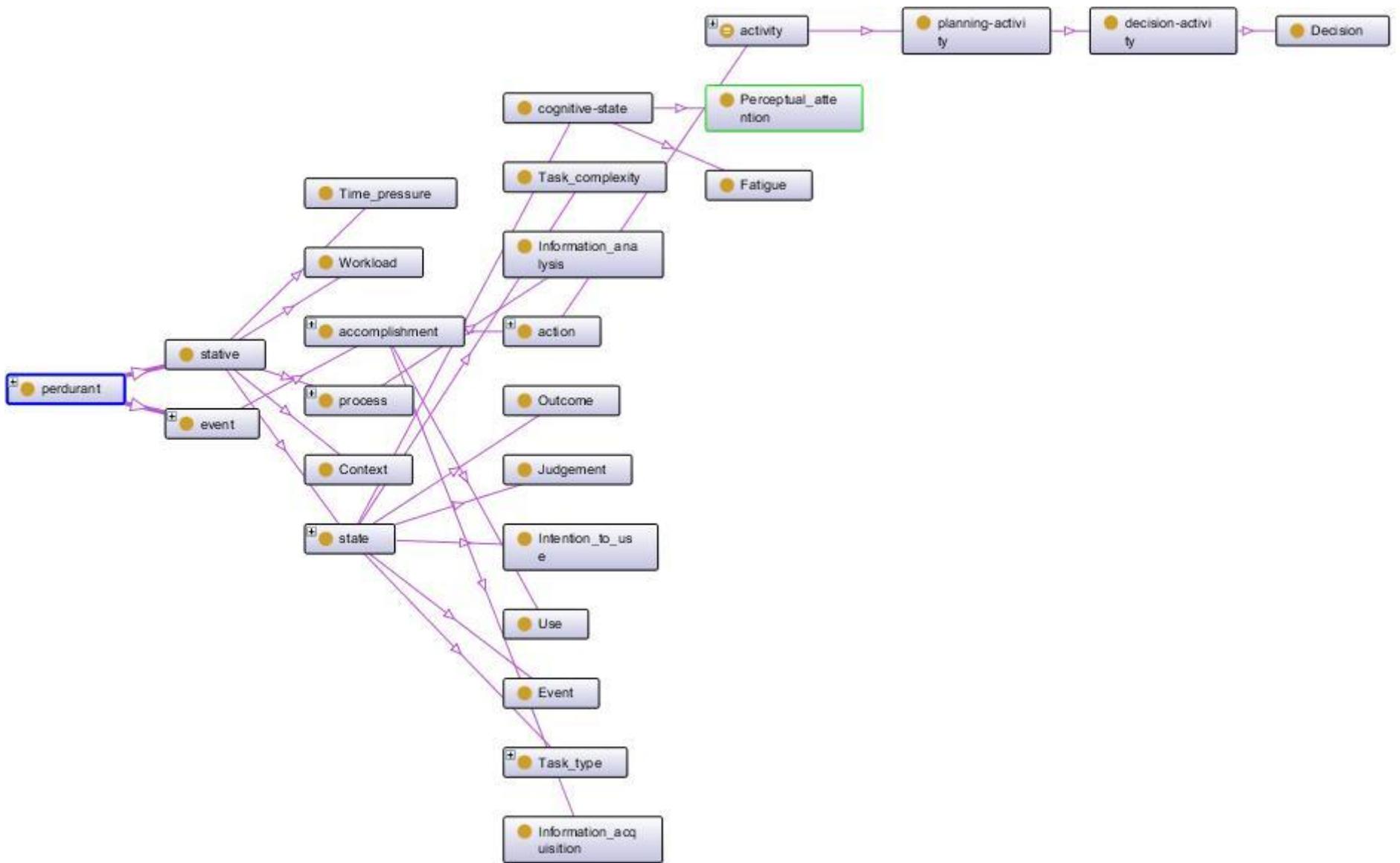


Figure 2.10 The perdurant branch of the ontology as illustrated in OntoGraf

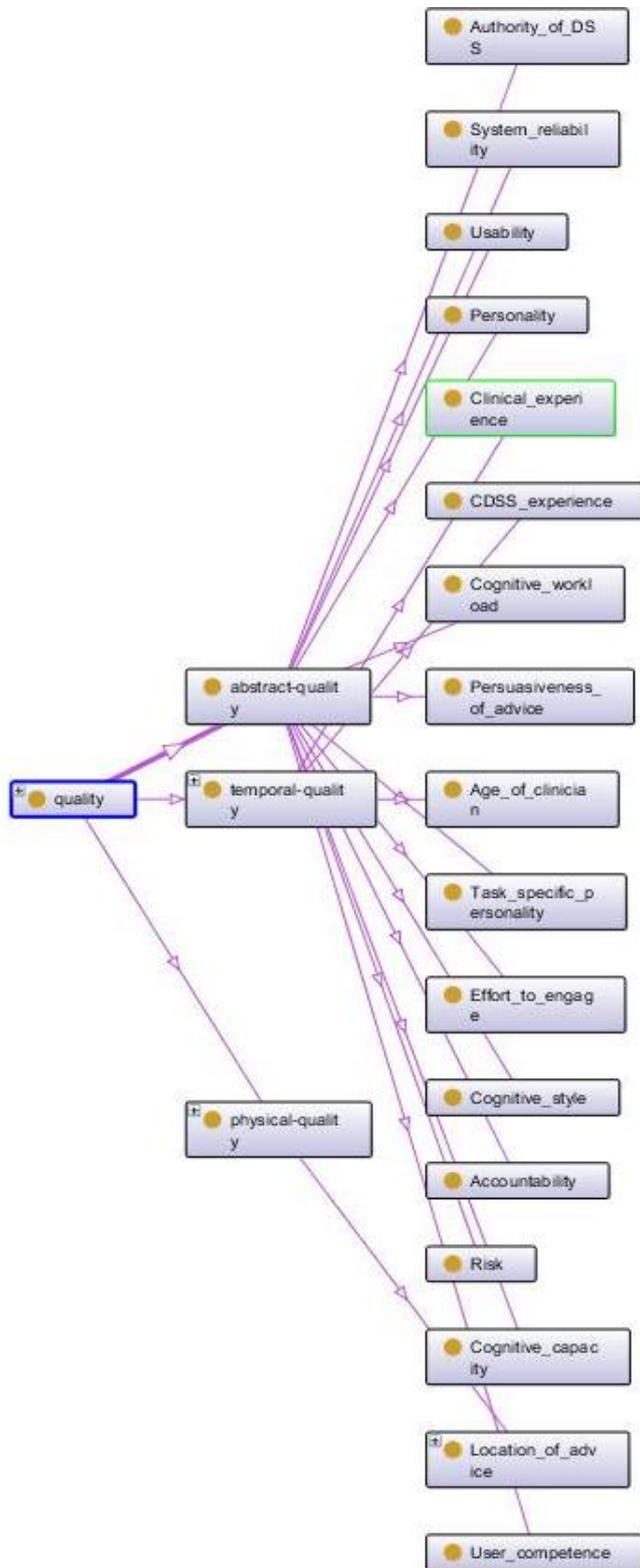


Figure 2.11 The quality branch of the ontology as illustrated in OntoGraf

This ontology is a first step, and requires further iterative evaluation and development. A number of papers e.g. Lopez et al (2008)<sup>222</sup> for emergent emotions, and Eapen (2008)<sup>242</sup> for the dermatology domain appear in the published literature as preliminary models calling on peers for their input and evaluation in the ongoing development of the ontology.

### **2.9.3.5 Stage 5: Evaluating the ontology**

The property and class taxonomy is formative, not exhaustive and needs evaluating. This validity of the evaluation could also be assessed at the levels of: Taxonomic and semantic relations, and structure.

A more valid ontology needs constant updating and maintenance, and matures over time with input from ontology and domain experts. A method by which this could be achieved would be to set up an online project page to open up to domain and ontology experts licensed under a General Public Licence (GPL) to contribute and enhance. Domain expert review would entail a more generic qualitative sense check (due to it being a new ontology). This could then be followed up with evaluation using a laddering technique<sup>243</sup> which has been used in both the psychology (e.g. Bannister and Fransella, 1989<sup>244</sup>) and knowledge acquisition (e.g. Shaw and Gaines, 1988<sup>245</sup>) domains. Briefly, this involves the creation, review and modification of hierarchical knowledge via tree diagrams. This is carried out by giving interviewees a guide to the domain under investigation – in this case the literature review with clear elements involved in the AB domain stated. The interviewees are then asked upwards, downwards or sideways probe questions to clarify their mental model hierarchy. Corbridge et al. (1994)<sup>227</sup> emphasised that the questions/probes used during the laddering process should be standardised. The general rules given by Stewart and Stewart (1981)<sup>246</sup>, for example, recommend adapting the use of the probe ‘why is that important to you?’ to take the participants higher up their pyramids, while probes such as ‘how is it different?’ will move lower. The structured questions would serve as probes to elicit the participants understanding of why AB might arise within the clinical context as input to the hierarchy. The hierarchy could be visually presented to the participant as it was developing according to their answers until they agreed that this was an accurate representation of the structures leading to AB.

The ontology could also be compared retrospectively with anecdotal evidence of situations where AB has occurred (more rigorous empirical evidence is scant, as previously discussed).

To decrease chances of selective bias the most useful evidence for the ontology's validity would be to prospectively capture instances of AB and match the situation and factors which are likely to have caused overreliance. Essentially, the ontology needs to be tested in situations where AB has arisen, and to match whether it accurately describes certain situations where this has happened.

There is no "gold standard" methodology for defining ontologies. Their development is a creative and fairly subjective process based on shared human understanding.

## 2.10 Summary

In this chapter, the literature surrounding the problem of AB was explored, in terms of the application of CDSS to mitigate medical decision making errors, particularly in the domain of prescribing.

There is an increasing concern with problems of prescribing; the GMC found on average 8.9% of NHS prescriptions contained error. There is also high variation of intra and international prescribing due to differences in evidence weighting, prescribing systems, drug names, culture (this also limits participant recruitment in proposed empirical studies). As part of the drive to prevent error, CDSS can be implemented to cover gaps in knowledge and promote best practice.

However, whilst most reviews and accepted knowledge adopt the view that CDSS are generally beneficial, the unexpected new errors that they can generate are a newer field of study. This research aimed to look at overreliance, or a bias towards decision support technology, which can result in systematic errors when the CDSS is inaccurate, which is little studied and poorly defined in the literature.

The review took a broad multi-domain exploration of potential factors which may drive or influence AB, and related models, and generated a conceptual model of factors and posited relationships. Factors were grouped into contextual, task-specific, CDSS and user overall components. It also yielded a preliminary ontology (for formative evaluation and development) as an output to help better define and provide a starting point for standardisation for reliance. The review also established that there was a gap in evidence in terms of deliberate empirical studies into the effect.

The results of the broader literature review lead into a more stringent search for evidence in the next section – a systematic review of the rates, influencing factors, and potential mitigators of overreliance. The results of these reviews will inform a testable model of overreliance later in the report to test generated hypotheses.

### 3. Systematic Review

#### 3.1 Introduction

Though a number of papers provide anecdotal evidence<sup>247,248</sup> and propose mechanisms for over reliance on automation, there is a paucity of deliberate empirical evidence for its existence, its scale and its causes. This chapter aims to more systematically investigate evidence for the rate of AB, the influencing factor, ways to mitigate this effect, and the gaps in the literature.

A broad preliminary literature review yielded papers which suggest possible factors involved in automation reliance and bias looking at both theoretical and empirical papers. The main concept surrounding the subject of AB is that of the DSS or automation – human interaction and the resultant task performance and error generation. These were the key themes used in a systematic search of the literature. Automation and Decision Support Systems are similar concepts and are thus both included in the overall technical concept to mitigate ruling out papers on the basis of this similarity of definition (fig 3.1), i.e. areas 1 and 2 are the areas involved in the search.

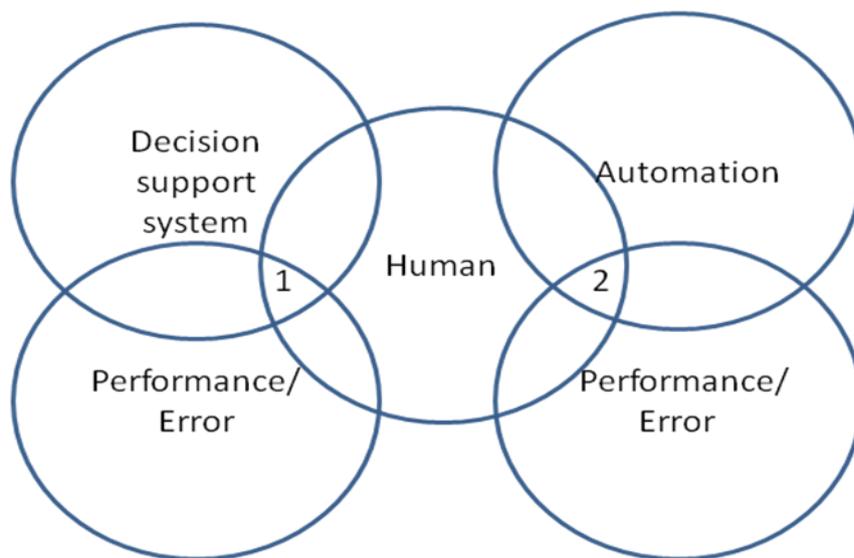


Figure 3.1 Venn diagram of search concepts and overlaps which comprised search target areas (1 and 2)

#### Automation Bias and Complacency

In a recent literature review, Parasuraman et al. (2010)<sup>249</sup> discussed AB alongside automation-induced complacency as overlapping concepts reflecting the same kind of automation misuse associated with misplaced attention; either an attentional bias towards DSS output, or

insufficient attention and monitoring of automation output (particularly with automation deemed reliable). They noted that commission and omission errors can be found as outcomes of both AB and complacency (although they mention that commission errors are more strictly the domain of AB). There is a lack of consensus over the definition of complacency.

Complacency appears to occur as an attention allocation strategy in multitask type situations where manual tasks are attended to over monitoring the veracity of output of automation. AB can be found outside of multitask situations, and occurs when there is an active bias towards DSS in decision making.

Though the focus of the review is on AB, due to the theoretical overlap and vagaries with current definitions, this review will not exclude papers which imply complacency effects as it will still inform the misuse or overreliance on automation literature. Similar outcomes in terms of commission or omission may mean that one effect may be conflated by or confused with another. Studies relating to automation bias are distinguished and separately examined from those relating to automation complacency.

### **3.2 Review aim and objectives**

The overall review aim is to systematically review the literature surrounding DSSs and AB, particularly in the field of healthcare.

The specific review objectives are to answer the following questions:

- What is the rate of AB and what is the size of the problem
- Does it vary in different studies and settings? i.e. focus on causes, risk factors, barriers and facilitators and types of users.
- Is there a way to avoid AB? What is the impact of various methods to reduce AB?

### **3.3 Review methods**

#### **3.3.1 Sources of studies**

The main concepts surrounding the subject of AB are that of the DSS intervention, the DSS–human interaction, and the resultant task performance and error generation. These were the key themes used in a systematic search of the literature. Given that initial searches indicated a relative paucity of healthcare specific evidence it was decided to include a number of databases

and maintain wide parameters for inclusion/exclusion to identify an optimal number of relevant articles.

The search took place between September 2009 and January 2010. The following sources of evidence were searched to identify articles relevant to this review: MEDLINE/PubMed, CINAHL, PsycInfo, IEEE Explore and Web of Science.

No timeframe limit was set for any database, the language filter was set to English language studies only.

### **3.3.2 Search strategy**

PRISMA methodology was used as guidance for the selection of papers, involving phases of identification, screening, assessing for eligibility and qualitative and quantitative assessment of final papers.

Combinations of subject-specific free text and index terms were used to search electronic databases. No timeframe limit was set for any database, the language filter was set to English language studies only. The study types included within the PubMed/Medline filter were Randomized Controlled Trials, and Comparative studies. All funding types were included in the search. Non-PubMed/Medline searches used these criteria if available, otherwise studies were chosen by hand.

### **3.3.3 Search terms:**

The index and freetext search terms included in the concepts were derived from MeSH and the preliminary literature search.

#### **From MeSH :**

*Concept 1:* The technical concept of automation or computerised decision support:

- i) Automation,
- ii) Clinical decision support systems, Decision support techniques, Computer-assisted decision support, Medical order entry systems.

*Concept 2:* The human factors concept: Humans, Human engineering, Psychological phenomena and processes, Behavior and behaviour mechanisms.

*Concept 3:* The task performance/error concept: Task performance and analysis, Medical Errors, Bias (epidemiology), Sensitivity and specificity.

***Additional CINAHL-specific indexed search terms:***

Concept 2: Psychology, applied, Psychological processes and principles, Behavior,

Concept 3: Diagnostic errors, Human error, Health care errors. Measurement error, Medication errors, Treatment errors

***Additional PsycInfo-specific indexed search terms:***

*Concept 2:* Human factors engineering, Psychosocial factors, Human machine systems, Personality traits, Human machine system design, Attitude formation, Psychology, Cognitive bias

*Concept 3:* Error of measurement, Errors, Prediction errors, Response bias, Sensitivity (personality), Interpretive bias, Performance, Error Analysis

***Other potential terms (non MeSH, freetext terms):***

*Concept 1:* Decision support, automated decision aid

*Concept 4:* Automation bias and synonyms: Automation bias, confirmation bias, automation dependence, complacency, over reliance

**3.3.4 Databases**

1. PubMed/MEDLINE
2. CINAHL
3. IEEE
4. Ebsco (PsychInfo)
5. WoS

**Table 3.1: Combinations of search concepts in the systematic review**

<b>Literature Search Terms for Review and Stage One of Retrieval Process</b>							
<b>Search Terms</b>			<b>Databases</b>				
			<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>
Concept 1	Concept 2	Concept 3	8888	169	2768	31	466
Concept 1	Concept 4		830	43	344	34	248
<b>Total</b>			<b>9718</b>	<b>212</b>	<b>3112</b>	<b>65</b>	<b>714</b>

In databases without controlled indexed terms, variations on the terms used in the 3 concepts explored were used (e.g. Boolean search terms and \*, \$ and ?). Again broad search terms were required in order to capture the widest range of articles

### **3.4 Eligibility criteria for studies**

It is clear from preliminary searching that the research should not be limited by a specific field. Investigators into decision support and automation from non-healthcare disciplines have valuable input to highlight factors in human-computer systems, the formation of cognitive biases and recommendations on how to debias individuals. Due to the exploratory nature of research a broad, multi disciplinary search was justified in order to seek out the most relevant cross-section of papers. Also, all study setting were considered whether “in the field” or laboratory studies. All participant types were accepted from naïve to those with field expertise. The first search of databases led to 14457 research papers (inclusive of duplicates).

#### **3.4.1 Inclusion criteria**

- Papers that examined human interaction with automated decision support were included from various fields (such as aviation, motoring and cognitive psychology). In particular, the field of healthcare was the focus of interest.
- Papers that studied empirical automation use were included, including those which had a subjective participant questionnaire or interview element.
- Papers which looked at the appropriateness and accuracy of the participant use of DSS were included

Particular attention was given to papers explicitly mentioning automation bias, automation misuse, or over reliance on automation, or terms such as confirmation bias, automation dependence or automation complacency. Initial research into the subject area revealed that though these effects are recognised, they are not generally considered in the literature in the clinical decision support domain. These effects also tend to go unnoticed with common assessment methods, effects may be lost to averaged results and, if reported at all, are often secondary or implicit findings.

### **3.4.2 Outcome measures**

Studies involving:

- Assessment of user performance: the degree and/or appropriateness of automation use. This included:
  - a) Indicators of DSS advice usage - consistency between DSS advice and decisions, user performance, peripheral behaviours (such as advice verification) or response bias indicators
  - b) Indicators of the influence of automation on decision making, such as pre and post advice decision accuracy (such as negative consultations, whereby pre-advice decision is correct and switched to incorrect post-advice decision), or DSS versus non-DSS decision accuracy (higher risk of incorrect decisions when bad advice is given by DSS, versus control group decisions), correlation between DSS and user accuracy (a relationship between falling DSS accuracy and falling user decision accuracy) such as user sensitivity<sup>7</sup> and specificity<sup>8</sup> with varying DSS accuracy.
- analysis of error types (such as those of commission or omission) and reasons for user error, or ineffective DSS use were included.

### **3.4.3 Exclusion criteria**

Purely technical papers were excluded e.g. detailing the development of CDSS as they omitted much of the human element. Non-english papers were excluded. Empirical studies of information seeking and decision making in group situations was often found, these were excluded.

---

<sup>7</sup> Sensitivity: the measure of correctly identified true positives, higher sensitivity is related to lower False Negative rate

<sup>8</sup> Specificity: the proportion of negatives which are correctly identified, higher specificity is related to lower False Positive rate.

- Did not include the use of DSS<sup>9</sup>
- Ignore purely final outcome based assessments (e.g. patient outcomes) i.e. make no mention of the nature of the automation usage and errors e.g. and false positive and false negative data
- Exclude papers which include solely the accuracy of the DSS and do not incorporate a human behaviour element, for example, papers which compare user performance to automation / DSS performance rather than assessing user performance *with* automation / DSS

Review papers were included in the stage 1 and 2 filtering process for possible insights and citations for other papers, but were excluded from the final selection.

### **3.4.4 Quality assessment**

Once the stage of initial reading of the full article had been carried out, the papers were scored for internal and external validity and for relevance to the review aims. The generic paper quality was scored according to items adapted from the PRISMA CONSORT checklist. The paper relevance was scored according to requirements for the outcome measures and inclusion criteria and structured according to Population, Intervention, Control and Outcome (PICO). Three papers were removed from the final sample because they were deemed not relevant to the final study. The quality assessment was kept fairly flexible - Juni et al (1999)<sup>250</sup> advise against overly rigorous scoring and selecting studies based on a threshold, advocating that papers are still taken on individual merits to avoid skewing results.

See Appendix B.

#### **3.4.4.1 Method for screening and deciding eligibility**

Table to show each stage of the literature retrieval process

1. References retrieved by search strategy (total and by database)
2. Number following visual review of titles and abstracts.
3. Relevant number identified for full text review
4. Final number of useful articles following reading of article

---

<sup>9</sup> Though these should be included for a more general literature review with indicative results – as it involves psychological linkages e.g. workload increases fatigue, but no DSS.

Any immediately obvious duplicate articles were removed at stage 1 and remainders were removed in stage 2 using Endnote and a visual search.

Table 3.2 Stages of the systematic review and numbers of papers found per search engine

<b>Literature Retrieval Process at Each Stage</b>				
	<b>Stage 1</b>	<b>Stage 2</b>	<b>Stage 3</b>	<b>Stage 4</b>
<b>Database</b>	<b>References retrieved by search strategy</b>	<b>Number following visual review of titles and abstracts</b>	<b>Relevant number identified for full text review</b>	<b>Final number of useful articles following assessment</b>
PubMed/MEDLINE	9718	254	54	34
Cinahl	212	26	3	
IEEE	3112	59	12	2
PsycInfo	65	41	19	10
WoS	714	87	31	18
Citation searching				9
Personal correspondence <sup>10</sup>				1
<b>Totals</b>	<b>13821</b>	<b>467</b>	<b>119</b>	<b>74</b>

#### 3.4.4.2 Reliability

The final papers were tested for reliability of extraction using Cohen's kappa statistic. Rater 1 created a sample of 100 paper abstracts and pseudo randomised a number of abstracts which they had rated as "hits" i.e. (varying the strength of apparent relevance to the set criteria - not all "hit" papers were immediately obvious). Rater 2 was given the same inclusion/exclusion criteria and asked to select the relevant articles from the sample. Rater 1 was not informed of the number of "hit" papers included by Rater 1. Rater 1 selected 13 papers to include in the sample. Rater 2 extracted 17 potentially relevant papers from this article including all 13 articles picked by Rater 1 (plus 4 extra articles). The crude rate of agreement was 87%. Cohen's kappa was 0.8436; according to Landis and Koch (1977)<sup>251</sup> (table 3.3), who

<sup>10</sup> Wyatt J. Acorn trial: Lessons learned from the field trial of ACORN, an expert system to advise on chest pain.

formulated a table for assessing Kappa significance; this result implies “almost perfect agreement”.

**Table 3.3: Landis and Koch (1977) table for interpreting k values**

<b>K</b>	<b>Interpretation</b>
< 0	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

### **3.5 Data extraction**

To help clarify the research question, a PICO taxonomy was used. For the results, the results were organised and tabulated according to an extended PICO framework to capture extra information.

- Population: all participants, any demography or background e.g. from naïve to field experts
- Design
- Intervention/exposure:
- Comparison: Groups not using automated decision support, or different forms of decision support (non-automated or automated but a different design), or before and after design
- Outcome measures: assessment of user performance; error types and reasons for user error
- Other relevant information

### 3.5.1 Draft data extraction form (i.e. PICO):

NB. Final version, including studies, is in Appendix C.

Not all of these heading were viable on checking all papers. Only headings which all papers could fill adequately were eventually included.

**Table 3.4: Column headings within the systematic review extraction table**

Year	Title	Author	Journal	Objective	population	intervention	control	outcome	methods	setting	Other info	Active/ passive support	Task supported	User type	User skill level	User age
------	-------	--------	---------	-----------	------------	--------------	---------	---------	---------	---------	------------	-------------------------------	-------------------	-----------	---------------------	----------

User Gender	Computer experience	Incentivised study (Y/N)	Study site	Methods to decrease AB	Rate	Difference	Target users representative	Participant response rate	Quality of gold standard	Study design	Number of citations	Was knowledge used in DSS Evidence based?	Conclusion
-------------	------------------------	-----------------------------	------------	---------------------------------	------	------------	--------------------------------	---------------------------------	--------------------------------	-----------------	---------------------------	---	------------

### 3.6 Findings

From an initial 13821 papers (after removal of duplicates), a total of 74 studies were found (table 3.5) which satisfied the inclusion / exclusion criteria. The concept of AB was first discussed and continues to be most explicitly explored in the aviation field. The main fields of clinical study are based around Computer Aided Detection (CAD) type DSS <sup>e.g.252</sup> followed by ECG use <sup>e.g. 253</sup>. Two studies were found with general diagnostic-based CDSS (e.g. QMR, Iliad)<sup>23,219</sup>. Other DSS include more domain-specific DSS such as ECG reading<sup>253, 254,255</sup>, skin lesions<sup>167</sup>, antibody identification<sup>256</sup>, chest pain<sup>257</sup>.

**Table 3.5 Profile of papers found in systematic review by research field and year of publication**

	Healthcare		Aviation	Generic HCI	Military	Other	Total
	CAD	Other					
1993 - 1996	1	3	6	1	0	0	11
1997 - 2000	1	4	8	0	0	1	14
2001 - 2004	7	3	6	4	1	2	23
2005 - 2009		3	2		1	2	8
2006 - 2009	5	4	5	1	1	2	18
Total	14	16	26	6	3	7	74

## 3.7 Results

### 3.7.1 Automation Bias rate

Human general error rates have also been studied at great length. In multiple contexts, and many scenarios humans have a general error rate of 0.5% to 1.0%. Generally, humans make a mistake up to 1% of the time when performing any given task<sup>258,259</sup>. Ruben et al (2003)<sup>260</sup> recorded error rates in 10 general practices in England; 940 errors were recorded relating to prescriptions, communication, appointments, equipment, clinical care, and “other” errors. The overall error rate was 7.56% of appointments – with 42% (397/940) were related to prescriptions.

The rate of error with an unreliable intervention can be hypothesised to be higher.

AB appears to be a fairly robust and generic effect across research fields. Overall improvement in user performance was found with DSS by most studies, some even when the advice given was inappropriate<sup>256</sup>, though some showed overall decrements to performance<sup>257</sup>.

In terms of outcome measures, errors relating to erroneous DSS output were recognised in terms of negative consultations<sup>23, 219, 255, 261,262</sup>, percentage of erroneous advice cases followed<sup>253, 254, 263, 264</sup>, and more indirect implied measures of AB such as a decrease in accuracy when DSS is inaccurate<sup>157,257, 265,266</sup>, or if there is a correlation between decreasing DSS accuracy, and decreasing user accuracy<sup>229,267, 268</sup>.

CAD studies focussed on AB effects, showing mixed results and distinguishing between errors of commission and omission, in terms of sensitivity<sup>11</sup> and specificity<sup>12</sup> respectively. Four studies showed a decrease in both measures with inaccurate DSS due to AB<sup>252,269,270,271</sup>. Four studies showed contrasting effects on sensitivity and specificity reporting increased sensitivity with a decrease in specificity with CAD<sup>272,273,274,275</sup>. It was found that CAD interventions can decrease specificity reported without decreasing sensitivity<sup>270,274,276</sup> but it has also been found to increase specificity with no effect on sensitivity<sup>277</sup>. Some studies explicitly state that no AB was found despite there being the opportunity for it to emerge<sup>277, 278,279</sup>. Moberg (2001)<sup>277</sup> stated this was mostly due to: False Positive (FP)<sup>13</sup> targets detected with CAD output being

---

<sup>11</sup> Sensitivity: the measure of correctly identified true positives, higher sensitivity is related to lower FN rate

<sup>12</sup> Specificity: the proportion of negatives which are correctly identified, higher specificity is related to lower FP rate.

<sup>13</sup> False Positive: A result that indicates that a given condition is present when in reality it is not.

generally different from those detected by human observers, thus it was relatively easy for observers to disregard FPs; the potential cost of higher automation error being mitigated by pilot strategy, whereby the sounding of an alert led to a closer scrutiny of the raw data.

Dixon (2006)<sup>280</sup> differentiated between inappropriate reliance and compliance; reliance pertaining to behaviour when the automation signals that “all is well”, whereas compliance refers to the human behaviour when the automation signals that action is required.

### *Interruptive DSS in aviation*

Interruptive DSS studies into AB were mainly found in the field of aviation research.

Generally, Skitka et al found commission errors to be higher than omission errors in 2 studies into AB<sup>10,25</sup>, conversely Mosier et al. (1997)<sup>281</sup> found 55% omission rates and 0% commission rates in an aviation study. Many studies did not distinguish this, reporting overall errors only. Studies reporting automation complacency error rates for interruptive systems have been shown to increase if a DSS is highly (but not perfectly) reliable, leading to overtrust and complacency<sup>203,282,283,284</sup> and less reliable (but not highly, obviously unreliable). Lower levels of reliability can paradoxically inspire better performance due to lower complacency levels<sup>284</sup>, for example, Madhavan (2007)<sup>209</sup> sets the optimal threshold at 70% reliability before performance degrades or the DSS is disused.

### **Meta Analysis**

The RevMan program was used to analyse the papers with the highest quality scores and the most homogenous methodologies and outcome measures. The results are shown in fig 3.2.

Four papers in the healthcare field found that participant accuracy decreased with erroneous DSS intervention in comparison with a non-intervention control group. Results from these four papers were pooled in a small, indicative meta-analysis on the basis that they assessed the percentage of erroneous decisions following incorrect advice when given by CDSS compared to a non-CDSS control. The studies had homogeneity in terms of methodology, control group, intervention type and field of study and had high quality scores. The CDSS analysed were non-interruptive in nature and the advice text-based. The studies also analysed commission errors, which are more clearly AB rather than complacency errors. These were included in a Mantel-

Haenszel method Random Effects, Risk Ratio analysis<sup>14</sup>, at the 95% confidence level. Studies are summarised in Appendix C. The Relative Risk, RR = 1.26 (1.11, 1.44), - erroneous advice was more likely to be followed in the CDSS groups than in the control groups and when in error the CDSS increased the risk of making an incorrect decision by 26%<sup>15</sup>. The RevMan program was used to analyse the papers with results shown in figure 3.2.

### **How often do people make errors following bad advice**

#### **Study 1 – Hillson (1995)<sup>254</sup>**

An investigation of the effect of computer assisted interpretation on electrocardiogram reading. Participants evaluated ten clinical vignettes accompanied by ECGs and reported their diagnostic impressions. Half of the subjects received ECGs with computer-generated reports, the other half received the same ECGs without reports. Participants who received erroneous reports were more likely to make an error corresponding to the advice.

#### **Study 2 – Tsai (2003)<sup>253</sup>**

Tsai studied the effect that the computer interpretation (CI) of ECGs had on the accuracy of physicians. Thirty physicians either interpreted an ECG with or without a CI. Participants erroneously agreed with the incorrect CI more often when it was presented with the ECG than when it was not, implying they had been misled.

#### **Study 3 – Southern (2009)<sup>263</sup>**

An examination of the effect computer misinterpretation might have on physicians' ECG interpretation and decision-making. Overall, of 105 erroneous CIs given, the interpretations of the physicians were also incorrect 49 occasions with no CI and 56 occasions with CI; the recommended actions of the CI were agreed with in 49 occasions of no CI versus 56 occasions with CI. The most significant difference in management decision was when 17 / 56 residents with the erroneous CI reading recommended urgent revascularization more frequently than the 5 / 49 residents without the erroneous CI.

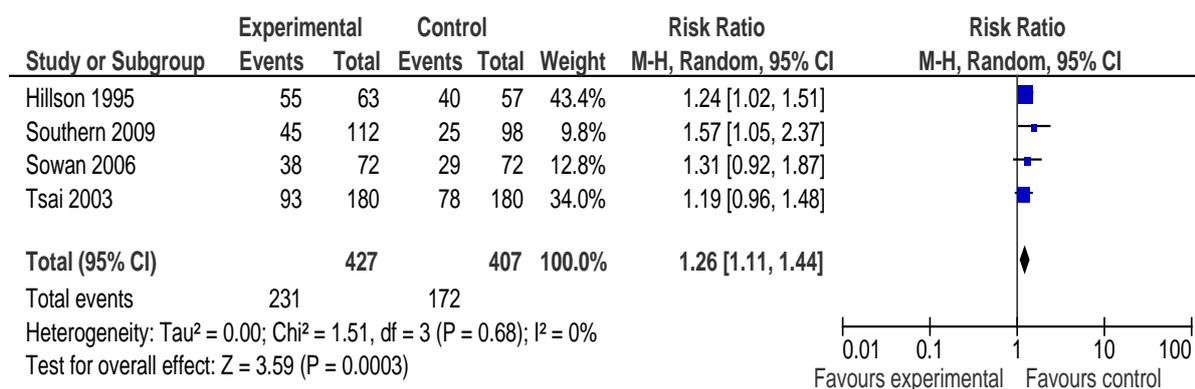
---

<sup>14</sup> The **fixed effects model** assumes that all studies come from a common population, and that the effect size (odds ratio) is not significantly different among the different trials. The **random effects assumption** (made in a random effects model) is that the individual specific effects are uncorrelated with the independent variables.

<sup>15</sup> The overall effect was significant,  $p < 0.0005$ . Tests for heterogeneity were not significant ( $p > 0.05$ ), implying that the variation in underlying intervention effects across studies was not significant.

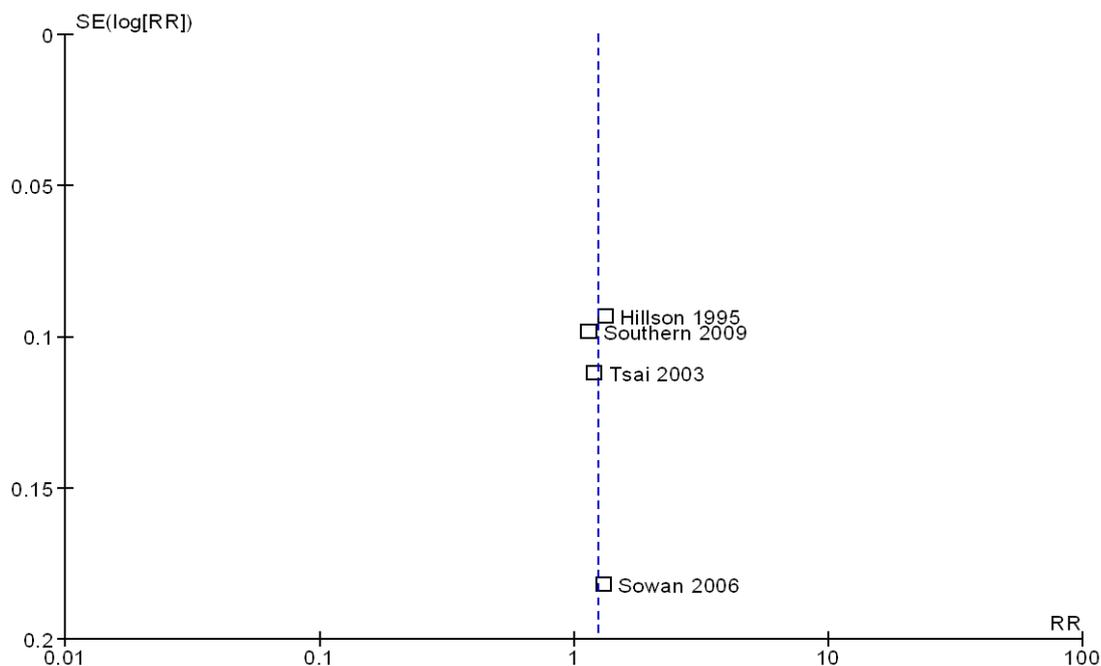
Study 4 – Sowan (2006)<sup>264</sup>

A study to assess healthcare professionals’ ability to detect medication administration errors by comparing CPOE orders with handwritten orders for paediatric continuous drug infusions in a simulated test environment. Despite nurses preferring the CPOE generated orders, they did not decrease the nurse’s ability to detect medication administration errors and nurses were less likely to detect them by CPOE than with manual orders.



**Figure 3.2 RevMan meta-analysis output of four papers showing erroneous advice followed (of total opportunities)**

There were insufficient studies to generate funnel plot assessing for publication bias, however the graph is shown below, fig 3.3, for information:



**Figure 3.3 Funnel plot to assess publication bias**

### **Non – controlled effects – negative consultations**

Negative consultations are the clearest measure of AB; for example, as compared to percentage of incorrect decisions following incorrect advice, which could be conflated by participants having the same incorrect pre-advice answer (thus no AB generated decision “switching” will have occurred, despite being included in the calculation). However the papers reporting it often do so incidentally, thus the effect is not controlled for and cannot be included as part of a meta-analysis. Four studies with similar designs (utilising text-based, non interruptive CDSS) exhibited this outcome. The proportion of decisions which demonstrated this ranged from 6% (Friedman 1999) - 11% (McKibbon 2006) of cases in prospective empirical studies.

Friedman (1999) found positive consultations, where the correct diagnosis was present after consultation but not before, in 12% of cases; negative consultations were observed in 6% of cases. The resultant net gain was 6%. Berner (2003) found that in 21 cases (of 272) the correct unaided answer was switched to an incorrect answer after DSS use; 8% were negative consultations. Westbrook (2005) found that system use resulted in a 21% improvement in clinicians' answers, from 29% correct pre- to 50% post-system use, however in 7% of cases correct pre-test answers were changed incorrectly. A similar study by McKibbon (2006), which examined how clinician-selected electronic information resources improved physicians' answers to simulated clinical questions, found a negative consultation rate of 11%.

### **3.7.2 Causes of Automation Bias**

When AB is reported, it ranges from a significant problem<sup>253,254</sup> which may render DSS not useful on balance (according to the rate and severity of automation error), to a small issue where it is still worthwhile given the benefits<sup>275</sup>. Looking at potential effect modifiers is crucial to understanding the underlying causes of AB.

#### *Experience*

General and DSS-specific *experience* has been shown to affect tendency for overreliance in 8 papers. Papers suggest that task inexperience may lead to automation-related errors in papers which focus on complacency<sup>136</sup>, and AB<sup>285</sup>, however inexperienced users concurrently showed the most overall improvement using DSS<sup>274</sup>. Linked to effects of training, experience may decrease overreliance on automation by different mechanisms; in complacency it may

familiarise users with baseline reliability, and in AB it may highlight the risk of accepting incorrect information, promoting verification of uncertain output. Walsham (2008)<sup>286</sup> showed that despite no apparent improvement in performance, CAD improved the subjective confidence of one less experienced user, whereas it improved only the overall performance of less versus more experienced users; CAD can lead to mismatched decision confidence against actual performance, and is of greater value to users with less task experience. AB occurs more often with task inexperienced users<sup>285,287</sup>, but can occur with more experienced users<sup>274</sup>. Physicians with greater experience may be less reliant on DSS and be more likely to recognise an incorrect piece of advice<sup>23,167,285</sup>. In an experiment looking at reliance on medication management systems, Ho (2005)<sup>157</sup> found that age was a factor in DSS related error, with older participants making more errors; though this may be an indirect relationship mediated by experience. Conversely, in terms of complacency, Bailey (2005)<sup>282</sup> found that specific DSS experience decreased monitoring performance – familiarity led to desensitisation and habituation effects.

### *Confidence and Trust*

Experience may be positively related to *user confidence*. In three papers incorporating multitask experiments, increased confidence<sup>282,284, 288</sup> in users' own decision decreased reliance on external support, whereas trust in the DSS increased reliance<sup>288</sup>. Similarly, Dreiseitl (2005)<sup>167</sup> showed that physicians were more likely to be biased by automation and accept DSS advice when they were less confident in their own diagnosis. Lee and Moray (1996)<sup>193</sup> state that automation reliance is essentially a trade off between self-confidence and trust in the DSS. *Trust*<sup>157,180</sup> and automation reliance is arguably the relationship with the most research in terms of complacent behaviours<sup>282,283,284,289, 290</sup> and automation bias<sup>151,172, 185,209,288</sup>. Trust is possibly the strongest driving factor in overreliance, when trust is incorrectly calibrated against system reliability. This may be a general trend in human judgement, for example, Dzindolet et al (2003)<sup>185</sup> demonstrated that participants had a predisposition to trust, or had a “positivity bias” towards the automated aid over a human one and commit AB error. Higher perceived automation pedigree<sup>209</sup> (for example novice versus expert systems) also affects reliance, increasing trust in the system.

### *Individual differences*

*Individual differences* in reliance have been found e.g. potential for complacency<sup>282,289,291</sup>, and also for predilection for certain decision strategies such as maximisation<sup>172</sup> or non compensatory decision strategies (versus compensatory) which use minimal information on which to base decisions<sup>292</sup>. Underlying personality and cognitive characteristics may make some participants more prone to committing automation based errors in terms of both AB<sup>151,157,172,281,293</sup> and complacency<sup>289,291</sup>. Producing DSS which provide good cognitive fit could decrease AB error rates.

### *Task Type*

The *task type* itself may affect how users rely on external automated advice. More complex tasks and higher workloads are posited to increase reliance<sup>282</sup> by placing stress on cognitive capacity. Users may become biased to overuse automated advice under increased workload<sup>151</sup> or may be prone to automation complacency<sup>280,294,295</sup>. Xu (2007)<sup>279</sup> however found that contrary to this, increased trial difficulty improved performance, suggesting it decreased participant complacency, and led them to inspect the data more closely. Factors which increase external pressures on an individual's cognitive processing capacity may produce a reliance shift towards external support. Prinzel (2005)<sup>289</sup> found that a relationship between perceived workload and complacency error was mediated by users' intrinsic complacency potential. Those with high complacency potential were more likely to report higher perceived workload and have lower monitoring performance. Sarter (2001)<sup>287</sup> suggested that high *time pressure* could bias a user towards DSS usage. Both AB and complacency errors are thought to stem from reallocation of attention<sup>249</sup>; putting pressure on cognitive resources could either bias a user towards heuristically using DSS output, or over-relying on automation to provide correct output so attention can be channelled towards other tasks.

These factors have in common that they place stress on cognitive capacity. As an adaptive measure users tend to then rely on DSS to compensate; if the DSS is reliable this is performance enhancing, if not, it can lead to new errors.

### **3.7.3 Automation Bias avoidance**

#### *Implementation Factors*

The research indicates that certain measures can be implemented to help prevent people over-relying on decision support. One study found that making users aware of the *DSS reasoning*

*process*<sup>185</sup> increases appropriate reliance, reducing AB. Increasing *accountability* for decisions may also prevent AB, however while two studies<sup>35, 293</sup> showed that external manipulations of accountability increased vigilance and thus decreased AB, another study<sup>281</sup> showed that external manipulations did not affect this, but that participants' internal perceptions of accountability did – people who perceived themselves to be accountable made fewer AB errors. Similarly, one study<sup>229</sup> found a positive relationship between DSS misuse and negative attitudes in a workplace and shared social norms; improving the *working culture* may help appropriate DSS use. Papers have assessed the effect of *training* on appropriate DSS use; linked to experience discussed above, training may increase the likelihood of recognising DSS error and thus reducing AB (particularly commission error)<sup>35</sup>. However, Parasuraman et al cite a Mosier et al. (2001)<sup>296</sup> study which implied training had no impact on AB. Complacency error<sup>136, 295, 297</sup> is more clearly reduced by training than AB.

#### *Decision Support System Design*

The *design* of the DSS can affect how participants use advice. To reduce complacency error, adaptive task allocation<sup>298</sup> - varying reliability rather than keeping it constant<sup>21, 28</sup> - was found to increase vigilant behaviour and improved appropriate reliance. The position of advice on the screen can affect the likelihood of AB. Berner et al (2003)<sup>219</sup> found that display prominence increased AB, affecting the likelihood of changing decision after advice – prominent incorrect advice is more likely to be followed. However, Singh (1997)<sup>212</sup> found that while DSS intervention produced more complacent performances compared to a manual control group, centrally (versus peripherally) locating the monitoring task made no difference to this performance. In another study into automation Complacency, Yeh (2001)<sup>290</sup> looked at system factors; too much on-screen detail makes people less conservative, thus increasing biases. This study also found conversely that increasing scene realism appeared to increase conservative decisions. McGuirl (2006)<sup>288</sup> found that updating the confidence level of the DSS alongside pieces of advice (as opposed to providing one overall fixed confidence level for the system) improved the appropriateness of participant reliance, decreasing AB. Sarter (2001)<sup>287</sup> suggested that status displays (versus command type displays) rendered imperfect DSS less causative of AB – while display helps with detecting a problem, command type advice cuts out a step in the decision making process and thus may be prone to overuse under time pressure.

Thus there is evidence that AB can be mitigated by decreasing the prominence of DSS output, but there is no evidence for this in complacency, while complacency can be reduced by adaptive task allocation. AB can be reduced by decreasing onscreen detail, updating advice confidence levels and providing supportive information rather than commanding advice.

### **3.8 Systematic review conclusion**

Though studies do exist which demonstrate the AB effect, there appear to be few definitive and deliberate studies into looking at how inaccurate DSS advice affects the user's decision. For example, the studies found rarely stated the percentage reliability/accuracy of the DSS, which can be assumed to have an effect on the rate of AB error.

There are a number of factors (in terms of user, DSS, task and environmental characteristics) which may directly or indirectly impact a user's tendency to accept inaccurate advice, and ways which this can be mitigated. The primary drivers for AB and complacency may be the user calibration of the trade off between trust and confidence. This is tempered by individual predispositions in terms of cognitive styles and technology acceptance. Task specific and previous DSS experience may act on primary drivers to impact reliance on DSS.

Environmental factors such as task complexity and workload, and time pressure can also place pressure on cognitive resources leading to more heuristic-based use of DSS output; if output is incorrect, this can lead to over-reliance. Methods to mitigate AB include implementation and DSS design factors. Increasing user accountability for decisions and DSS training improves appropriate reliance. Additional information such as up-to-date confidence levels of DSS can improve appropriate reliance, as can design factors such as the position of advice on the screen and mode of advice (for example, information versus recommendation) can affect reliance.

Parasuraman and Manzey (2010)<sup>249</sup> carried out a broad literature overview incorporating theoretical and anecdotal papers, outlining complacency and AB in several research fields. The focus and scope of this review systematically expands on empirical evidence for AB rates, causes and mitigators within the healthcare field.

There are many factors involved in AB and complacency effects which are likely to be interlinked. Even though the nature of AB is not clear, enough studies, discussion papers and anecdotal evidence exists to imply that it is a frequently occurring effect. It is postulated frequently but lacks clear empirical evidence. In the following study, the AB effect will be under study (despite the overlap with complacency), as the CDSS simulation will focus on a

more passive system, where automation errors are unlikely to be due to attentional lapses and likely to be commission based.

### **3.9 Summary**

The systematic review provides further evidence for the existence of AB, factors which affect it and potential ways to prevent it from causing systematic error. The next chapter attempts to draw together the evidence and gaps from both the literature and systematic reviews in Chapters 2 and 3 and use this to generate a testable model with hypotheses to be investigated in an empirical study.

## 4. Discussion and Conclusions of Reviews

### 4.1 Introduction

The literature and systematic reviews found that though some studies do exist which demonstrate the AB effect, AB is poorly defined in the literature and there appear to be very few definitive and deliberate studies into looking at how inaccurate DSS advice affects the user's decision. Where studies do exist, a lack of homogeneity of studies in terms of study domain (for example, the original field of study was aviation, which is a qualitatively different format of DSS (broadly more active and visual), DSS type and experimental design affects the ability to directly compare and contrast rates of overreliance and potential precipitating factors. The AB finding is also generally incidental to other primary aims, though may be available in raw data, it is rarely analysed and reported explicitly. Receiving far more deliberate study in the clinical field has been the subject of disuse of DSS. This can be linked to the confidence and egocentric discounting literature which can inform why users may misuse DSS. Many studies which look at reliance and disuse, and a "bias to distrust"<sup>299</sup> practically and theoretically, but conversely there is a paucity of empirical studies specifically examining at over reliance / automation bias a priori, in particular within the clinical field. Self-reliance may provide participants with an illusion of control. Langer (1983)<sup>300</sup> found that people often behave illogically in order to have an illusion of control.

At the same time, even though the nature of automation bias is not clear, enough studies, discussion papers and anecdotal evidence exist to imply that it is a common effect. It is postulated frequently but lacks empirical evidence.

Despite this, there are many theoretical factors (in terms of user, DSS, task and environmental characteristics) which may directly or indirectly impact a user's tendency to accept inaccurate advice. This study should aim to bridge some of the gaps between the literature review showing many potential factors, and the systematic review indicating a scarcity of strong empirical data for AB, particularly in the healthcare field by providing a deliberate study of this effect.

This research aimed to investigate some of the primarily top down (such as confidence and trust) and contextual factors (time pressure and task complexity) which affect reliance, so we can better design the bottom up factors (which can be manipulated more directly) such as the DSS interface.

#### **4.2 Rate and influencers of Automation Bias**

*Rate of AB:* The systematic review described two outcome measures for AB error (section 3.7.1). The first assessed the chance of the clinician making an incorrect decision with incorrect automated advice versus no advice. The meta analysis found that the presence of automated advice increased the likelihood of an incorrect decision by 26%. The second outcome measure involved negative switching, or switching from a previously correct opinion to an incorrect one on the basis of incorrect automated advice; this was found in four papers and ranged from 6 – 11% of cases (this is compared with the rate of negative switching as found in this study in section 6.6). A caveat here is that very few of the healthcare field studies explicitly stated the reliability of the CDSS under study (it may be assumed that this will affect the rate of AB related error); this study will be one of the first to explicitly control and state the reliability of the CDSS.

*Influencing factors generated by literature reviews:* The systematic review (section 3.7.2) indicated that the user's attitudinal factors affected automation reliance, with the strongest evidence coming from the most studies appearing to be more direct attitudinal factors of trust, and self-efficacy (or self confidence). Trust in the DSS and self- confidence may be in a trade off; it may be found that users display higher trust in relation to lower confidence are more likely to over-rely on the DSS.

The non-attitudinal user factor of experience was also implied as a factor in reliance behaviour, which included both task and DSS experience. These would provide users with a mental context on which to base their judgement, which may allow for more appropriate calibration of reliance. It may be that users with higher task experience rely less on DSS advice, and possible that they have a more complete knowledge base with which to compare the DSS advice for using or dismissing (this may interact with self-confidence with more task experience giving users more confidence in their decisions).

DSS experience relating to the effect of DSS training (which has strong evidence for improving appropriate reliance) may work similarly, by giving users a better context for judgement of the accuracy and validity of the DSS advice, and thus better ability to accept correct and recognise and reject incorrect advice.

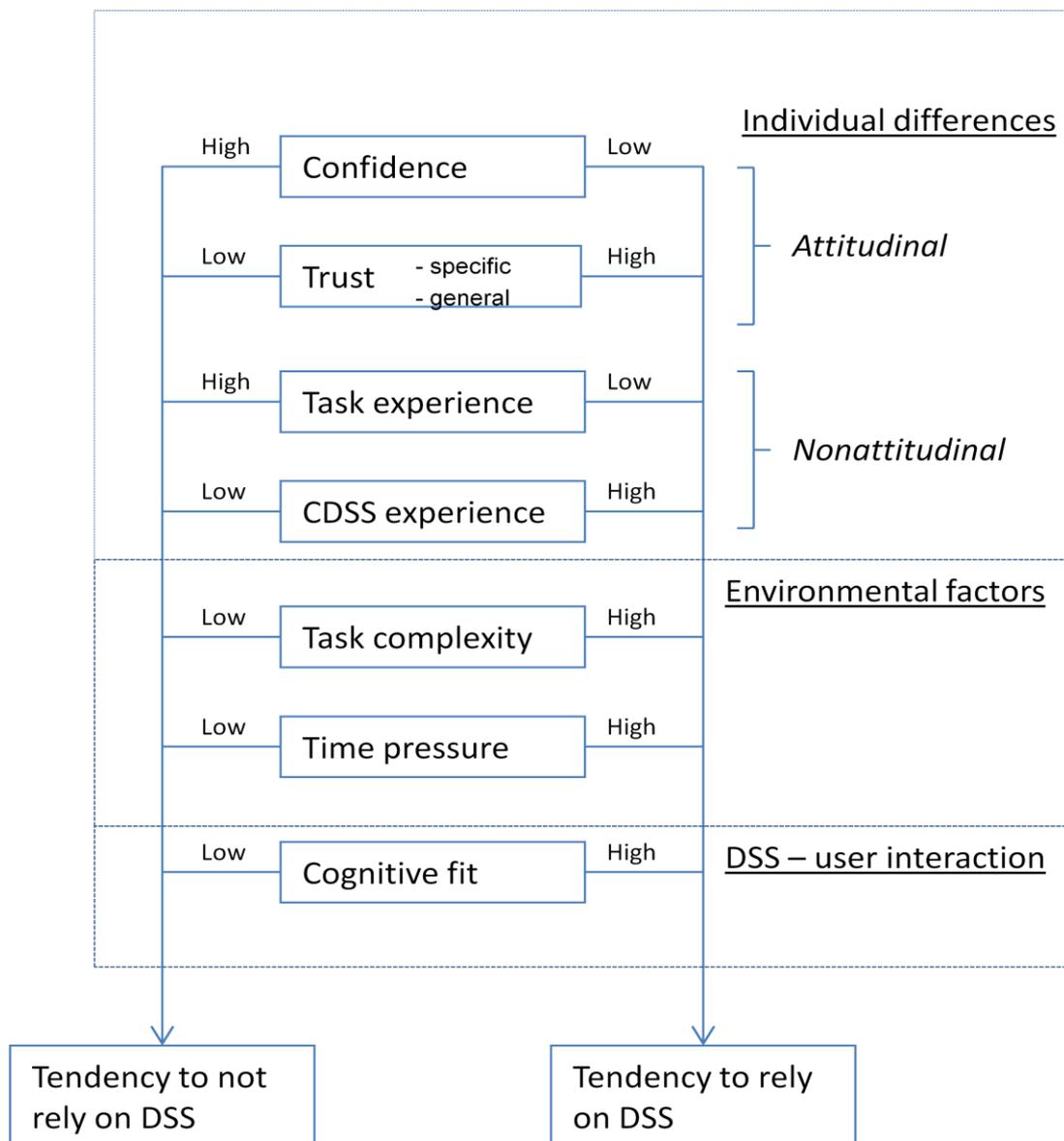
The strong task and environmental factors of time pressure and task difficulty may affect reliance; these were posited to constrain judgement and decision making and increase satisficing behaviour and heuristic use. Thus if the DSS advice is being used as a “rule of thumb” for the correct answer, increasing both these factors may produce an increase in AB.

Cognitive and decision style e.g. compensatory versus non-compensatory decision styles are also theorised to have a strong underlying effect. The interaction between task, cognitive style and DSS – “cognitive fit”, is likely to affect the degree to which the user trusts the DSS and its perceived usability. However, this is a more complex factor to measure and manipulate, and while included in the following model, is beyond the scope of this thesis.

Numerous models of general reliance on technology were found to help explain the phenomenon of overreliance. A model which appears to fit many of the factors found in the literature and systematic review is the TTD theory (fig 2.4). Authors of the TTD model highlights primary factors found in the literature review which lead to overreliance on automation, but the authors also state that it is probable other factors exist. This research model attempts to use and adapt this theory as a basis to augment with the additional stronger evidence in the literature reviews, and use this to test factors involved. The main drivers towards overreliance found in the systematic review have been highlighted above and are illustrated in fig 4.1.

### 4.3 Automation Bias influencing research factors

The theoretical model with the closest fit to the aim of this study, overreliance on DSS, is the TTD model (in terms of salient factors and the focus on overreliance)<sup>301</sup>. This was utilised and augmented with the findings from the literature and systematic reviews to formulate a preliminary simple research diagram (fig 4.1) from which testable hypotheses were drawn:



Factors are assumed to be independent but interactive.

**Figure 4.1** Diagram of factors to include in study and possible directions of relationships

#### 4.4 Diagram hypotheses

- Confidence – Higher confidence leads to less reliance on DSS advice and thus lower propensity for AB error.
- Trust - Higher trust leads to more reliance on DSS advice and thus higher propensity for AB error. This is potentially the case for either general levels of trust in CDSS, or a specific level of trust in a CDSS. This may be exacerbated if there is concurrent low confidence.
- Task experience – Higher experience relates to people being less likely to rely on DSS or make AB errors
- DSS experience, for example, training or frequency of use of DSS - Higher familiarity (measured by previous frequency of CDS use) means people may be more likely to rely on DSS but not necessarily to make AB error i.e. should lead to more appropriate reliance.
- Task difficulty - Higher difficulty relates to people being more likely to rely on DSS and make AB errors
- Time pressure - Higher time pressure relates to people being more likely to rely on DSS and make AB errors
- Cognitive fit – Higher cognitive fit will lead to greater and more appropriate reliance. Better cognitive fit should decrease misuse of DSS advice (*this is a complex effect which is beyond the scope of this investigation; it may be a matter for further study*).

It is likely that these factors are related or interact e.g. there may be a suggested trade off between confidence and trust. Some factors may be compensatory e.g. DSS familiarity and the tendency for more appropriate reliance could be mediated by low confidence leading to higher AB errors; others may be non-compensatory e.g. trust may overrule all other factors if sufficiently high, to lead to higher AB errors. It is also possible that *trust* and *confidence* are ultimate drivers for overreliance (while other factors impact these).

Participants were not presented with the rationale for asking for their stated measures of the suggested influencers in this study.

#### 4.5 Study design

The research hypotheses were initially to be tested using a 2x2 mixed factor study<sup>16</sup>, with the manipulated factors being time pressure (time pressure (approx 20-30 seconds per decision, subject to pilot) vs no time pressure, between subjects), and task difficulty (high versus medium, within subjects).

Other study designs were discussed to investigate the AB effect, for example examining AB in “real” situations using existing CDSS. This could have been carried out by monitoring AB error in real systems, possibly by capturing prescribing screens and then relating them to correct/incorrect answers. In the early stages of the study, preliminary discussions with clinicians associated with the Centre for Health Informatics, and developers of the ISABEL system (online paediatric diagnostic CDSS) including consultant clinicians; this yielded preliminary interest in collaboration with the study but difficulty in follow up. One of the suggested reasons for this was that at this stage (pre full literature and systematic review) there was little robust empirical evidence to justify the implementing the study within a “real” environment and the potential disruption to real life prescribing. It is likely a number of GP surgeries would need to be monitored to get adequate numbers of GP participants and possible numbers of occasions where incorrect advice is given and an AB error could potentially occur (seeing as there is little information available about overall reliability of these systems, this factor is random in terms of predicting the size of the sample required for any meaningful results), which again increases potential difficulty of recruitment. Also the possible technical expertise required to capture screens within a pre-existing prescribing/CDSS system was not at the time available to the study.

A controlled study with a simulator allowed other variables posited by the literature to influence reliability to be manipulated, such as task difficulty and time pressure. It was also much more feasible to record factors such as trust and confidence.

The trade off between the increased control granted by a simulated study versus a field study is that there is a lower level of ecological validity. With a controlled study, it is possible to

---

<sup>16</sup> The resulting low response rate to the study (as described in section 5.10) meant that to attempt to maintain study power, the between subjects factor of time pressure had to be dropped. This finally resulted in a one factor within subjects study (medium and hard task difficulty).

control factors such as task difficulty and system reliability and much easier to record non-controlled variables such as trust and per-case decision confidence.

As discussed in the systematic review, the outcome measure which best demonstrates AB error is negative switching. The current study therefore followed a similar before-after controlled design as these studies for in the systematic review, which also tended to be simulated studies.

The system reliability for the following study was set at 70%. The literature and systematic reviews revealed that (to the researcher's knowledge) no healthcare studies demonstrating an AB effect stated the reliability of the CDSS under study. This issue was also discussed with Health Informatics experts and researchers also looking at the AB effect within a secondary hospital setting, and the same conclusion was reached through personal correspondence. The rate was set at 70% as one particular aviation-field paper was found which explicitly recorded DSS reliability and noted AB type effects (Madhavan, 2007)<sup>209</sup>. This paper suggested that 70% reliability was the threshold at which any lower reliability actually enhanced user performance (due to increased vigilance and awareness of bad advice), negating AB errors. This study was the first seen by these researchers, looking at healthcare CDSS, which explicitly sets and states the reliability of the (simulated) CDSS. The validity of textual case simulations has previously been demonstrated in medical education exercises<sup>302,303</sup>, and during the assessment of mock clinical decision making<sup>304</sup>.

#### **4.6 Summary**

This chapter has discussed the rates and influencing factors for AB which can be tested in an empirical study.

This study aimed to provide some empirical support for the presence of AB error in primary care prescribing when the CDSS is not wholly reliable, and for its influencing factors. As a result of this the researchers hope that which the increased deliberate evidence base, further studies will be more possible, for example field studies suggested in this section looking AB errors in "real" situations.

The following chapter describes the empirical study employed to test the rate of AB in the context of primary care prescribing, with a CDSS simulator set at a 70% reliability level, and the hypotheses outlined in section 4.4.

## **5. Empirical Study**

### **5.1 Introduction**

Following the exploration of the literature, an experimental study was carried out to investigate the rate of AB when a CDSS is implemented to aid primary care prescribing. This was also carried out to investigate some of the strongest factors affecting reliance found within the literature and systematic reviews of the evidence, using a simulated CDSS: User self-confidence and trust in the automation, task difficulty and environmental pressure (in this case time pressure), experience (both in terms of CDSS and clinical).

### **5.2 Setting**

Primary data were gathered via a web-based CDSS simulation. As previously mentioned, the validity of textual case simulations has previously been demonstrated in medical education exercises, and during the assessment of mock clinical decision making (section 4.5).

The CDSS simulator was designed to be put online, so participants were able to access the study remotely.

### **5.3 Participants**

#### **5.3.1 Sample size calculations**

Estimating the required sample size can be one of the most important aspects of the recruitment process. Before a study is designed, it is crucial to calculate or make an informed estimation of the sample size necessary to show a significant result. However, there are very few studies which are directly applicable to this experiment, thus a crude sample estimation was carried out with the caveat that it may not be truly representative of the actual sample required. The sample size is calculated below, but was formatively adjusted during data collection.

The following methodology was verified through personal correspondence with a senior academic statistician. The sample size was calculated for a 2x2 mixed factors study (between subjects = time pressure, within subjects = task difficulty). A meta-analysis which was carried out on randomised controlled trials comparing number of bad pieces of advice followed whether there was automated decision support present (intervention condition) versus none (control condition), found that significantly more bad pieces of advice were followed with the

presence of decision support (fig 3.2). The effect size for this was calculated by using the Cohen's d calculation. Cohen's  $d = 0.596$ . This is classed as a "medium" effect size (Cohen 1988)<sup>305</sup>. Using the G\*Power programme, and assuming a power size of at least 80% as adequate, the project should aim for 72 participants (36 per group). If a power of 95% is required, 124 participants will be required (62 per group).

Other sources cite much lower rules of thumb for samples in quantitative data<sup>17</sup>.

- Quantitative studies
  - 30 participants for correlational research
  - 15 participants in each group for experimental research
  - Approximately 250 responses for survey research

### **Sample size required**

#### **Step 1: Calculate effect size**

A small indicative meta-analysis was carried out on randomised controlled trials comparing number of bad pieces of advice followed whether there was automated decision support present (intervention condition) versus none (control condition), found that significantly more bad pieces of advice were followed with the presence of decision support. See output in fig 3.2, in the systematic review.

#### **Step 2: Standard deviation was calculated from weighted means of percentages (working below):**

Experimental group =  $(55+45+38+93) / (63+112+72+180) = 231/427 = 54.1\%$  (SD = 20.8%)  
Control group =  $(40+25+29+78) / (57+98+72+180) = 172/407 = 42.3\%$  (SD = 18.8%)

#### **Step 3: Effect size was calculated**

**Effect size (between subjects) =** Input data provided:

Mean 1: 54.10

SD 1: 20.80

Mean 2: 42.30

SD 2: 18.80

Output:

*Cohen's d: 0.596*

---

<sup>17</sup> E.g. Participant, Subjects and Sampling. [cited September, 2011] Presentation available at: [http://people.uncw.edu/caropresoe/EDN523/523\\_Spring\\_08\\_Spring\\_09/McM\\_Ch5-Rv.ppt](http://people.uncw.edu/caropresoe/EDN523/523_Spring_08_Spring_09/McM_Ch5-Rv.ppt)

This is based on the average SD from two means.

**Step 4: G\*Power program was used to calculate sample size needed at Powers of 0.80 and 0.95**

**t tests - Means: Difference between two independent means (two groups)**

<b>Analysis:</b>	A priori: Compute required sample size		
<b>Input:</b>	Tail(s)	=	One
	Effect size d	=	0.596
	$\alpha$ err prob	=	0.05
	<b>Power (1-<math>\beta</math> err prob)</b>	=	<b>0.80</b>
	Allocation ratio N2/N1	=	1
<b>Output:</b>	Noncentrality parameter $\delta$	=	2.5286138
	Critical t	=	1.6669145
	Df	=	70
	<b>Sample size group 1</b>	=	<b>36</b>
	<b>Sample size group 2</b>	=	<b>36</b>
	<b>Total sample size</b>	=	<b>72</b>
	Actual power	=	0.8048843

**t tests - Means: Difference between two independent means (two groups)**

<b>Analysis:</b>	A priori: Compute required sample size		
<b>Input:</b>	Tail(s)	=	One
	Effect size d	=	0.596
	$\alpha$ err prob	=	0.05
	<b>Power (1-<math>\beta</math> err prob)</b>	=	<b>0.95</b>
	Allocation ratio N2/N1	=	1
<b>Output:</b>	Noncentrality parameter $\delta$	=	3.3183876
	Critical t	=	1.6574395
	Df	=	122
	<b>Sample size group 1</b>	=	<b>62</b>
	<b>Sample size group 2</b>	=	<b>62</b>
	<b>Total sample size</b>	=	<b>124</b>
	Actual power	=	0.9510426

The effect of negative consultations was also considered, however due to this outcome not having a control condition in previous studies, a meta-analysis of this was not possible. However, comparable studies which represent the outcome of negative consultations corroborated this number e.g. Westbrook (2005)<sup>261</sup> found a rate of negative switching of 7% using 75 participants (and 557 completed cases) in a repeated measures design (though this was not an a priori primary outcome measure, thus the required sample size may have been lower), thus 72 participants should cover this aspect. Berner (2003)<sup>219</sup> found 70 participants (and 272 completed cases) elicited an 8% negative consultation rate. McKibbon (2006)<sup>262</sup> needed fewer participants; 26 participants carried out 46 cases each – 11% of which were negative consultations.

Time pressure effects on heuristic use and advice taking (advice may be seen as a heuristic) are consistently shown to be a "large" effect e.g. Rieskamp (2008)<sup>306</sup> found an effect size of  $\eta^2 = 0.56$ ; for this measure of effect size Small = 0.01; medium = 0.06; large = 0.14 (Kittler, Menard & Phillips, 2007<sup>307</sup>). Thus theoretically, fewer participants were needed to demonstrate this effect with statistical significance, and 72 participants should have covered this.

Task complexity also tends to have "large" effects on heuristic use e.g. Gino and Moore 2007<sup>308</sup> found that task difficulty had a significant effect on advice taking heuristic use:  $F(1,23)=6.193$ ,  $p=0.021$ ,  $\eta^2=0.212$ . Thus again, the participant number should have been adequate to provide sufficient power to the experiment to produce meaningful results.

## 5.4 Design

### Aims:

- a) To assess the influence of a simulated prescribing decision support system on clinician prescribing performance. Particularly we were interested in the negative impact of CDSS: will clinicians follow incorrect advice; is there negative switching
- b) To examine the impact of manipulated task difficulty and time pressure on CDSS usage
- c) To examine clinician characteristics related to prescribing performance with DSS: attitudinal factors of trust in CDSS, confidence in decisions, and non-attitudinal factors of clinical and DSS experience (non manipulated variables)

**Rationale:** According to the heuristics literature, satisficing behaviour increases in the context of environmental pressure (e.g. time constraints, task difficulty), so decisions were to be analysed under these constraints to see if automation bias is increased. These factors were to be manipulated within a 2x2 multifactorial study. User factors are also posited to influence the propensity to take CDSS advice, including attitudinal factors (trust, confidence), and non-attitudinal factors (task and DSS experience).

**Hypotheses:** a) Rate of AB – Hypothesis: physicians will show some bias towards accepting incorrect DSS advice.

b) Causes of AB – Hypothesis: certain factors will affect propensity for AB e.g. time pressure, task difficulty

**Factors under investigation:**

**Design:** 2 x 2 part manipulated mixed/multifactorial experiment, before-after design, following the typical JAS paradigm found in the psychology literature.

*Independently manipulated variables:*

- Task difficulty – two within subjects levels: medium and difficult (within subjects)<sup>18</sup>;
- Time pressure – two between subjects levels: time pressure versus no time pressure (between subjects)
- Control: “Before” condition for the effect of intervention, the medium difficulty and no time pressure as primary control comparator

*Non-manipulated variables:*

- Confidence: Measured alongside every prescribing decision by asking for the GP for their decision confidence level. Measured on 6 point Likert scale (highest confidence- lowest confidence)
- Trust: Participants were be asked for their trust in CDSS generally before taking part in the study, and asked for their trust in the simulated CDSS after the experiment. Both factors were measured on 6 point Likert scale (completely trust – completely distrust)
- Participant self-reported estimation of their CDSS and clinical experience:
  - Clinical experience: Estimated number of years [freetext]
  - DSS experience will be assessed by asking for self reported frequency of DSS use: 6 point Likert scale (very frequently – never)

N.B. Likert scales: According to King and Epstein a rating scale can be as reliable as a ranking scale<sup>309</sup>; to measure perceived importance of generated values a questionnaire was created using Likert-type scaling.

To prevent the middle point effect being interpreted as a neutral point, a six item scale was generated as per recommendation from Fowler (2003)<sup>310</sup>.

---

<sup>18</sup> An “easy” condition could be omitted, as it is assumed participants would find it too simple to detect bad advice and may make participants suspicious of other advice

**Primary outcome measures:**

- Decision switching: changes in decisions from correct pre-advice, to incorrect post-advice, indicates a negative consultation and an AB error.
- Decision performance was assessed under various experimental conditions (e.g. high versus low trust).

**Secondary outcome measures:**

- The relationships between time pressure, task difficulty and decision switching and AB
- Other factors (e.g. trust, confidence, DSS and clinical experience) were tested for their relationship to AB, though it must be noted that these were not controlled for.

**5.5 Procedure**

Participants were told that they could assume the preliminary diagnosis was correct.

Participants were invited to participate in the study by email (see Appendix D).

On viewing the information sheet and consent form (Appendices E and F) an initial example of a scenario was given to give participants some familiarisation in the accuracy of the DSS advice.

Participants were asked for age (in 5 year bands), gender, clinical experience (estimated number of years) and CDSS experience (frequency recorded on a 6 point Likert scale), and general trust in DSS (recorded on a 6 Point Likert scale), see section 5.4.

The participants were asked to view 20 hypothetical primary care medical scenarios (10 labelled “hard” and 10 labelled “medium” difficulty). Half the participants were to view the 20 scenarios under time pressure; the other half with no time pressure (allocation was to be randomised). Section 5.9 describes how the 20 final scenarios were developed.

Fig 5.1. illustrates the procedure for each scenario presented to a participant. Participants provided their initial prescription in freetext format<sup>19</sup>. Alongside this they were asked for their confidence in the decision (according to a 6 point Likert scale; 1 Lowest confidence to 6 Highest confidence). Following this, they were be given a piece of advice from a simulated

---

<sup>19</sup> Freetext so clinician is not constrained by fixed choices – more ecologically valid

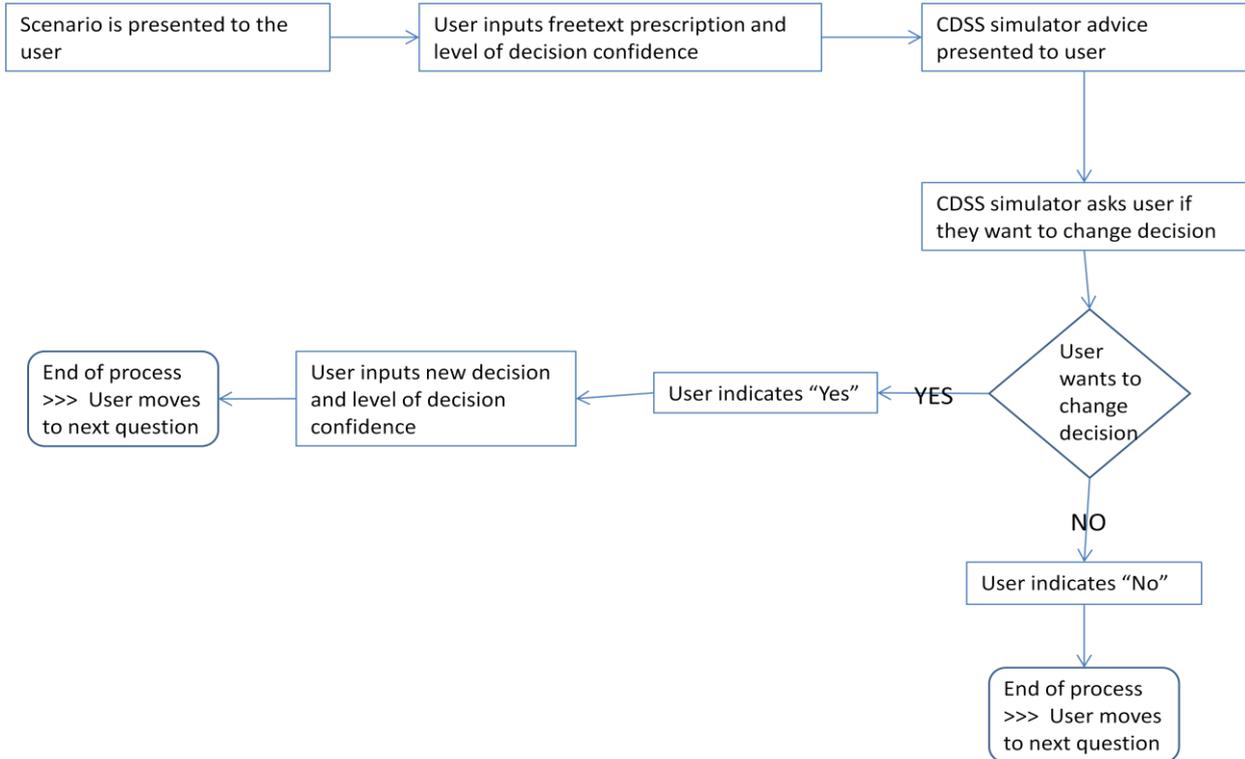
decision support system. Of the 20 scenarios, 3 scenarios in the “hard” condition were accompanied by incorrect advice, and 3 scenarios in the “medium” condition were accompanied by incorrect advice (i.e. overall there were 14 scenarios with correct advice and 6 with incorrect advice) – rate of **70%** correct.

After viewing the advice they were asked if they wanted to change their prescriptions. If they did, then another freetext box was presented in which to type this. They were also asked for their confidence in this decision.

Following the 20 scenarios, they were asked for their trust in this specific decision support system (using the same 6 point Likert scale as at the start of the study for general trust in CDSS), and 3 qualitative opentext questions:

- a. How reliable did you find the advice?
- b. Did you always follow it?
- c. What are important factors in the design of DSS to you?

Participants were then debriefed by the simulator and by an email (Appendix G).



**Figure 5.1 Test procedure for each case (20 cases presented)**

## **5.6 Resources**

A computer scientist was employed to develop simulated web-based DSS and data capture forms.

A mailing list company was commissioned to find GP participants.

Incentivisation for participants included a prize draw for a 32GB iPod Touch and £100 donation to a charity of choice.

Healthcare professionals were approached to help develop a scenario database i.e. provide a professional opinion to validate scenarios, and eventually validate GP sample answers.

## **5.7 Clinical Decision Support System simulation**

The CDSS used in this study was simulated, to allow for manipulating different facets (e.g. interface variables, output accuracy) of the system according to the experimental aim.

Development of the simulation was formative, with advice coming from experts in the programming and healthcare field providing opinions during pilot phases of the study.

This simulation was based on decision support for Primary Care prescribing tasks. Projects such as the Microsoft Common User Interface (CUI) / Connecting for Health for the NHS CUI Programme<sup>311</sup>, as utilised in Scott et al (2011)<sup>12</sup>, were investigated for ideas for development of this simulator. Also investigated was the interface for the ScriptSwitch system<sup>20</sup>. The interfaces had in common that they were simple, and provided a suggestion for drug, dose, and frequency. This was aimed at for the CDSS simulator (see section 5.7.1.1).

### **5.7.1 Simulator**

An aim of this study was to build a CDSS simulator rather than a real tool for prescribing decision making. In this way the system's comments or advice could be manipulated as necessary, for example according to difficulty of the patient scenarios, and according to the time pressure condition. The aim, as mentioned, was to study the effect of incorrect CDSS advice on users' decision making and to measure how bad advice can affect physicians' performance. In this way the system's accuracy could also be manipulated by the researchers.

The CDSS simulator was developed alongside a database which stored the final 20 clinical scenarios, each with a corresponding correct and incorrect piece of prescribing advice. The scenarios were programmed to be presented in a random order accompanied by the correct or

---

<sup>20</sup> ScriptSwitch online demonstrations at: <http://www.scriptswitch.com/see-demos.html>

incorrect advice (scenarios accompanied by incorrect advice were also randomised). The rate of incorrect advice could be manipulated by the researcher (in this case 30% of the scenarios were accompanied by incorrect advice). Users were asked whether they agreed to the advice or not and their answers were recorded into the database.

The CDSS simulator tool was a simple, dynamic web based application. The scenarios were stored in a relational database (i.e. multiple tables with multiple relations between them), which was developed in MSSSQL Express Edition. All users were provided with a username and a password to login to the system. The aim was to make the application as simple as possible, showing just the patients' scenarios and advice to the users i.e. not presenting further pictures or graphs etc.

The CDSS Simulator was developed in a Microsoft .NET Framework, an environment for developing both web based and desktop applications, which can use Visual Basic.Net or C#. This simulator was developed as a web based programme in Visual Basic.Net.

A "middle tier" was developed between the web forms and the database to improve performance and security (by letting web forms access the database indirectly through it). For all data manipulation functions, appropriate Stored Procedures (SP) were built and the middle layer summoned those stored procedures only, for querying and manipulating data.

The simulator was hosted at a virtual machine server at the university (CHIVM).

### 5.7.1.1 Screenshots from the simulator

Example scenario:



**School of Informatics**  
CITY UNIVERSITY LONDON

**Clinical Scenario:**

*Question number 2*

A 32 year old male complains of recurring nosebleeds over the past couple of months; they occur almost daily and can last for around 20 minutes. It is currently not bleeding. He mentions he is asthmatic and also has an allergy to peanuts.

Preliminary diagnosis: Epistaxis

What would you prescribe? [if you wouldn't prescribe medication on this occasion, please indicate this in the text box below]

**Type your initial answer below:**

**How confident are you with your answer**

Lowest Confidence  1  2  3  4  5  6 Highest Confidence

**Save & show me the system advice**

Advice page:



**School of Informatics**  
CITY UNIVERSITY LONDON

**System's Advice**

- Naseptin, four times a day, 10 days, supply 15g (NHS cost £1.90)

**Would you like to change your initial answers?**

No  
 Yes

Second answer page:

School of Informatics  
CITY UNIVERSITY LONDON

Type your second answer below:

How confident are you with your answer?

Lowest Confidence  1  2  3  4  5  6 Highest Confidence

SAVE & Continue

## 5.8 Ethics

Ethical approval for this study was gained from the City University Research Ethics Committee. Ethical approval was also necessary from any NHS bodies approached. The study was granted Ethical Approval by the London – Bentham Research Ethics Committee.

*City Ethics* – gained 17<sup>th</sup> Feb 2011, no revisions (see Appendix H).

*NHS Ethics* – full approval granted 5<sup>th</sup> September 2011 after one set of revisions outlined in below (see Appendix I).

The structure of the survey was amended to include:

- a. A disclaimer in the information page explaining that there may be some incorrect answers. This was to mitigate prescribers finishing the study with the impression that the incorrect pieces of advice given purposefully were correct.

- b. Provision of measures was required during the course of the study itself to ensure that misinterpretation of incorrect information was not an undesirable outcome of participating.

As a result the study was amended to take place in one sitting (to prevent the participant pausing having seen an incorrect answer). Previously an option to login and out again was included.

As soon as the participant logged in, a read-receipt email was sent automatically, to debrief them (to ensure that participants who did not complete the study were still debriefed full). In the Consent Form, the participant agreed to check the email immediately after the experiment. If the participant completed the experiment they saw the list of “incorrect” scenarios they saw alongside the incorrect and correct answers. If the participant did not complete the experiment, the email asked them clearly to reply to receive the scenarios they were given with the incorrect answers.

- c. The study had to include a provision that participants would have to view the summary and undergo debriefing as an essential part of the study. The participants were debriefed after the study and asked to indicate (by tick box) that they had read it. The debrief also outlined all the incorrect scenarios and their incorrect responses, alongside examples of validated correct answers.

Participants were also directed to the systematic review paper (Goddard et al. (2011)<sup>14</sup>) as a further source of information for the basis of the study.

In addition further details about the statistical analyses planned, and the scenario validation were required, and were provided satisfactorily.

## **5.9 Scenarios**

A literature search and correspondence with experts in the Health Informatics, Prescribing and Healthcare fields revealed that there was no pre-existing bank of prescribing scenarios with “gold standard” answers – a small bank of primary care prescribing scenarios had to be generated and validated by the researcher.

### **5.9.1 Scenario generation**

The researcher generated 35 scenarios based on internet searches of anonymised real-life cases as a basis for the scenarios. These were then heavily modified to further alter the demographics

and history of the patient, and the prevailing specific symptoms, whilst maintaining the preliminary diagnosis. The 35 scenarios were then validated as described in the following section; twenty scenarios which were deemed to have the most consensus in terms of difficulty rating and agreement on the correct and incorrect prescribing options were then selected.

Certain parameters needed to be considered in the generation of the scenarios. Factors were later validated by clinical professionals.

- 1) Disease prevalence – Ideally all scenarios would have featured similar prevalences. However most illnesses are fairly common or at least well known. By default some of the harder scenarios developed are also rarer disorders. It had to be taken into account that AB may be as a result of a lack of previous knowledge, rather than a more positive draw towards the automated advice (i.e. some conflation of the two is possible).
- 2) Scenario length – variations in the lengths of scenarios could confound the results by adding a fatigue effect for the more lengthy scenarios. Paragraph lengths were kept fairly brief (1 or 2 short paragraphs) and the wording was uncomplicated. Some differences in length were inevitable, particularly as more difficult cases can involve more variables to take into account.
- 3) Difficulty of scenarios – the aim was to find a number of scenarios which would not elicit floor or ceiling effects, but which had some variation to compare performance between scenario difficulties.
- 4) Type of scenario - more acute cases that needed less complex management were developed. More chronic conditions could have more complex treatments and management.

The format and broad content of the scenarios was adapted from WHO guidelines<sup>21</sup> i.e. demographic information and the fact that the preliminary diagnosis can be taken as the correct one. These are shown in Box 2.

Sources of information: Clinical Knowledge Summaries (CKS) and the BNF were the primary sources for prescribing information as it is possibly globally the most respected source of prescribing information<sup>22</sup>. BMJ Best Practice was also used as a source.

**Box 2: Factors taken into consideration:**

**Demographics**

Age  
Gender  
Ethnicity

**Lifestyle**

Occupation  
Habits e.g. smoking, alcohol  
Allergies  
Pregnancy  
Other vaccinations, drug use

**Medical history**

Date – Illness - Treatment

**Presenting symptoms**

Short paragraph to describe scenario

These then had to be validated by people with experience and qualifications in the prescribing domain.

### 5.9.2 Scenario validation

Thirty-five acute, primary care scenarios were assembled by the researcher.

#### *Stage 1*

A preliminary sense check of 35 scenarios by a physician and a pharmacist was carried out. All scenarios passed this stage as deemed sensible. No account was taken of more specific information such as the judged difficulty level.

#### *Stage 2*

A more specific round of validation was carried out aiming to gain consensus over the validity of the scenarios and the “correct” and “incorrect” answers, and also attempt to quantify the level of difficulty. The 35 scenarios were validated using 3 questions:

1. How difficult is the task of prescribing in this instance?  
1 – Very Easy

<sup>21</sup> Teacher's Guide to Good Prescribing: <http://hinfo.humaninfo.ro/gsd/healthtechdocs/documents/s15940e/s15940e.pdf>

<sup>22</sup> E.g. [http://www.bma.org.uk/patients\\_public/Youmedicinesusefulsourcesofinformation.jsp#.T04wrfUmSSo](http://www.bma.org.uk/patients_public/Youmedicinesusefulsourcesofinformation.jsp#.T04wrfUmSSo)

- 2 – Easy
- 3 – Somewhat Easy
- 4 – Somewhat difficult
- 5 – Difficult
- 6 – Very difficult

2. In your opinion, what is the “gold standard” correct answer?
3. Any other improvement suggestions or comments about the scenario and the answers.

Two GPs and one pharmacist were asked for their opinions.

A statistical analysis for reliability of answers was carried out for perceived stated difficulty levels of the scenarios (table 5.1).

To carry out tests for inter-rater reliability, a Spearman’s correlation was carried out in SPSS for all 35 scenarios (this is an appropriate test for ordinal over interval data and more than two raters).

**Table 5.1 Correlation of difficulty ratings between 3 raters over 35 scenarios**

			Correlations		
			Pharma	GP1	GP2
Spearman's rho	Pharma	Correlation Coefficient	1.000	.241	.311
		Sig. (2-tailed)	.	.163	.069
		N	35	35	35
	GP1	Correlation Coefficient	.241	1.000	.174
		Sig. (2-tailed)	.163	.	.317
		N	35	35	35
	GP2	Correlation Coefficient	.311	.174	1.000
		Sig. (2-tailed)	.069	.317	.
		N	35	35	35

\*. Correlation is significant at the 0.05 level (2-tailed).

There was non-significant correlation between the 3 raters in this instance. To improve the inter-rater reliability, a Spearman’s correlation test for ordinal data was carried out for the 20 most apparently correlating scenarios chosen (experts deemed most valid and most consensus in terms of correct/incorrect prescriptions and difficulty rating), table 5.2.

**Table 5.2 Correlation of difficulty ratings between 3 raters over 20 scenarios**

Correlations			Pharma	GP1	GP2
Spearman's rho	Pharma	Correlation Coefficient	1.000	.440	.561*
		Sig. (2-tailed)	.	.052	.010
		N	20	20	20
	GP1	Correlation Coefficient	.440	1.000	.452*
		Sig. (2-tailed)	.052	.	.046
		N	20	20	20
	GP2	Correlation Coefficient	.561*	.452*	1.000
		Sig. (2-tailed)	.010	.046	.
		N	20	20	20

\*. Correlation is significant at the 0.05 level (2-tailed).

The correlation improved to significant (or marginal), implying that these scenarios had more consensus. To further increase the validity of the remaining 20 scenarios, a 3<sup>rd</sup> round of validation was carried out.

### Stage 3

Stage 3 included the final check after editing, and removal of scenarios which were deemed outliers (too easy or difficult), lacked consensus over the treatment types. Some less valid scenarios were also removed due to too many cases of contraindications due to pregnancies or allergies (i.e. the aim was to be able to generalise these results as much as possible).

The twenty remaining scenarios were validated by 2 NHS GPs.

A Spearman's rho correlation (between 2 raters) was carried out, see table 5.3. A significant correlation was found.

**Table 5.3 Correlation of difficulty ratings between 2 raters over 20 scenarios**

Correlations			GP1	GP2
Spearman's rho	GP1	Correlation Coefficient	1.000	.487*
		Sig. (2-tailed)	.	.029
		N	20	20
	GP2	Correlation Coefficient	.487*	1.000
		Sig. (2-tailed)	.029	.
		N	20	20

\*. Correlation is significant at the 0.05 level (2-tailed).

These 2 Stage 3 validators had overall consensus with the Stage 2 validators, see table 5.4. Non significant agreements occur between the pharmacist rater and GPs 2 and 3.

**Table 5.4 Correlation of difficulty ratings over all 5 raters over final 20 scenarios**

Correlations			GP1	GP2	GP3	GP4	Pharma
Spearman's rho	GP1	Correlation Coefficient	1.000	.487*	.486*	.647**	.673**
		Sig. (2-tailed)	.	.029	.030	.002	.001
		N	20	20	20	20	20
	GP2	Correlation Coefficient	.487*	1.000	.538*	.543*	.357
		Sig. (2-tailed)	.029	.	.014	.013	.122
		N	20	20	20	20	20
	GP3	Correlation Coefficient	.486*	.538*	1.000	.440	.411
		Sig. (2-tailed)	.030	.014	.	.052	.071
		N	20	20	20	20	20
	GP4	Correlation Coefficient	.647**	.543*	.440	1.000	.548*
		Sig. (2-tailed)	.002	.013	.052	.	.012
		N	20	20	20	20	20
	Pharma	Correlation Coefficient	.673**	.357	.411	.548*	1.000
		Sig. (2-tailed)	.001	.122	.071	.012	.
		N	20	20	20	20	20

\*. Correlation is significant at the 0.05 level (2-tailed).

\*\* . Correlation is significant at the 0.01 level (2-tailed).

### 5.9.3 Scenario presentation

An example of the final 20 scenarios is shown in Table 5.5 below (for all scenarios, see Appendix J). This scenario outlines how the scenario and scenario answers were categorised:

**Table 5.5 Example scenarios with difficulty and sample correct and incorrect answers**

<b>Scenario</b>	<b>Difficulty</b>	<b>Correct answers</b>	<b>Incorrect answers</b>
<b>Septic olecranon bursitis</b> - A 23 year old man present with pain at olecranon and down posterior arm which started 6 weeks ago after a fall in which the patient banged their elbow. He feels mild intermittent and increasing pain. The patient has begun to experience a mild fever with chilling and some sweating. On inspecting the skin around the olecranon, there is redness and swelling, the patient reports tenderness. Aspiration of the bursa reveals a leukocyte count above 100,000/mL.	Medium	- Flucloxacillin, 500mg, 1 capsule 4 times a day, supply 28 capsules (NHS cost £3.21) - Plus appropriate analgesic	- Ibuprofen, 400mg, 3-4 times a day, supply 84 tablets (NHS cost £1.72) - Diclofenac sodium, 25mg, 3 times a day, supply 84 tablets (NHS cost £1.14) - Naproxen, 250mg, 2 times a day, supply 56 tablets (NHS cost £2.70)

*Randomisation:* All 20 scenarios were presented to each participant. The presentation of the scenarios was randomised to prevent the occurrence of fatigue effects. The advice given for each scenario was also randomised with respect to being correct or incorrect.

The rate was set that for each participant 3 randomly chosen “Medium” difficulty scenarios would be presented with incorrect advice, and 3 randomly chosen “Hard” difficulty scenarios would be presented with incorrect advice. Thus 6 of 20 scenarios were presented with incorrect advice.

### 5.10 Pilot study

The study required piloting to validate the study protocol, contents and simulator, and to develop it further on the basis of any recommendations.

The study was evaluated by sending 6 people including 2 Health Informatics field experts, 2 clinicians, 1 pharmacist and one academic peer (who had not carried out the scenario validation) the link to the study. They were asked to carry out the study in one sitting, recording spontaneous thoughts they had about the design and content whilst they carried out the experiment.

Following this, they were also sent a short list of open-ended questions to gauge opinion on some key issues:

### **Questions:**

1. Is the overall difficulty of the prescribing scenarios not too hard or too easy? (to make sure there are no floor or ceiling effects)
2. Timer: approximately how many seconds are needed to view each scenario and give a prescription?
3. Approximately how long does it take to go through the study overall (if finished)?
4. Are the instructions clear?
5. Did you notice any/many incorrect pieces of advice?
6. Is the advice appropriate? How could this be improved?

### **Results**

By question:

1. Overall people felt that, though there was apparent variation in the difficulty of the scenarios, there were no scenarios that felt overly simple or impossible. Some users stated that they did still feel the urge to check other resources, such as the BNF.
2. Most people stated that they would need approximately 30 seconds per scenario. This fits within the approx. 30 minutes already stated as estimated time to complete the study.
3. Users corroborated the timing estimation; all believed they had completed the study within 20-30 minutes.
4. Some commentary was made that the instructions could be more brief and to the point. It was noted that the study would benefit from explicitly stating the variables required for a prescription (namely drug, strength of dose, and frequency). Also it needed to be made more explicit that if they felt that no drug treatment was necessary, or that they should refer to secondary care, they could state this. The wording was altered to include the more holistic “management” of the complaint.
5. Some participants noticed that they were at least unsure about incorrect pieces of advice. Interestingly, it did not tend to be a binary correct or incorrect judgement in

most cases; it appeared more that they felt that the incorrect answers may have been correct, but had diminished confidence/trust in the advice (this however was not strongly reflected in the confidence ratings in the results from the later experiment). One stated “I felt quite strongly influenced by the machine advice in terms of level of detail/amount of info, irrespective of the content”

6. The consensus view was that the advice was appropriate (notwithstanding the lowered trust in the incorrect advice).

#### *Other observations*

Some users were concerned that they were unable to see the advice again once they had moved to the next page. This feature was included to mitigate a second look effect (albeit at the potential expense of ecological validity). For the sake of comparison of different variables (such as task difficulty and time pressure) without an added uncontrolled conflating value of a second look bias, this was not allowed. This is however, also discussed as a limitation in the discussion section.

On occasions when more than one drug option was given in the advice section, one user noted that it should be made clear that these were alternatives rather than conjunctive treatments.

The pull of AB was acknowledged:

“Definitely I like the feeling that I was 'persuaded' on the depression question to move from fluoxetine to duloxetine (the wrong advice) – a very real sense of falling for the automation bias!”

### **5.11 Participant recruitment**

Recruitment of participants for any study aims 1) to recruit sufficient respondents to provide enough power to generate meaningful results, and 2) be representative of the group in question to aid validity and generalisability.

For reasons discussed in section 2.3.2, participant recruitment was limited to UK NHS GPs. Involving other prescribers (e.g. nurses) or nationalities may have confounded results. Due to the anticipated difficulty in recruiting from this participant sample, there was no limitation placed in terms of gender, age, experience level, but these were recorded to describe the sample. The results recruitment process is further described in section 6.3.

Sample size was calculated in section 5.3.1. An inadequate sample size could lead to not detecting differences which exist and could lead to wasteful studies. Results may show that there is no difference between groups or association between variables, where in reality there is one (a Type II error). A sample size that is too large can lead to unnecessary expenditure of time, effort and money.

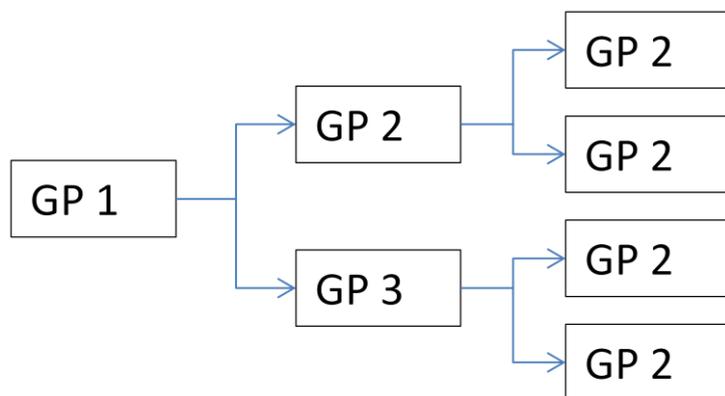
A caveat in this study is that because there are so few studies into this subject, the calculated sample size is used as guidance, but is by no means the definitive requirement – the empirical experiment will be somewhat exploratory.

#### **Types of sampling**

There are various methods by which sampling could be carried out. In terms of generalisability, GPs are known to be a group which has a low response rate in terms of participating in research. Obtaining a fully representative sample was anticipated to be a challenge; the opportunistic nature of sampling meant that a degree bias may have been encountered.

This research utilised opportunistic volunteer sampling both asking personal contacts to participate in the study, and also contacting participants via an email through a healthcare mailing company (Appendix D). The GP participants were primarily gathered via non-probability quota sampling; the probability of selecting a participant is unknown, and subjects are non-randomly chosen from the GP subgroup of the population.

The main specific type of non-probability sampling method was Exponential Non-Discriminative Snowball Sampling (or Chain referral process) illustrated in Figure 5.2.



**Figure 5.2 Exponential discriminative snowball sampling**

For example Avery, 2007<sup>312</sup> used this method to invite participants to interviews concerning improving general practice computer systems to enhance patient safety. In the debrief email participants were asked to forward the email to debrief participants on the study on to 5 GP colleagues who they might have felt would be interested.

Personal correspondence with the Research Governance coordinator at South West London Primary Care Trusts (SWL PCTs) and the Mailing List company response rate estimations imply that the expected response rate for this sample could be between 5-8% or lower. The actual approximate response rate is later compared to this estimation (see section 6.3).

### **Methods of improving response rates**

To assess methods by which overcoming the recruitment challenges associated with Web-based research could be overcome Gordon et al (2006)<sup>313</sup> carried out a study which evaluated several methods. The recruitment channels were (a) Thematic promotional "releases" to print and broadcast media, (b) Google ads, (c) placement of a link on other Web sites, (d) limited purchase of paid advertising, (e) direct mailings.

Self-reports revealed that of 2533 eligible respondents at least half (50.6%) of participants were recruited via mailings, 34.6% from Google ads or via search engines or links on another Web site, and 14.8% from all other methods combined. As part of the recruitment process, this research used e-shot mailing as a method of recruitment as one of the routes for recruitment.

The difficulties in recruiting this particular sample of the population are well acknowledged. A number of experimental modifications have been posited as methods that have evidence for improving response rates<sup>314</sup>.

The following methods were deemed appropriate for this study and were implemented in the recruitment drive:

- 1) Incentivising participation with both a personal prize of an iPod, and a charity prize of £100.
- 2) Simplifying the introductory brief for the study
- 3) Allowing self-registration
- 4) Clarifying the approximate length of the study
- 5) Contacting respondents mid-week (Mondays and Fridays tend to be busier)

### **5.11.1 Stages of recruitment**

#### **Stage 1**

Personal contacts were asked to circulate the invitation email, including :

- The Primary Healthcare Specialist Group within the British Computer Society. The mailing list of the GP network was made available and was used to email to send a call for study volunteers.
- A contact at the Scottish Clinical Information in Management group (SCIMP) circulated the email to advertise the study.
- The Clinical Innovation and Research Centre (CIRC) at the Royal College of General Practitioners (RCGP) showed preliminary interest in this study, but did not follow up.
- Dr Mike Bainbridge, the Clinical Architect for the NHS CUI project, offered to contact people on his mailing list (over 50 people).

The response rate was extremely low – with only 1 respondent for the study.

#### **Experiment Design Alteration:**

Due to the extremely low sample size in Stage 1, the Time Pressure element, which required the design to have 2 between subjects groups, was removed. This allowed the remaining resources to be channelled into testing the remaining factors without the time pressure element.

At this point it was noted that it may facilitate participation if the GP could self register, rather than approach the researcher. To help improve the response rate, this was altered for the next stage of recruitment.

### **Stage 2**

A mailing list company was employed to send the invitation email to 3000 UK NHS GPs. The e-shots were split over 2 weeks – 1500 sent in week 1, and 1500 in week 2.

### **Stage 3**

Four GPs (who were contacts of the Centre for Health Informatics at City University) were contacted and asked to circulate the invitation to colleagues.

Response rates were estimated – see section 6.3.

## **5.12 Summary**

This chapter described the process of developing the empirical study. A potential sample size requirement was calculated, with the caveat that, due to heterogeneity of previous studies and relative paucity of direct evidence this was an approximation.

Study tools were described. The development of the prescribing scenarios which were presented to participants was outlined, as was the simulator through which it was presented to participants. The ethical application process was discussed, including required changes to the design protocol.

The piloting of the study generated suggestions for improvements in study design and validation of scenario difficulty levels i.e. no obvious outliers being too easy or difficult, and timing i.e. study took about 20-30 minutes overall.

Once ethical approval had been gained and the study was piloted, the recruitment process began. The difficulties in recruiting from the GP group were demonstrated and the resulting alterations to the study design were described.

The following chapter describes and analyses the results that were generated from this study.

## **6. Results**

### **6.1 Introduction**

In this chapter the results of the study described in Chapter 5 are laid out. The coding process and validation of the responses gained in the study are outlined.

The overall response rate is estimated. The demographic profile of the resulting participants is presented, alongside their overall stated trust in CDSS and frequency of CDSS use.

The purpose of this chapter is to present an overall rate of AB error as shown by negative consultations, and take CDSS advice to switch decisions in this specific situation. The impact of the factors outlined in Chapters 4 and 5 is investigated.

### **6.2 Answer coding**

Answers were given in an open ended fashion for the sake of ecological validity, and also to not constrain the clinician by forcing them to choose between pre-determined options. With prescribing, there are grey areas with the appropriateness of prescription, which can range from “gold standard” to fatally dangerous. Most prescriptions and advice given during the course of the study which were deemed “incorrect”, were inappropriate due to contraindications, incorrect dosage, or prescribing the wrong drug for the diagnosis.

The sources referred to were the CKS web resource, and the BNF. As a third reference tool the BMJ Best Practice resource was used, though less emphasis was placed here, as it was not NHS/UK specific.

#### **Coding validation**

Due to the necessity of open ended answers (versus forced choice), coding of answers into correct and incorrect was necessary (as discussed this can fall along a spectrum, but for the purpose of this study needed categorising). This can include a degree of subjectivity, thus validation from field experts was required.

Consistency of marking was checked by filtering each case and ensuring that answers were marked as correct or incorrect consistently on a case-by-case basis (there is variation and argument over “correct” prescriptions, even with the validation, there may be some variation and thus borderline cases). Here, the CKS, BNF and were consulted for the “best” answers, and

in some cases a comparison between the appropriateness of drugs for a condition, as they all may be beneficial, but with different levels of efficacy and cost. The importance of context and relativity was noted here; the important factor was the direction of correctness in pre- and post-advice answers as opposed to absolute “gold standard”.

When a GP stated they would not prescribe medication, unless a brief and acceptable management method was outlined, it was marked as incorrect, as the CKS and BNF recommended some form of treatment in all but one of the 20 final cases (not for carpal tunnel syndrome). If correct medication was given without a dosage, on balance it was decided that this could be marked as “correct” or “incorrect” dependent on context. The GP, if unsure of this (and not including the dosage was seen and treated as uncertainty of the dosage), could look up the correct dosage later (if the user changed the answer to include the correct dosage, then the previous answer was marked as “incorrect” to illustrate the positive direction of the change). Similarly, if no dosage was given pre-advice, and then incorrect dosage advice was followed, then the pre-advice condition was marked as correct, as the automation had driven an incorrect answer. If there was no change, then the pre-advice answer was marked as correct as default, again, as the correct dosage could be looked up (assuming this was correct). Thus coding was somewhat context dependent, and the important factor was the change in response, over the actual “correct”/gold standard answer.

A random sample of the answers were taken and given to 3 domain experts who had not seen the previous experiment (1 pharmacist and 2 hospital clinicians), who coded the participants’ answers as correct or incorrect. The coding needed to agree with the researcher and with the other experts. The general reasoning for the researcher’s coding (e.g. the pertinent facts) was given alongside the scenarios with prevalidated “correct” and “incorrect” answers, without the actual researcher’s coding, to aid the validation. Twenty cases were randomly selected and given to the experts. For ease of calculation the correspondence between raters for only the “before” decision was required (though the participants’ answers to both before and after was given, to give context, as mentioned coding was somewhat context dependent).

Cohen’s kappa is usually used for categorical data reliability between 2 raters; for multiple raters Fleiss’ kappa is used to test reliability. Fleiss’ kappa is used with binary or nominal-scale ratings.

Agreement can be thought of as follows, if a fixed number of people assign numerical ratings to a number of items then the kappa will give a measure for how consistent the ratings are.

From Wikipedia<sup>23</sup>, kappa,  $\kappa$ , can be defined as,

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

The factor  $1 - \bar{P}_e$  gives the degree of agreement that is attainable above chance, and,  $\bar{P} - \bar{P}_e$  gives the degree of agreement actually achieved above chance. If the raters are in complete agreement then  $\kappa = 1$ . If there is no agreement among the raters (other than what would be expected by chance) then  $\kappa \leq 0$ .

Fleiss' kappa uses this principle but applies it to multiple raters. There was 85% crude agreement between raters; using an online calculator Fleiss' kappa was calculated at 0.7, which Landis and Koch describe as “substantial agreement” (see table 3.3).

### 6.3 Response rate

Stage 1 of the recruitment process elicited a very low response rate. An unknown number of people were contacted via contacts at SCIMP and PHSG. Only 1 respondent was gleaned from this.

The second stage of recruitment employed a mailing list agency. They contacted a total of 3005 GPs in 2 waves, with c.1500 in each wave, 1 week apart. Table 6.1 shows the overall response rate, 7 days after the invitation emails were sent. This resulted in 23 respondents.

**Table 6.1: Response rate (opening and displaying email, and clicking to website)**

Total Emails Sent	Displays	Display Rate	Click-Throughs	Click-Through Rate (of those displayed)
3005	119	3.96%	<b>23</b>	19.33%

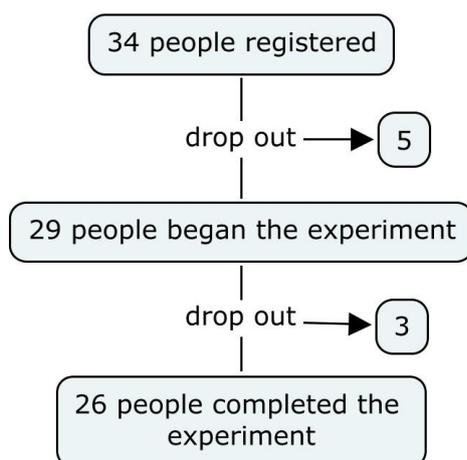
In the third wave, again, 4 GP contacts were asked to circulate the email invitation to members of their practice and any other peers. Again, an unknown number of potential respondents was contacted, but the estimated number of respondents from this phase was around 10 people (it is not precise, as we cannot tell if GPs contacted in phase 2 accessed the study after the 7 day

<sup>23</sup> Fleiss' Kappa [cited January 2012] Available from: [http://en.wikipedia.org/wiki/Fleiss%27\\_kappa](http://en.wikipedia.org/wiki/Fleiss%27_kappa)

limit, or participants were recruited by chain referral), as the final number of people registered was 34.

After dropouts, and partial completions the study recruited **26 full respondents** (see fig 6.1).

The overall response rate was below 5% of people contacted if people who did not display the invitation email (in stage 2) are included i.e. 3005 people were sent an email (see table 6.1.), but only 119 of the 3005 people actually displayed the email, and eventually 23 of these people clicked to see the study website. If the people who did not display the email are excluded then the response rate for the 2<sup>nd</sup> wave of recruitment was 19.3% i.e. depending on the definition of the denominator included, the response rate is either over or below the estimate of 5-8%. This nevertheless low rate will pose some limitations e.g. the representativeness of the sample, which will be taken into account in the discussion.



**Figure 6.1** Flow chart of recruitment dropout rate

#### **6.4 Data preparation**

1. Missing data: Usually substitution or deletion is required when data is missing. All participants who fully completed the experiment were included in the analysis. Three participants dropped out very soon after beginning to complete the scenarios, and were not included in the analysis. The study was designed so that to move through the scenarios, users had to complete the responses, so there were no missing data.

2. Outliers: The data were primarily analysed by nonparametric statistics (parametric statistics were used for tests of interval value measures, some of which had adjustments for low sample sizes, or uneven groups, and tests for homogeneity variance). Due to the low response rate and

relatively small sample size we could not consistently assume normality of distribution or homogeneity of variance. Nonparametric statistics reduce data to an ordinal rank, which reduces the impact or leverage of outliers.

### 6.5 Descriptive statistics of participants

The methods of analysis were validated and approved by 2 professional statisticians. Data were analysed using IBM SPSS statistics v19, MS Excel 2007, and the Vassarstats website for online statistics calculators<sup>24</sup>.

All respondents were UK based NHS GPs. Most respondents fell into the 46-50 years age band; 14 males and 12 females responded (see fig 6.2).

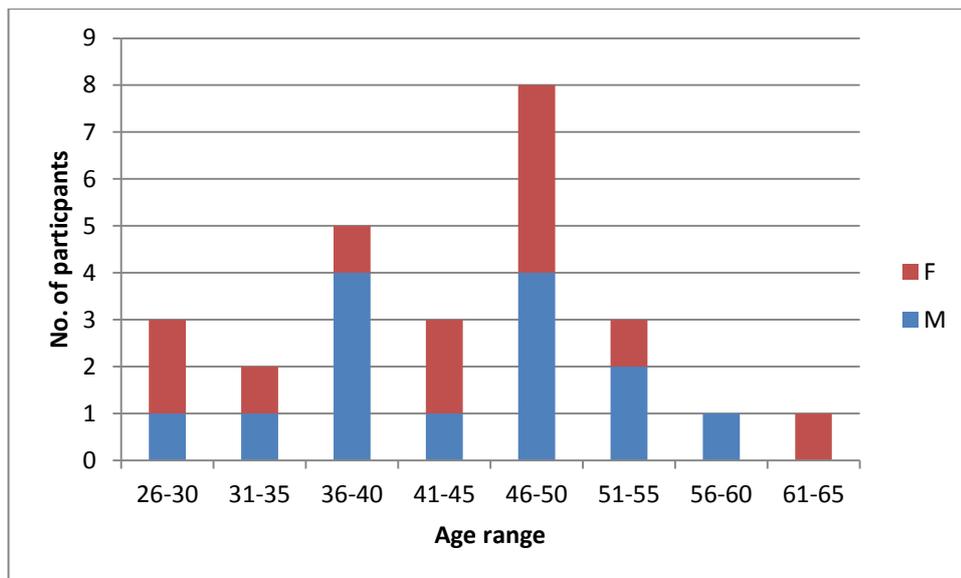


Figure 6.2 Age range and gender of participants

### Clinical Experience

The average stated years of clinical experience was 16, but this was highly variable; SD 10.9 years, range was 2 - 40 years.

<sup>24</sup> VassarStats: Website for Statistical Computation [cited May-Jul 2012] Available at: <http://vassarstats.net/>

### Frequency of Clinical Decision Support System use

The frequency distribution of participant-stated frequency of use is shown in fig 6.3.

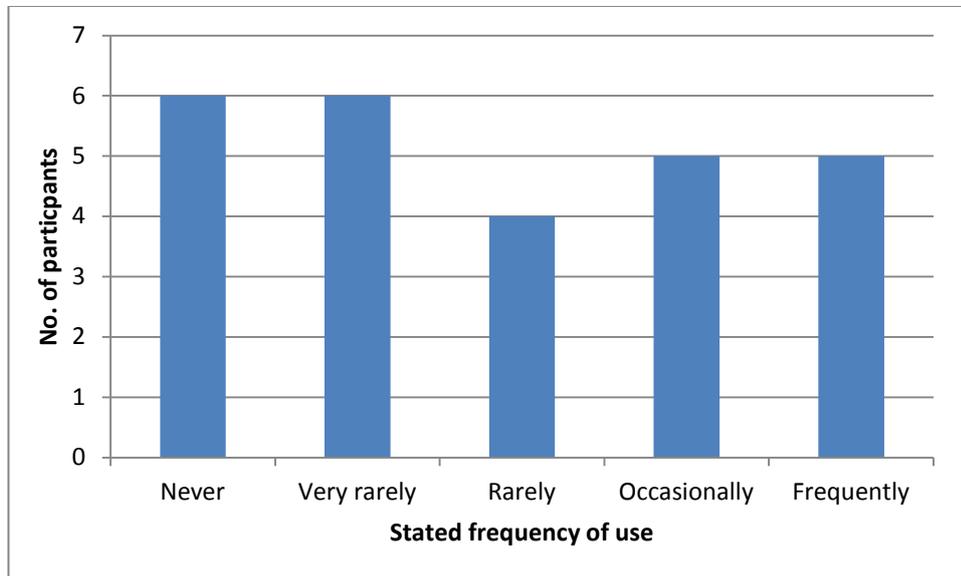


Figure 6.3 Participant-stated frequency of use

### General trust in Clinical Decision Support Systems

The frequency distribution of participant-stated general trust in CDSS is shown in fig 6.4.

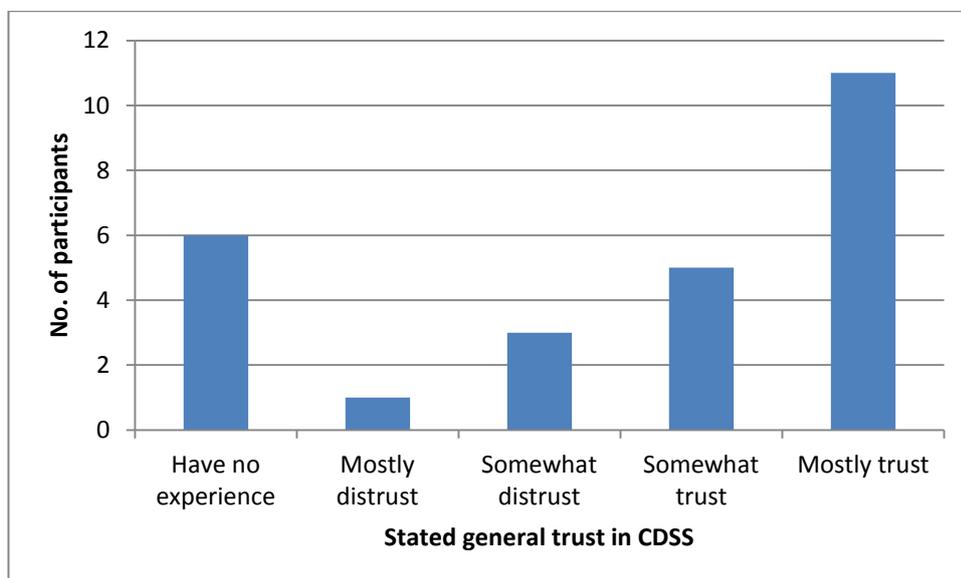


Figure 6.4 Participant-stated general trust in Clinical Decision Support Systems

## 6.6 Results

Primarily nonparametric testing was utilised, as homogeneity of variance (for comparisons) and a normal distribution could not be consistently assumed. In cases where parametric tests were used homogeneity of variance and normal distribution of data were tested and confirmed.

The number of correct pre- and post-test answers was calculated. Broadly, the following tests were applied: Wilcoxon and Mann-Whitney U tests were used to assess significance of differences between paired and independent groups respectively. The Kruskal Wallis test was used when there was testing over more than two groups and groups were independent. Spearman's rho was used to test correlations between variables. A two-way ANOVA tested the effect and interaction of trust and confidence on decision switching.

NB. Overall statistics (i.e.  $N = 520$ ) have been included at the top level to use as a comparison with literature review papers which also present similar overall statistics. However, this type of analysis does not take into account within participant variation. These overall statistics must therefore be taken with the caveat that these results may not be generalisable, but they demonstrate AB and its influencers in this specific situation. To take this into account, tests were carried out by participant ( $N = 26$ ).

Twenty-six participants completed 20 randomised scenarios each; 520 prescribing instances with DSS simulator advice were presented at a set accuracy rate of 70% (364 cases were presented with correct advice, 156 presented with incorrect advice). Overall decisions were switched in 22.5% (117/520) cases<sup>25</sup>.

### 6.6.1 Proportion of correct advice

Overall the DSS intervention improved accuracy from 262/520 (50.4%) correct before advice to 303/520 (58.3%) correct after advice; Wilcoxon test for matched comparisons was significant,  $z = -3.44$ ,  $p < 0.0005$ , suggesting that there was a significant improvement in number of correct decisions after advice.

---

<sup>25</sup> Noting that a decision switches can be Right to Right, Wrong to Right, Right to Wrong, or Wrong to Wrong

To examine changes in the direction of answers pre- and post-test, scenario answers were categorized using the classification of the 4 possible answer situations when incorrect advice is presented, is shown below (as in Westbrook, 2005)<sup>261</sup>:

1. Wrong Wrong (WW): Wrong answer before DSS advice and wrong answer after system use [system did not help]
2. Wrong Right (WR): Wrong answer before but right answer after [system helped]
3. Right Wrong (RW): Right answer before but wrong after [system leads to error]
4. Right Right (RR): Right answer before and right after use [system possibly helped to confirm answer]

These are displayed in contingency table 6.2.

**Table 6.2 Changes in scenario prescriptions pre and post CDSS simulated advice**

Scenario answers			
Before advice	After advice	%	Total number
Wrong	Wrong	36.5%	190
Wrong	Right	13.1%	68
Right	Wrong	5.2%	27
Right	Right	45.2%	235
		100%	520

There were 117 answer switches overall. Answers were switched from correct to incorrect 27 times; there were 156 opportunities (incorrect pieces of advice given) to commit a RW error; 17.3% of incorrect advice was followed.

A binomial test was used to see if participants were taking incorrect advice more or less than chance; at test value 0.3 (30% chance choosing incorrect advice vs 70% correct advice), participants switched from correct to incorrect 28.4% (27/95; 95 being the total number of RW or WR answer switches) of the time, which is not significantly different from chance,  $p > 0.05$ . Concurrently answers were switched from incorrect to correct 71.6% (68/95) times, again, not different to chance  $p > 0.05$  (at test value 0.7). This implies there was not an overall greater distinction than chance of correct and incorrect advice.

The primary test for overreliance was if the physician switched from a correct to an incorrect decision following a piece of incorrect advice (so called “negative consultation” – as demonstrated by RW answers). The RW rate here is 5.2% (27/520) - i.e. in 5.2% of 520 cases a correct prescription was switched were changed incorrectly.

Fifteen of the 26 participants made an RW error. Twenty-four of 26 participants changed some of their initial prescriptions after seeing advice; Figure 6.5 illustrates decision switching frequency by participants and decision type. In 13.1% of cases, the prescription accuracy was improved. Thus there was a net improvement of 8% in user accuracy after DSS simulator use. When advice was correct, decisions were more likely to be switched to a correct prescription,  $z=-4$ ,  $N=26$ ,  $p = 0.0001$ .

When advice was incorrect, decisions were more likely to be switched to an incorrect prescription,  $z=2.1$ ,  $N=26$ ,  $p < 0.05$ .

Spearman’s rho demonstrated there was a significant correlation between decision switching and RW error,  $R_s=0.69$ ,  $N=26$ ,  $p<0.0005$ .

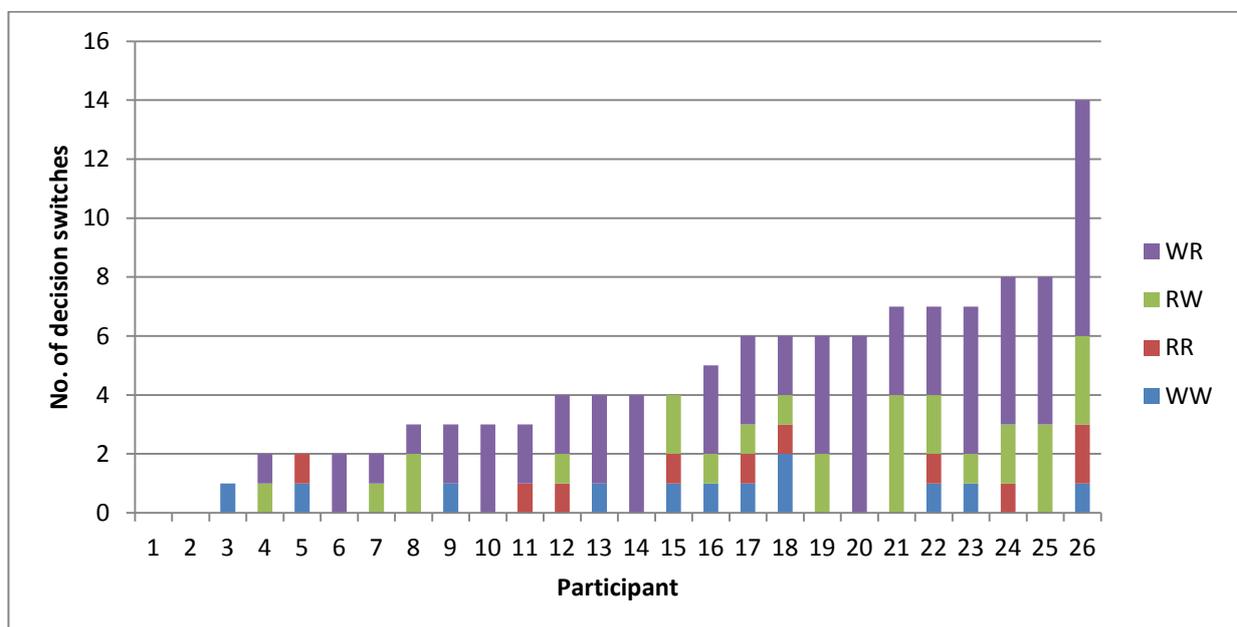


Figure 6.5 Number and type of decision scenarios by participant

## 6.6.2 Influencing factors:

### Task difficulty

The study aimed to measure how users responded to advice according to questions' degree of difficulty, i.e. if users accepted advice in hard cases significantly more than advice in medium cases.

Decision accuracy – the number of correct pre-advice decisions was greater in the medium than the hard difficulty condition,  $z=3.73$ ,  $N = 26$ ,  $p<0.0001$ .

Decision switching – decisions were switched more in the hard condition, than the medium condition, however this was not significant at the  $p=0.05$  level,  $z = -1.54$ ,  $N = 26$ ,  $p=0.06$ .

Negative consultations – there was no significant difference in number of RW switches between medium and hard conditions,  $z=-0.64$ ,  $N = 26$ ,  $p>0.05$ .

Task difficulty may influence decision confidence<sup>26</sup>. Mean pre-advice confidence in the hard condition was significantly lower than the medium condition,  $z = 4.14$ ,  $p<0.0001$ .

### Trust

#### *Trust in Clinical Decision Support Systems generally:*

Most participants ( $N = 16$ ) indicated a degree of trust (see fig 6.4) in CDSS in general.

Participants were divided into two groups with higher and lower general trust levels; the higher trust group consisted of those who Mostly or Somewhat trusted DSS in general ( $N=16$ ), those with lower trust consisted of those who stated they had No experience, or Distrusted DSS ( $N=10$ )<sup>27</sup>

Decision switching – there was no difference in decision switching between higher and lower trust groups,  $z = 0.45$ ,  $p>0.05$ . There was also no significant difference when the No experience group was removed from the test,  $p>0.05$ .

---

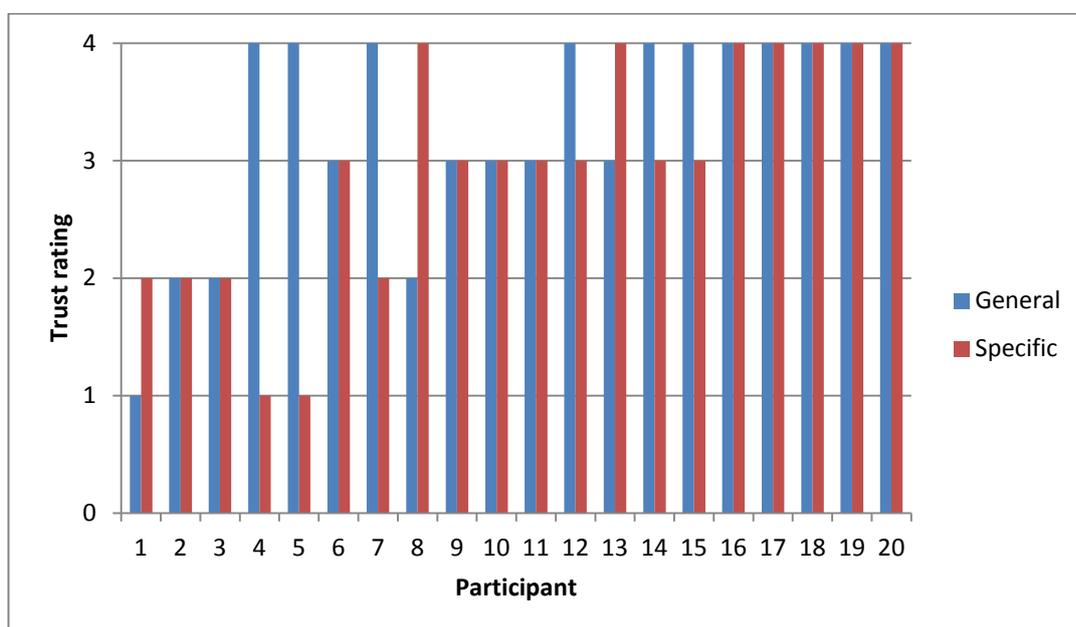
<sup>26</sup> Confidence was measured on a 6 point scale: Very confident – confident – somewhat confident – somewhat confident – unconfident – very unconfident

<sup>27</sup> This assumption must be taken with a caveat – having “No experience” ( $N=6$ ) does not necessarily directly relate to trust levels in DSS. Assumption made as novelty is often associated with aversion and distrust e.g. the phenomenon of neophobia

Negative consultations – there was no difference in RW switching between higher and lower trust groups,  $z = 0.26, p > 0.05$ .

***Trust in Clinical Decision Support System simulator specifically:***

After completing the study, participants were asked how much they trusted the DSS simulator they had encountered. Participant general trust in CDSS was compared to trust in the specific CDSS simulator used in this study (fig 6.6<sup>28</sup>).



**Figure 6.6 Comparison of trust ratings by participant**

No significant difference was found between the participants’ general trust in CDSS and the trust in this CDSS in particular,  $z = -1.2, p > 0.05$ .

Three users said they trusted CDSS generally but did not trust the CDSS simulator. One user placed higher trust in the simulator than CDSS generally. For all other users who had experience with CDSS, if they trusted CDSS generally, they trusted the simulator, and vice versa.

<sup>28</sup> Where 4 = Mostly trust, 3 = Somewhat trust, 2 = Somewhat distrust, 1 = Mostly distrust

Participants were divided into two groups with higher and lower specific trust levels; those who stated they Mostly or Somewhat trusted the CDSS simulator (N=18), and those who said they Mostly or Somewhat distrusted the simulator (N = 8).

Decision switching – there was a higher number of decision switches in the higher trust group than the lower trust group,  $z = 2.17$ ,  $p < 0.05$ .

Negative consultations – more RW switches were made in the higher trust group than the lower trust group, however, this was not significant at the  $p = 0.05$  level,  $z = 1.47$ ,  $p = 0.07$ .

### **Confidence:**

The confidence the participant had in each prescribing decision made was recorded, both before and after advice. The average confidence rating was compared between decision confidence (4.36) versus after decision confidence after advice (4.54); there was a significant increase in decision confidence post advice,  $z = -1.66$ ,  $p < 0.05$ .

Of participants that made any decision switches (N = 24), mean pre-advice decision confidence was significantly lower in decisions where switching occurred (3.76) than when decision switching did not occur (4.5),  $z = -3.61$ ,  $p < 0.0005$ .

A Spearman's rank correlation found there was no significant relationship between pre-advice decision confidence and years of clinical experience,  $\rho = 0.085$ ,  $p > 0.05$ .

### ***Trade off between Trust and Confidence***

To assess whether there was a trade-off between specific trust in the CDSS simulator and confidence and whether higher trust coupled with lower confidence was associated with more decision switching, a 2 way ANOVA was carried out. The dependent variable was number of decision switches; the independent variables were trust (with 2 levels of high and low) and confidence (with 2 levels of high and low). The highest mean number of decision switches was found in the high trust-low confidence condition; the lowest mean number of decision switches was found in the high confidence-low trust condition (fig 6.7).

A significant main effect of trust was found,  $F(1, 22) = 6.45, p < 0.05$ . There was no significant effect of confidence,  $F(1, 22) = 2.89, p > 0.05$ . There was no significant interaction between trust and confidence on decision switching,  $F(1, 22) = 0.001, p > 0.05$ .

Levene's test was not significant; homogeneity of variance can be assumed  $F(3,22) = 1.05, p > 0.05$ .

The Shapiro-Wilk test was not significant,  $p > 0.05$ , thus normality of distribution can be assumed.

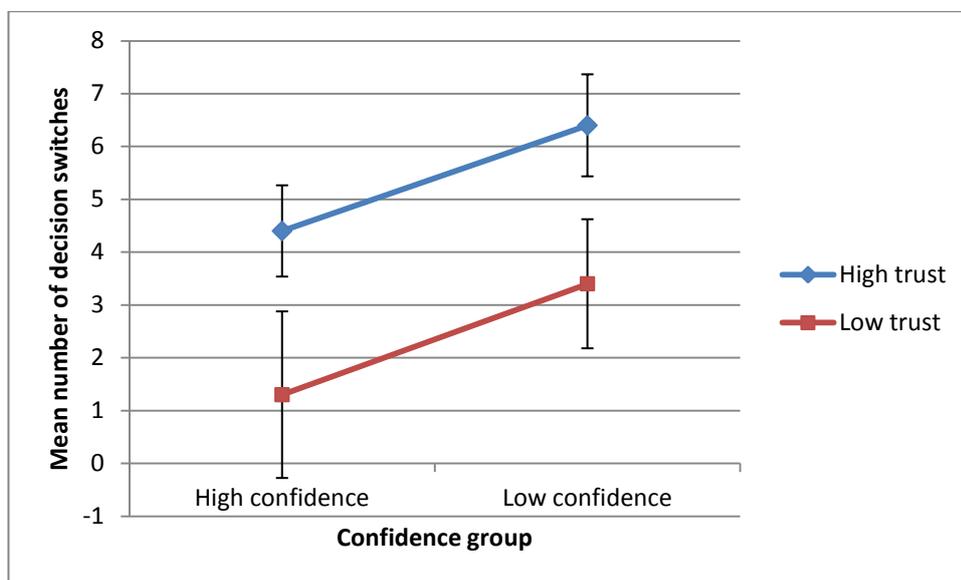


Figure 6.7 Mean number of decision switches by high versus low trust and confidence

## Experience

### *Clinical Decision Support System Experience:*

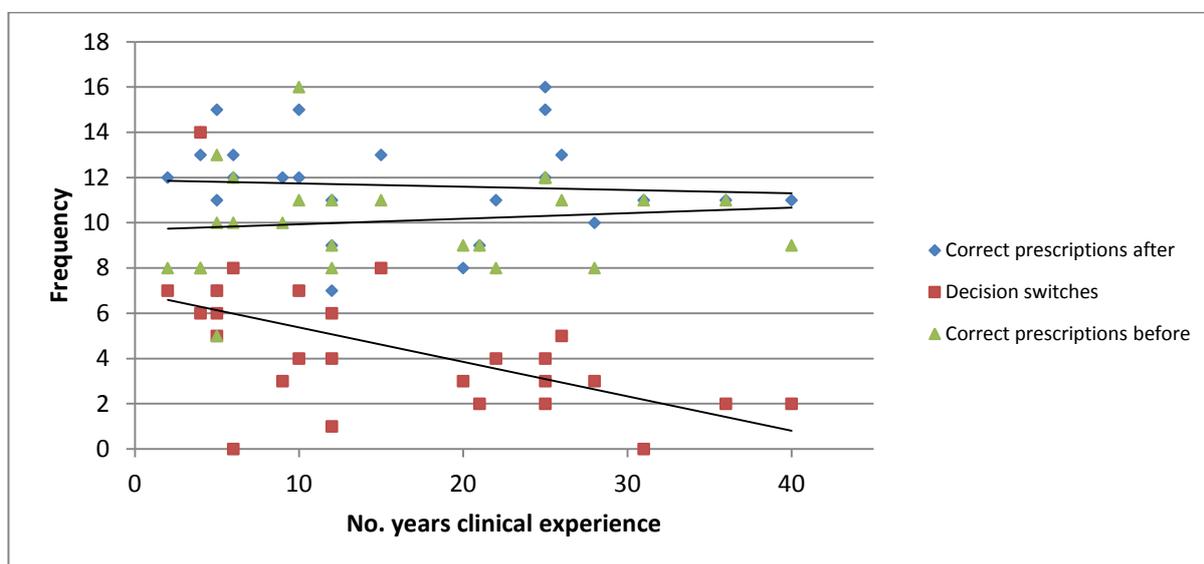
Most participants (N=20) stated that they had some experience of using CDSS (fig 6.3). The participants were split into 2 groups: those who Never – Very Rarely used DSS (N= 12), and those who Rarely– Frequently used DSS (N=14).

Decision switching – there was no significant difference between the number of switches in the group with higher stated use compared with those with lower stated use,  $z = 0, p > 0.05$ .

Negative consultations - there was no significant difference between the number of RW switches in the group with higher stated use compared with those with lower stated use,  $z = -0.93, p > 0.05$ .

**Clinical Experience:**

To see if there is a relationship with clinical experience, the stated number of years of clinical experience by participant was correlated with the number of correct prescriptions (before and after seeing advice). There was no significant correlation between the years of clinical experience and the number of correct pre-advice answers,  $R_s = 0.23$ ,  $p > 0.05$ , or post-advice answers,  $R_s = -0.19$ ,  $p > 0.05$ , but there was a significant negative correlation between years of clinical experience and number of answer switches,  $R_s = -0.61$ ,  $p < 0.005$  (fig 6.8).



**Figure 6.8 Relationship between years of clinical experience, and decision switching correct prescriptions**

Participants were split into two clinical experience groups; lower group (range of 2 – 12 years clinical experience,  $N = 14$ ) and a higher group (range of 15 – 40 years,  $N = 12$ ).

Decision switching – significantly more decision switches occurred in the group with less stated clinical experience than the higher clinical experience group,  $z = 2.13$ ,  $p < 0.05$ .

Negative consultations – more RW switches occurred in the group with less stated clinical experience, but this was not significant,  $z = 1.49$ ,  $p = 0.07$ .

## **Age**

A Kruskal-Wallis test indicated that there was no difference across the 8 age groups (in 5 year bands) in terms of number of RW switches  $p > 0.05$ , but there was a significant difference in terms of switches in general  $p < 0.05$ .

Decision switching was compared between participants with lower age (26 – 45 years,  $N=13$ ) and the upper age groups (46 - 65 years,  $N=13$ ).

Decision switching – there was no significant difference between younger and older age groups,  $z = 1$ ,  $p > 0.05$ .

Negative consultations – there was no significant difference between younger and older age groups,  $z = 1.31$ ,  $p > 0.05$ .

## **Validation of Analysis**

There are known to be many individual differences with regard to prescribing (see section 2.3.2). This is just one of the factors that need to be taken into account when looking at these results. To investigate any other potential biases or confounding variables, certain checks were carried out.

- 1) Categorisation of difficulty: whether the categorization of the questions (medium versus hard) was valid. A proxy measure for this may be the difference in accuracy between difficult and easy scenarios overall. It could be hypothesised that prescribing for scenarios classified as difficult would have been more taxing for knowledge resources for the users – comparatively less accuracy in “difficult” cases, as was demonstrated in the results, may imply that this categorisation was valid, thus adding confidence in the categorisation.
- 2) Fatigue effects: Prescribing for 20 scenarios over around 30 minutes may cause decreases in user performance due to tiredness, boredom. To mitigate this, the order of the scenario was randomised. The dropout rate once users had begun the experiment was low (only 3 users dropped out at this phase with very few scenarios completed), it could imply that the study did not experience overwhelming fatigue effects.

### 6.6.3 Qualitative analysis

Qualitative techniques were used to explore, guide and provide context to the quantitative analysis. Through carrying out a qualitative evaluation, we aimed to capture certain factors and opinions involved in the prescribing decisions and DSS use that were not quantitatively measurable, in this case it was important to investigate awareness of reliability and stated use as it may have bearing on propensity for AB. The end of the study comprised some open ended freetext boxes with questions to ascertain whether the user was aware that the advice was sometimes unreliable, whether they used it, and what are important factors in a DSS. This page also asked how much they trusted the advice on a 6 point Likert scale from Completely Trust to Completely Distrust.

Qualitative questions:

1. How reliable did you find the advice?
2. Did you always follow it?
3. What are important factors in the design of DSS to you?

1. There was a mixed opinion about the advice. Most participants noticed that some of the advice was dubious and stated that whilst most of the advice seemed appropriate, there were some cases which they were not confident to use the advice in. Nine participants stated that they found the advice reliable, 13 participants expressed mixed opinion about the reliability of the simulator, with the remaining 4 participants stating that they found the simulator unreliable. The latter 4 all still made decision switches (ranging from 2 – 4), and two participants made 1 or 2 AB switches – the implication perhaps being that automation bias could still occur in people who explicitly state they distrust a DSS.
2. None of the participants stated that they followed the advice all the time, most stated “no” or “Rarely”. Some stated that they followed when they were unsure of what to prescribe, or the advice seemed to be a better alternative. This question was potentially loaded in hindsight, and perhaps could have been better worded, essentially to try to get information about reasons for following or not (e.g. asking “What were your reasons for following advice”).

3. Twelve broadly independent factors were cited by participants as important in the design of DSS:
  - a. Provision of rationale (7 mentions)
  - b. Reference to evidence base (5 mentions)
  - c. Simplicity (5 mentions)
  - d. Accuracy (4 mentions)
  - e. Integration into workflow (4 mentions)
  - f. Choices/ multiple suggestions (3 mentions)
  - g. Flexibility (2 mentions)
  - h. Provision of non-drug treatment advice (2 mentions)
  - i. Cost effectiveness (1 mention)
  - j. Non intrusive (1 mention)
  - k. Speed (1 mention)

NB. The results from this empirical study are to be submitted to the Journal of the American Medical Informatics Association.

## **6.7 Summary**

The response rate in this study was very low. This has a potential effect on the representativeness of the sample of the general UK GP population.

The demographics of the participants were recorded. As an assessment of how representative of the population under investigation the results may be, these will be compared to the general demographic profile of UK GPs.

The analysis of the results suggests that AB is a replicable effect, which may be affected by more immediate influencing factors, such as trust in the CDSS advice, decision confidence, and task difficulty. More general factors, such as age, and CDSS experience (as measured by stated frequency of use) did not impact advice taking/decision switching and negative consultations in this study. Awareness of the factors that influence reliance and overreliance on CDSS advice may help identify cases in which there may be a higher risk of AB error. With the increasing emphasis on using technology to address knowledge gaps in healthcare, the introduction of new types of errors may become more prevalent and need to be accounted for.

The following chapters discuss the results and limitations of the study.

## 7. Overall Discussion

### 7.1 Introduction

This thesis explored the little-researched AB effect, following the MRC recommended methodology for investigating complex interventions; firstly a literature and systematic review was carried out looking at potential factors which could affect AB, which allowed a model to be generated to hypothesise how the system involved works (in this case, generating a conceptual model (fig 2.5), ontology (section 2.9.3) and simple experimental model (section 4.3). This informed the following empirical studies (including a pilot). The empirical study provides the first a priori study into the effect, finding that AB exists in a simulated primary care prescribing context and is potentially more influenced by more immediate factors such as trust in the simulator, confidence in the decisions being made, and the perceived difficulty of the task.

### 7.2 Results summary and discussion

#### 7.2.1 Reviews and models

The starting steps for the research involved forming the evidence base. The literature review (feeding into the systematic review) unearthed varying rates and ways to measure AB, and a large number of potential factors and thus situations in which it could arise. This evidence base fed into development of a conceptual model, a pilot ontology for AB and a testable model (demonstrating different ways of modelling results).

#### 7.2.2 Empirical study

The empirical study sought to investigate the rate of AB within a specific domain, and some of the more compelling, testable factors unearthed by the reviews. The results showed a mixed level of support for the testable model in fig 4.1. The results of the empirical study will be discussed in relation to the evidence base.

##### *1. Decision switching and Automation Bias*

The decision switching rate was 22.5% of all cases – this compares well with the finding by Dreiseitl and Binder (2005)<sup>167</sup> that in their study in 24% (86/357) of dermatology cases the

physician switched their decision after receiving CDSS advice. The accuracy of the CDSS was not recorded in their study however, so this cannot be compared.

The rate of switching in cases with “correct” and “incorrect” advice was similar. This implies there was not an overall greater distinction than chance of correct and incorrect advice. This highlights the importance of reliability of CDSS – to a certain extent, people may not be able to differentiate between good and bad advice (when confronted solely with the advice, and no additional information).

The baseline average accuracy of the participants was 50.38% (which is similar to a recent ePrescribing alert study where the control group was on average 48.2% correct (Scott, 2011)<sup>12</sup> (of 24 junior doctors)), which improved to 58% after advice; 13.1% of cases saw an improvement, 5.2% of cases saw a worsening of decision outcome resulting in a net improvement of 8%. The rate of RW decisions in this study is similar but slightly lower than those found in the systematic review, ranging from 6 - 11%. A caveat here is that the papers in the systematic review did not mention or discuss the rate of reliability for the decisions aids under study (in this case, the overall rate was set explicitly at 70%). This slightly lower rate could also be due to the study potentially being carried out in a less distracted environment (the study was online, and thus could be carried out when the participant had free time). The primary care environment could be a domain for lower AB rates (compared to more acute care, or alert type situations), as there may be less environmental pressure, possibly more familiar cases, and a greater range of appropriate drugs. The variation could also be a product of the huge variation in prescribing patterns (section 2.3.2) and decision styles and differences in physician type and representativeness due to low response rate (see Limitations section below, section 7.3). This final explanation is why the caveat was included for overall results (N = 520), and why, though they demonstrate an effect in this situation, they may not be generalisable to other situations.

## *2. Automation Bias related factors*

The factors found in the literature were grouped into the four causal areas of context and task (which could be grouped into environmental factors), user, and CDSS.

## Environmental factors

### *Task difficulty*

Task difficulty (of which task complexity is a factor) is posited to increase/exaggerate reliance on more heuristic style decision making (Bin, 2009)<sup>146</sup> (for example, there is support that it affects the trust levels in a DSS, Daly, 2002)<sup>150</sup>, whereby more non-compensatory strategies are used i.e. more reliance on fewer cues (perceived to have an adequate correlation with the measure in question) to make a decision (Chinburapa, 1993)<sup>149</sup>.

In the empirical study, task difficulty impacted decision switching, with higher switching occurring the more difficult prescribing scenarios. However, the number of RW switches was not significantly different between “hard” and “medium” conditions. This, may however been as a result of low overall number of AB decisions rather than there being no real effect. An alternative explanation for the lack of significance may have been that users could differentiate between good and bad advice. However, this was not supported by the results – overall, there was no significant proportional difference between AB errors in the good and bad advice cases.

Confidence is linked to task difficulty with significantly lower decision confidence in “hard” prescribing cases. This was not unearthed in the literature review, so not included in the conceptual model, but it can be hypothesised that confidence mediates the relationship between task difficulty and propensity to use CDSS advice.

Concurring with the literature e.g. Berner (1999)<sup>152</sup>, participant accuracy was also higher in the “medium” condition, adding validation that cases were properly coded.

### *Repetitive tasks and fatigue effects*

Fatigue effects can erode task performance (Langhals, 2001)<sup>153</sup>. Fatigue effects were not explicitly investigated in this study due to the blind randomisation of the 20 scenarios per user, however, a proxy measure could be the low dropout rate once the study had been started (3 people early in the experiment). Feedback from the pilot study also indicated that the study length was appropriate so as to not elicit significant fatigue effects.

*Time pressure*, unfortunately, could not be investigated due to low response rate.

## User factors

### *Trust*

There was a difference between global versus specific trust in CDSS in terms of effect on decision switching and AB. Trust in a DSS can increase the bias towards its output, as a clinician will perceive it as a correct course of action (Muir, 1994)<sup>180</sup>.

In *general*, 16 participants stated that they somewhat / mostly trusted CDSS (though 4/26 stated distrust in CDSS generally, which may corroborate with the literature on disuse of CDSS). However, number of switches was not significantly different between those stating higher versus lower general trust levels.

No difference was found in the levels of trust in CDSS generally, versus the stated trust in the CDSS simulator (implying that the study was a valid simulator of CDSS). However, differences were found between participants stating higher versus lower trust in the CDSS simulator *specifically*. Overall rate of switching was significantly higher for those who stated a higher level of trust in the simulator. There was a higher number of RW decisions in the higher trust group, but this was not significant at the  $p = 0.05$  level (in this case  $p = 0.07$ ). This supports the notion that trust may be a strong influencing factor in AB (de Vries, 2003<sup>171</sup>; Wu et al, 2008<sup>200</sup>).

### *Confidence*

Decision confidence was significantly higher overall after advice was taken. Pre advice confidence was significantly lower in cases where decisions were switched, implying, that along with trust, confidence is a psychological concept that has a significant influence on reliance on advice and decision switching e.g. confidence heuristic (Price, 2004)<sup>116</sup> – physicians are more willing to accept a CDSS recommendation when they are less confident in their prescription (Dreiseitl, 2005)<sup>167</sup>.

High trust in DSS and low decision confidence has been posited as a trade-off which may lead to reliance and AB (Lee and Moray, 1992)<sup>193</sup>, however this was not supported by the results in this study – individually trust and to a lesser extent confidence levels predicted decision switching, but not necessarily taken together. This may have been due to low number of cases, and the levels of confidence taken as a mean per participant to test against the stated level of

trust in the simulator per person (N = 26 cases), which only included 4 ordinal levels of trust (from Mostly distrust – Mostly trust) which may not have allowed enough variation for a more sensitive analysis. Perhaps a better measure would have been to ask participants for the trust in advice at each scenario and inputted the confidence in pre-advice decision and trust in advice to the model.

Of the cases in which decisions were switched, again, there was a significant increase in decision confidence after advice was taken. This is a potential danger if, as posited by Westbrook et al (2005)<sup>192</sup> - clinicians' confidence in their answer is not always related to the answer being correct.

It was also found that there was no significant positive correlation between clinical experience and decision confidence, contrary to suggestions in the literature. This may have been due to a smaller study sample (lower statistical power), or an unrepresentative sample skewed towards participants with somewhat similar levels of decision confidence. The length of clinical experience also did not distinguish between types of clinical experience, so as an improvement in hindsight the length of primary care clinical experience could have been investigated. Also, stated confidence can be divergent from observed confidence, and there is the common effect of central tendency with Likert scales<sup>315</sup>. Overconfidence is a decision making bias – it may be that for users of a CDSS a piece of advice is used as a heuristic cue for a “correct” prescription, leading to a premature stopping of cognitive search processes and the choosing of an “incorrect” answer, which leads to overconfidence.

### *Experience*

Two types of experience were investigated in relation to decisions with CDSS advice: experience with CDSS (with frequency of use used as a proxy), and estimated length of clinical experience.

#### *a. Clinical Decision Support Systems*

The results indicated that participants overall used CDSS in their working lives fairly infrequently. There were no differences in performance between those who stated more frequent use in terms of switching, and those of RW answers. It is possible the measure was

not adequate for the purpose of gauging DSS experience. For example, another proxy measure could have been to ask for participants' self-reported computer skills; however, similarly to this study Westbrook et al (2005)<sup>261</sup> found that there were no performance differences in decisions between different reported skill levels. It is possible that the effect of experience on CDSS reliance is more CDSS specific than general computer or CDSS experience i.e. more experience with a specific CDSS improves reliance rather than all DSS.

#### *b. Clinical*

Experience was found to not be related to overall performance in this study (as measured by before and after advice correct decisions), but negatively related to number of answer switches; participants with fewer stated years of experience were more likely to switch. Dreiseitl (2005)<sup>167</sup> found there was a slight but significant negative correlation between susceptibility to change and experience level of the physicians. There was also more RW switching in lower clinical experience groups. The literature may suggest that this may be a product of experience increasing clinical knowledge and thus confidence, however confidence was only somewhat correlated with clinical experience. As previously mentioned this could be a product of central tendency; if experience is linked positively to confidence then this variation may be masked, but as in this case be revealed more behaviourally (rather than stated confidence) by rate of decision switching. This may be an effect of global versus specific confidence however i.e. experience may have a stronger effect on overall self confidence than individual decision confidence.

#### Clinical Decision Support System factors

The qualitative section of the study elicited a number of factors which the sampled GPs suggested were important in terms of CDSS design, a number of which are supported by the literature a potential design factors to promote appropriate reliance. These involved CDSS content, format and implementation factors.

Five main *content* focussed factors were mentioned by participants. The most mentioned factor was suggesting that provision of rationale for the advice would improve prescribing performance. This improves the cross verification process and allows clinicians an extra route to identify questionable advice (the trade off is the time it takes to verify the advice) e.g.

Kawamoto (2005)<sup>20</sup>, Dzindolet (2003)<sup>185</sup>. Similarly, providing users with a reference to the evidence base for the prescription in question allows the user to check the advice. Participants also felt that, particularly in light of recent controversies with over-prescribing<sup>29</sup> (for example, the Medicines and Healthcare Products Regulation Agency has launched an investigation into concerns about the lack of training of doctors in pharmacology), it would be useful to have additional non-drug treatment advice.

Flexibility of CDSS was cited in terms of allowing for plenty of variables so that can adjust to current clinical scenario.

This can be a component leading to increased accuracy of the content, which was also frequently mentioned as important. This factor is fundamental to the trust that a user can place in the advice given e.g. de Vries (2003)<sup>171</sup>, Muir (1994)<sup>180</sup>. Ensuring accuracy is the main aim of content factors, in terms of allowing the user to formulate an informed decision based on accurate CDSS output.

Four *format* focussed factors were mentioned. Most frequently mentioned was that the format had to be simple and non-distracting, which is linked to perceptual attention and decreased propensity to disuse (Kirlik, 1993)<sup>208</sup>. Similarly, participants preferred the CDSS to be non-intrusive; intrusiveness of advice can, for example, affect the trust in a CDSS (Bliss, 2003<sup>170</sup>, McGuirl, 2006<sup>288</sup>). Some participants suggested having multiple advice options, this could be seen positively as debiasing the clinician from one particular response and encouraging more thoughts on alternatives (however it could also increase cognitive load). This could relate to the use of information rather than recommendations as advice (i.e. less active directing of the prescription) (Sarter, 2001)<sup>287</sup>. The speed of the system was also seen as an important factor, which would link to workflow integration and decreasing cognitive load (for example reducing pressure on short term memory); some evidence implies that users value speed more than any other parameter<sup>316</sup>, though this is not duplicated here.

Not adding to cognitive and perceptual load is key to these format factors.

In terms of CDSS implementation, cost effectiveness and integration into the workflow were cited. Implementing in a manner that allows for evaluation and does not assume benefit will lead to better outcome<sup>317</sup>.

---

<sup>29</sup> For example: <http://www.independent.co.uk/life-style/health-and-families/health-news/nhs-told-to-end-culture-of-overprescribing-2175179.html>, <http://www.nature.com/bdj/journal/v188/n12/abs/4800571a.html>

### **7.3 Limitations**

A number of potentially limiting factors need to be taken into account with respect to the research carried out within these studies.

#### **7.3.1 Reviews of literature**

The major unresolved issue encountered during both the literature and systematic reviews is the incidental nature of the reporting of AB. Key papers do not set out to examine this phenomenon and thus it is not mentioned explicitly in the title, abstract or often even in the full text. In addition – both AB and Complacency processes remain ill defined; the posited overlap and similarity in error types implies more research and theory is required to understand the distinction and relationship between the concepts (such as the integrated model proposed in the Parasuraman et al. (2010)<sup>249</sup> paper). The effect is usually found in a post hoc analysis of data, and the data reported are often indirect, implicit, evidence of AB. This also means that papers with this finding are likely to have high heterogeneity in their search engine indexing. Another issue is that of the heterogeneity of results, which allowed only for a smaller meta-analysis. Heterogeneity within papers, in terms of materials and methodology, and outcome measures, can render direct comparisons difficult<sup>318</sup>. In this instance the systematic review may be best taken within the context of the preceding literature review of hypothetical factors to give a broader context and meaning to these results.

To address the gaps in empirical evidence relative to the available anecdotal evidence for AB, the systematic review focussed on quantitative evidence. However, randomized controlled trials may not be the most ecologically valid method of assessing over-reliance on technology in real world settings. Studies based on fieldwork, such as that reported by Campbell et al (2007)<sup>42</sup> should be looked at in conjunction with more controlled evidence to fully understand the nature of AB.

The reviews aimed to provide an evidence base for the existence of AB. Awareness of the nature of automation-induced errors should be used to inform DSS designers, policy makers, implementers and users. Given the potentially serious outcomes of medical decision error, it

would be beneficial to examine negative impacts of introducing automated clinical advice, as well as, the overall positive effects of CDSS on medical decision making.

### *Generalisation*

While these results come from divergent fields (e.g. aviation, motoring etc.), it can be reasonably hypothesised that the same underlying psychological processes are at work in terms of human propensity to overuse advice, particularly from automated sources e.g. satisficing and heuristic cognitive processing as described in Chapter 2. Different DSS types could mediate the exact nature and extent of AB. But it does appear to be a genuine problem across the fields surveyed in the systematic review. There may be, for example, differences in terms of whether advice is interruptive or non-interruptive. In this systematic review, the heterogeneity of papers may result in it being difficult to carry out a direct comparison.

### **7.3.2 Ontology of Automation Bias**

The skeleton ontology suggested in this research as an extension and application of the formal model is still formative and thus in pilot stage, to be added to and validated.

### **7.3.3 Empirical study**

Before the study was carried out, a number of potential issues were highlighted to be taken into account as potential factors to consider which may bias or confound the results in terms of internal validity:

- The “second look” effect, i.e. might decisions be more accurate at the user’s 2<sup>nd</sup> attempt, even without DSS advice. Friedman et al<sup>319</sup> described this effect as “On their second attempt at each problem, the students have additional time to think about the problem, to jog their memories and perhaps recall some additional relevant personal knowledge.” (pg 6)

To mitigate this potentially affecting some cases and not others, participants were not able to read the case text again after advice had been given, preventing them from processing the case a second time, thus avoiding 'second look' bias.

This is also potentially a limitation - as a trade off this may have affected the ecological validity, as ordinarily, GPs would be able to double check details that had slipped their short term memory.

- Potential conflation of the size of AB – it is possible that a switch from a correct to an incorrect answer may have been purely down to uncertainty rather than an effect of the automation. This was somewhat recorded by using confidence in decision as a proxy – users tended to switch more when they had less confidence in their pre-advice decision. This conflation is also likely to occur in real life, thus is still recording a real effect.
- Scenarios had to be developed that were challenging enough to allow there to be a measurable benefit to users from the correct system advice to avoid a ceiling effect. The baseline performance was recorded at 50.4%, which was appropriate to allow decision variation but avoid floor or ceiling effects.
- Sequence of scenario presentation - A considerable number of studies have investigated the effects of reliability on human behaviour e.g., Parasuraman et al (1993)<sup>21</sup>; Bliss et al (1995)<sup>170</sup>; Vries, et al (2003)<sup>171</sup>; Wiegmann et al (2001)<sup>172</sup>. In all these studies, reliability was manipulated by changing the overall error rate of the automation. However, none of these studies presented or mentioned the distribution of automaton errors across the experimental session. However systems with the same average reliability can have different patterns in time of human error. The location of errors within a specific range of time can have different effects on the way automation reliance and on the overall trust that humans report at the end of a session (Wickens and Xu, 2002)<sup>173</sup>. It must be noted that there was no significant difference in stated general CDSS trust and specific trust in this study's simulator, which may be an indication that sequence effects in this experiment may have been small, however this indication is insufficient to support the notion that there was no effect, thus this limitation remains a significant consideration. It is also noted in section 6.6.3 that only 4 participants stated they did not find the advice reliable, the remaining participants found the advice either reliable or of mixed reliability. It is not certain at which point participants noted their “first” unreliable piece of advice. Wickens and Xu would argue that humans interacting with an initially reliable system would have a different perception of the first automation error than humans interacting a system that is less reliable later on in time. This difference in perception is a product of experience which results in different expectations of the automation. Wickens and Xu (2002) suggest that experience is an important factor in the effects that errors by the automation have on human behaviour. They further argue that the first automation failure can result in a more pronounced drop of trust and reliance on the automation than subsequent failure - the “first

failure effect” (p. 8). Currently there is mixed evidence to support the existence of the “first failure effect” and the impact that it can have on the way humans perceive and interact with the automation (for evidence of the first failure effect see Molloy and Parasuraman, 1996<sup>174</sup>; for evidence against it see Wickens, et al, 2002<sup>175</sup>). The possible existence of the first failure effect, suggests that the distribution of errors in time is an important component in the relationship between automation reliability and human behaviour by affecting expectations and perceptions of the automation. Sanchez (2006)<sup>168</sup>, for example, found that participants who were exposed to automation error at the beginning or end of a series of cases relied more on automation than participants who were consistently shown error. This, it was suggested, implied that when automation frequently and randomly generates errors, humans’ reliance is more likely to remain lower than if the automation behaves reliably for an extended period of time.

Future research should perhaps focus more on the consistency of automation reliability rather than “overall reliability”.

The external validity / generalisability may have been affected by a number of factors including:

- Does it matter which clinical task the DSS addresses, e.g. prescribing, diagnosis, test ordering? Prescribing was chosen for reasons already outlined in section 2.3, but it is possible that different cognitive processes are involved in decision making in other domains, and would a bias towards automation manifest itself similarly – this could be a matter for further investigation. Due to the AB effect being demonstrated across a number of domains, it is reasonable to assume the effect is a general one and this is a frequently occurring underlying cognitive effect, but may differ in nature (quality, quantity) over different task types.
- The Hawthorne effect may be a problem in any study of this type i.e. participants may alter their behaviour due to the presence of researchers and the knowledge that they are carrying out a study, rather than due to the effect of experimental manipulations.

### **Low response rate**

After the study was carried out, one of the most obvious limitations to consider was the low response rate. This made it necessary to drop the time pressure condition (which was used as a proxy for environmental pressure). In terms of the results it impacts on the *power* of the study.

Some of the non-significant results in this study may prove to become amplified and more significant with a larger study sample size e.g. confidence and clinical experience, and statistical differences between AB instances in different conditions (e.g. task difficulty). The response rates outlined in the review into GP response rates in section 2.4 are higher than predicted by the experts involved in this study i.e. 5-8% predicted by the latter. This could be a factor of the review studies having more endorsement from the relevant higher authority, and a more insistent approach, with a better targeted sample, or different cultural attitudes towards research. It may be a factor of postal versus online surveys (survey results implied that GPs preferred postal contact), or (due to the slight deception of the true nature of the experiment) the lack of links to elaborate on the background for the study. It may be that the CDSS subject itself is not familiar to the majority of GPs, therefore there is less incentive to answer. There is also the possibility that research with low response rates can go unpublished, therefore there is a publication bias, underestimating the likelihood of obtaining an extremely low response rate in a study.

Despite the lower than expected sample size, the results largely appear to conform to previous literature and hypotheses.

In terms of number of cases, data was given for 520 cases (26 participants carrying out 20 scenarios each), which compares well for overall number with a similar study with 75 participants and 8 scenarios each (N = 557 overall) (Westbrook, 2005)<sup>261</sup>. The Scott et al (2011)<sup>12</sup> study used 24 participants carrying out 30 scenarios each (N= 504). Thus the number of cases and number of participants for this study compares favourably. Between participants variation is a factor to bear in mind, however, as there can be a huge amount of variation in prescribing behaviour between doctors (see section 2.3.2, international prescribing differences); more participants would allow this to be taken more into account. A strong caveat should be given here for caution generalising results which have been analysed by case (N = 520) as individual variation is not accounted for in these analyses.

The other effect is that low response rates may introduce sampling biases (in particular a self selection bias, as participation was voluntary and anonymous), which could affect the representativeness and generalisability of the results. For example, are people who responded more automation friendly, or more likely to be from a particular sub-demographic of GPs?

There may be a systematic reason why certain personality types completed the study. People who did not participate may be less likely to use DSS and thus less likely to commit AB. The latest Department of Health statistics on GP gender demographics are from 1999-2004<sup>30</sup>, which showed that in 2004 39% of GPs were female and this was an increasing trend (in this study 12/26, 46% of the participants were female). In 2005/2006<sup>31</sup> the age group with most GPs was 50-59 years (10946/33808 – 32.38%) – this is slightly older than the study sample with the age group with most participants - 42.31% (11/26) - being the 41-50 age group. To the researcher's knowledge there were no national data on years of clinical experience. The sample appears fairly representative, however bearing in mind that the results indicated that age and clinical experience are related to decision switching, it could indicate that the switching results in this study may be exaggerated. Another example of systematic difference between the sample and the general population is that they may have greater interest in healthcare technology, and a higher level of technology acceptance than “average”, which may have inflated certain effects. However no relationship was found between performance and frequency of DSS use (as a proxy for DSS experience). Stocks and Gunnell (2000)<sup>320</sup> found that UK GP non responders to postal surveys tended to be older, and less likely to possess a postgraduate medical qualification or belong to a practice that is involved with postgraduate or undergraduate training.

Some evidence exists which indicates that low response rate may not significantly affect the representativeness of the results. Holbrook (2005)<sup>321</sup> assessed whether lower response rates are associated with less unweighted demographic representativeness of a sample. By examining the results of 81 national surveys (by News Media and Government Contractor Survey Research Firms) with response rates varying from 5 - 54 %, they found that surveys with much lower response rates were only minimally less accurate. That said, results from very low response rates need to be framed with the caution that though the results may be valid and important they may not be wholly representative of the general population under study.

The ecological validity of the project also needs to be taken into account to assess if the results will transfer to real situations (though as previously discussed, simulations can be valid

---

<sup>30</sup> /[www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsStatistics/DH\\_4106726](http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsStatistics/DH_4106726)

<sup>31</sup>

[www.ic.nhs.uk/webfiles/publications/earnex0506/2005\\_06%20GP%20Earnings%20and%20Expenses%20Final%20Report%20TSC13rev2%2019%20mar.pdf](http://www.ic.nhs.uk/webfiles/publications/earnex0506/2005_06%20GP%20Earnings%20and%20Expenses%20Final%20Report%20TSC13rev2%2019%20mar.pdf)

methods of study). There may be less trust in the DSS than would be in real situations here – an effect of participants realising the study is simulated (thus less accountability to results of decisions), and the fact they were using unfamiliar “technology” in a more unusual situation. The manipulation of the DSS simulator accuracy may also be an issue of ecological validity (and thus potential for AB) and has pros and cons. Pros include that researchers can control for accuracy (as found in the literature review, there is a “U” curve in performance where performance is better with perfect or very imperfect DSS, but worse with middling), BUT it is not necessarily ecologically sound. None of the literature to the researcher’s knowledge shows explicitly the rate of accuracy in relation to AB – this can be more explicitly investigated in future studies, for which this research is a grounding.

The validation of the scenarios, though felt to be thorough, has to be again taken with the fact that there is often a degree of subjectivity with regard to “correct” and “incorrect” answers. Medicine is not an exact science, and there is unlikely to be a real “gold standard” prescription in most situations, as opinion can change in light of new evidence. In relation to ecological validity, the scenarios would require less validation if they had been taken as a random, anonymised set of real-life cases from a Primary Care practice (but this may be difficult to obtain, may have issues gaining ethical approval, and we would not have been able to manipulate variables to increase or decrease task difficulty). There is also the case that 20 different primary care conditions were encountered. In real life, physicians are likely to see more similar cases more frequently (especially due to seasonal differences, such as hayfever and influenza) i.e. the prevalence of problems presented here is unlikely to be representative of usual random sample of GP cases.

#### **7.4 Summary**

This chapter discussed the results for the empirical study in the context of previous research. Most findings conformed to hypotheses built from previous research. Where result diverged, possible explanations were given in terms of the limitations of the study.

Limitations included low response rate, potential influence and non-recording of sequencing of errors, and possible questions of the ecological validity of the simulated experiment.

The next chapter discusses the degree to which the study aims were achieved, additions to literature and potential benefits to different groups, and possible future directions research could take.

## 8. Conclusion

### 8.1 Introduction

This PhD has investigated rates, influencers and potential ways to mitigate AB using a combination of primary and secondary research within the MRC framework for complex interventions.

The following chapter discusses how the work has achieved objectives set out in section 1.3, the contributions to knowledge for different groups, and potential directions for future work.

### 8.2 Achievements based on aims and objectives

The study aim is ultimately to improve the safety, usability, clinical acceptance and effectiveness of CDSS by investigating potential rates of AB error, and helping highlight factors that may contribute towards AB related errors, through:

- Investigating the ability of users to detect bad advice. The systematic review, in particular explored quantitatively the rates of AB by two different types of measures: negative consultations (ranged from 6 – 11% of decisions), and comparisons of variable-accuracy CDSSs versus non CDSS groups – a small meta analysis showed that erroneous advice was more likely to be followed in the CDSS groups than in the control groups and when in error the CDSS increased the risk of making an incorrect decision by 26%. In this specific study, the negative consultation rate was 5.2%. Direct comparisons between percentages of negative consultations are tenuous, as the level of reliability in the meta-analysis studies were not stated, and in these overall figures in different papers within subjects variation was not taken into account. There was no difference in the post-advice confidence level whether the switch was a positive or negative consultation switching implying a lower level of awareness of the incorrectness of advice (switching was as likely to occur in “bad” advice conditions as “good” advice conditions). This however did not corroborate the pilot where, in retrospect, users expressed lowered decision confidence when the advice was incorrect – it may be that stated and observed behaviour is different. This corroborates some research e.g. Yeh and Wickens (1988)<sup>202</sup> who found that subjective perceptions and objective measurements are often dissociated.

- Investigating the risk factors leading to overreliance on automation via a literature and systematic review. This elicited many factors which are potentially directly or indirectly involved with AB, broadly grouped into contextual factors, task factors, user factors, and CDSS factors. These were organised into a conceptual model (an application of which was demonstrated by hierarchically organising the factors into a pilot ontological model grounded in DOLCE upper ontology), and also a model from which to test some hypotheses elicited about factors and their effects on AB (fig 4.1).
- Investigating the influence of potential risk factors in an empirical study. Some of the most compelling and feasible influencing factors were tested in the empirical study. More immediate factors such as trust in the CDSS simulator, decision confidence, and task difficulty were related to propensity to switch decisions, whereas more global factors such as overall trust in DSS, and DSS experience and participant age appeared to have little, or non-significant, effect. This corroborates the findings of Singh et al (1993b)<sup>291</sup>, who found that there was little correlation between generic attitudes towards all automation and automation use. Clinical experience was an exception, with a significant negative relationship with decision switching.
- Proposing follow on studies to investigate the effect of interventions to avoid AB e.g. the addition of confidence information, or source for advice etc. This research concentrates on investigating on the rate and possible involved factors of AB. The next section briefly outlines suggestions for future study and methodology, concentrating on design factors which were suggested by participants and are supported in the literature to mitigate inappropriate reliance – adding source information and simulator confidence in the advice. The relevance and usefulness of CDSS can be improved by improving data quality and decision support logic, however very little is known about user interface design and the impact on prescribing errors; there are few empiric studies investigating different approaches<sup>322</sup>.
- Formulating recommendations for DSS developers about how to make the output from the DSS more transparent to users. With the current findings, adding to the literature

surrounding the nature of AB, it can inform developers of the type of situation in which AB can arise. Recommendations could also follow on from the qualitative portion of the investigation and further studies – GPs suggested content and format factors surrounding usability (e.g. speed) and transparency (e.g. advice source information), which concurs with the literature for methods of improving appropriate reliance. In future studies GPs could be asked which design features could encourage appropriate usage with the context of AB, to make these suggestions more specific to reduce overreliance.

### **8.3 Factors to consider / Recommendations**

This research has shown by simulation that CDSS can improve prescribing decisions overall, but that switches from correct to incorrect decisions with the influence of incorrect CDSS advice can decrease the net improvement. As a result of this PhD a number of recommendations can be made for future research, and potential CDSS design and implementation.

In terms of who is most prone to AB, implementers should perhaps be aware of CDSS use with:

- Users with less clinical experience. Despite many studies showing that less experienced users benefitted most from CDSSs, this demographic could have the highest propensity for AB.
- People carrying out more difficult tasks may also be more likely to commit AB errors, more complex cases should potentially come with more suggestions to cross verify advice.
- Factors such as trust and confidence could still be primary drivers of the user reliance on automated advice. The dangers of incorrectly calibrated trust and confidence should be noted and measures could be taken to promote awareness.

In terms of CDSS design and implementation:

1. Reduce information load. Cognitive overload can increase heuristic use, which could potentially increase the use of automated advice to a higher extent than its reliability warrants.

2. Create an aesthetic and minimalist interface design. Users stated that they preferred simpler systems which fit in with the workflow. Linked to decreasing information load, a complicated interface may increase cognitive load and interrupt the cognitive process.
3. Support internal locus of control of users. A balance must be sought between overuse and underuse. Despite certain conservative bias e.g. egocentric bias, AB is a consequence of users relinquishing too much control and power to automated systems. Users should still be made aware that ultimate accountability and control remains with them. This point is supported by the next point
4. Provide concise additional information to aid cross verification and highlight less certain advice (increase transparency) – this increases transparency and supports users to make more informed decisions, particularly if they are uncertain of their decision, or the DSS's advice. For example McGuirl and Sarter (2006)<sup>288</sup> found that adding updated reliability information to a piece of advice improved reliance. This is recommended for further research (section 8.5).

#### **8.4 Addition to literature**

1. This research added a comprehensive literature review and systematic review of the literature to the evidence surrounding AB.
2. The research demonstrated AB in the Primary care prescribing domain for the first time.
3. On the basis of this a broad conceptual model of the literature, incorporating aspects of smaller reliance-based models was developed. This could then be used as a basis for developing a skeleton ontology for formative evaluation and a testable model of AB.
4. The results from the empirical study corroborate some findings in the AB literature: trust, confidence, task difficulty and clinical DSS are all influencers on decision switching, if not AB rate. This also added more information about AB rates to the body of literature.
5. This research used a JAS paradigm to test the salient factors involved in AB (from the reviews of literature and previous empirical evidence) which had not been explicitly tested in conjunction with overreliance. Not all hypotheses could however be fully tested, due to low response rate.
6. A new CDSS simulator was generated which allows experimenters to set the accuracy level of DSS (this is generally not reported in the Healthcare AB literature – may affect AB rates). The effect of varying DSS error rates could be assessed.

### 8.4.1 Outputs and benefits of the literature reviews

Table 8.1: Output benefits by target group – literature reviews

<b>Output</b>	<b>Target group</b>	<b>Benefit</b>
Review findings	Health professionals	Better day to day decisions
Review findings	Health policy makers	Evidence based policy
Review findings	Patients, the public	Better, safer care
Questions that were not answered; study problems	Researchers	Clearer aims, better methods, easier to get funding
What research is really needed	Funding bodies	Clarity about what research is really needed

### 8.4.2 Benefits of the study overall

Table 8.2: Benefits for different groups from the overall study

<b>Group who may benefit</b>	<b>How they may benefit</b>
DSS users, e.g. health professionals	Better insights into when to follow or ignore DSS advice More accurate decisions and better quality of care Better job satisfaction Lower liability exposure
NHS organisations purchasing DSS	Greater confidence in the usability and acceptability of DSS delivered as part of the national procurement
Patients	Safer, better quality decisions made by health professionals
DSS developers	Enhanced user acceptance Lower liability exposure; better quality product
Health informatics researchers	Insights into why people do or do not follow advice

## 8.5 Future work: Designing Clinical Decision Support for appropriate reliance

Future work could focus more on CDSS design to *mitigate AB* rather than *causes* (as this research focuses more on causes leading to AB). This could occur in the form of adding different information to the advice (e.g. in the qualitative analysis top mentions for improving decision support were to include advice rationale and evidence base).

### 8.5.1 Possible extra information to include

- Labelling the system clearly with its purpose, scope and intended users and user skills
- Giving an explanation of the advice. If the user detects that the explanation is suspect, then they are more likely to realise that the advice is wrong too.
- Using a matching algorithm to issue an alert when the case data do not closely match the type of cases used for training a data-derived system (e.g. Bayesian or neural net) or for deriving and testing the knowledge base of an “expert” system.
- Using two or more distinct methods to reason about the case (e.g. neural net and expert system) and issuing an alert when these methods come to differing conclusions.
- Giving the user a calculated probability estimate of the system’s certainty about its advice, e.g. “The most likely diagnoses are acute appendicitis (60%) and non specific abdominal pain (30%)”

Other outstanding questions relating to experimental variables include:

- Is the benefit of the additional information dependent on the frequency of bad advice? By increasing the rate of bad advice from, say, 5% to 20%, to make the study more feasible, will this bias the study results?
- What instructions should be given DSS users about the system scope and performance? This alone may cause them to accept or ignore bad advice, so needs to be strictly controlled
- Might the impact of added information vary by task?
- What impact would the sequence of errors have on CDSS reliance and AB?

An example of empirically testing some of these manipulations is given below.

## 8.5.2 Background

As previously mentioned some factors that increase advice utilisation have been found to be judge's self-confidence, trust in the source of advice, and judge or advisor expertise level (e.g. Azen and Budescu, 2003)<sup>34</sup>. Research has shown that judges' post-advice decision accuracy is related to the weight the judge gives to each advisor's recommendation (Humphrey et al, 2002)<sup>36</sup>. When judges and advisors have more decision-relevant information, they are on average more accurate - judges become more capable of discriminating between good and bad advice (weighing the former more highly).

There are a number of ways that DSS developers might seek to help users recognise when the DSS advice is likely to be wrong and so avoid these types of errors of dismissal and AB. Transparency of process can be one way of achieving this aim; by providing concise and relevant additional information, users can verify the advice given. This may improve the ability to recognise "bad" advice by the noting of discrepancies between the information and advice given.

Madhavan and Wiegmann (2007)<sup>209</sup> hypothesised that the visible behaviour of a decision aid affects its perceived reliability. Leaving the user out of the decision loop can lead to a lack of system understanding and loss of situational awareness, which can lead to unanticipated effects for more complex tasks. Even partially automated systems can result in measurable costs in human performance, such as loss of situational awareness, complacency, skill degradation, and decision biases<sup>323</sup>.

In a series of laboratory studies, Jamieson et al (2008)<sup>214</sup> examined the effects of system reliability information and interface features on human trust in, and reliance on, individual combat identification systems. It was found that providing updated reliability information led to more appropriate reliance on that feedback. This may be linked with the evidence that providing an "overall reliability" for automation is insufficient; the sequence of errors can also affect appropriateness of trust.

McGuirl and Sarter (2006)<sup>288</sup> found that updating the confidence level of the DSS alongside pieces of advice (as opposed to providing one overall fixed confidence level for the system) improved the appropriateness of user reliance, decreasing AB.

These methods involve giving the user extra information in addition to the advice, for example:

- Giving a brief explanation of the advice e.g. how the prescriptive information was inferred (decision relevant information). If the user detects that the explanation is suspect or there is a missing/inaccurate component to the knowledge base, then they are more likely to realise that the advice is wrong too
- Giving the user a calculated probability estimate of the system's certainty about its advice, e.g. for a diagnostic DSS "The most likely diagnoses are acute appendicitis (60%) and non-specific abdominal pain (30%)". These may be drawn from simple base rates.

Discordant information may lead to better monitoring/checking/validation of advice. Though it is still not foolproof, see the swiss cheese model (or cumulative error effect)<sup>324</sup>.

The DSS must strike a balance between providing users with enough information to produce an accurate decision and minimising the information load to facilitate the decision making process. Information given has to be clear and concise, in this case advice rather than alert-type support – optimal information load. Information load is a measure of the degree to which a user's memory is being used to process information on the display screens. It is a function of the task being performed, a person's familiarity with the task, and the design of the user interface itself.

Advice from CDSS can be given without context, or additional information. A key question for developers, users, patients and those who purchase decision support systems is what kind of additional information is most likely to lead to users correctly adhering to correct advice and ignoring incorrect advice? Some of the extra information is likely to affect users in avoiding faulty advice; others to encourage trust in "correct" advice.

There are also plenty of more design factors /format which may be relevant, particularly in relation to alert based DSS, such as colour and placement of advice<sup>325</sup>. This study could focus on more content related factors, rather than format of information.

As a caveat, information may also lead to overtrust - providing the user with a rationale as to why the system might err increases trust and reliance in a system. After observing the automated aid make errors, participants in a Dzindolet (2003)<sup>185</sup> study distrusted even reliable

aids. This was only mitigated if an explanation was provided as to why the aid might err. Knowing why the aid might err increased trust in the decision aid and increased reliance, even when the trust was unwarranted.

### **8.5.3 Aim**

To investigate how to enhance the ability of users of decision support system users to detect bad advice – focus on design content factors.

#### **Objectives:**

1. Does decision accuracy increase with more information?
2. Does automation bias decrease with more information?
3. Does confidence in decision increase with extra information?
4. Does advice verification behaviour increase with more information? Or is there an interaction depending on the consistency of the advice with the initial answer (It was found that physicians were strongly anchored by their initial diagnoses prior to using the CDSS (Berner 2003)<sup>219</sup> - this corroborated Teich et al<sup>326</sup> who found that physicians were more receptive to advice that did not require a change in initial plans)

### **8.5.4 Methodology**

#### **Two stage study:**

##### **1<sup>st</sup> stage – Think Aloud Protocol**

Participants: 10-12 GPs to represent the target population (half randomly allocated to CDSS with no extra information, other half to CDSS with extra information).

*Aim:* To qualitatively explore and compare how decisions are made with and without CDSS, and with and without extra information, and when correct or incorrect advices is given.

*Method:* A think aloud protocol allows researcher to better understand the internal mental processes of a participant while they carry out a task.

Usually one subject is observed by one or two observers. Two observers is optimal to obtain multiple perspectives, but may be off-putting for the subject, thus the session is often recorded for further analysis (with the permission of the subject).

### *Think Aloud process*

The subject should have the think aloud process and goal of the process (without the steps required to complete it). A practice session may be run e.g. to describe the characteristics of the test room, to familiarise subjects with the idea.

The subject should be informed that the DSS is being tested, not the users, and any difficulties are the fault of the DSS, that they can stop the process at any time, and that they can ask questions at any part of the process but that the observer cannot answer them. If necessary the subject can be prompted to carry on talking, if there is a lapse in the monologue.

Once the task is finished, the subject should be debriefed and asked for any additional feedback.

This will also serve to pilot the DSS advice with the additional information. For example, it must be ensured that the additional information is not overwhelming for the participants (it may bias participants to disuse).

### **2<sup>nd</sup> stage – Quantitative study (current PhD work may serve as control) testing the impact of additional information outlined above.**

**Participants:** The scenarios used would be the same as in the empirical experiment in this PhD (i.e. prescribing), therefore GPs will be approached as participants.

### **Design**

- This PhD work may serve as control for the proposed study. However there is a caveat in that the order of scenarios in future work would ideally be randomised but that errors would be designed to be distributed evenly through the experiment to account for potential differences in reliance stemming from differences in the sequence of error presentation. This PhD study may have to be repeated, controlling for error distribution.
- Within subjects design: Within subjects condition is before vs. after decision advice, with the before condition serving as the control.
- Other variables: monitoring (whether the participant uses the link to the official guidance for the clinical problem in the scenario), confidence (the confidence in decision before and after advice, and with vs. without extra information), demographic information (gender, experience)

## **Procedure**

The method will follow a typical JAS paradigm found in the psychology literature. Initial training may be incorporated to give participants some familiarisation in the accuracy of the DSS advice. The following procedure may take place:

1. Participants read information page followed by consent to participate page
2. Ask participants for experience using DSS and clinical experience (age band and gender will also be recorded)
3. Give participants 20 cases in random order – 70% with correct advice, 30% incorrect, alongside the extra information
4. Ask for their pre advice answer
5. Show the advice (whether correct or incorrect) and also give a link to official advice for prescribing for the clinical problem (the “monitoring step” – and record whether this is used)
6. Record their decision and confidence with decision before after receiving the simulated advice
7. Follow up questions: Which advice did you use to make your final decision? [percentage confidence , source of information , neither , both  ]
8. Debrief participant, including letting them know about scenarios with incorrect advice.

## **8.5.5 Results/ Analysis**

Primary outcome: The primary test for overreliance will be if the physician switches from a correct to an incorrect decision following a piece of incorrect advice (so called “negative switching”).

## **Statistics**

### **Primarily nonparametric testing**

#### **- Primary outcome:**

As in the PhD study, clinicians’ written responses to the scenario questions pre and post-system use will be compared. Scenario answers provided before displaying the DSS advice (stage 1) and after usage (stage 2) will be coded as “correct” or “incorrect” according to pre-validated scenario answers.

- **Secondary outcomes:**
- Overall improvement/decrease in performance (percentage correct before advice versus after advice).
- Similar statistical tests can be carried out as in section 6 of this PhD.

## 8.6 Summary

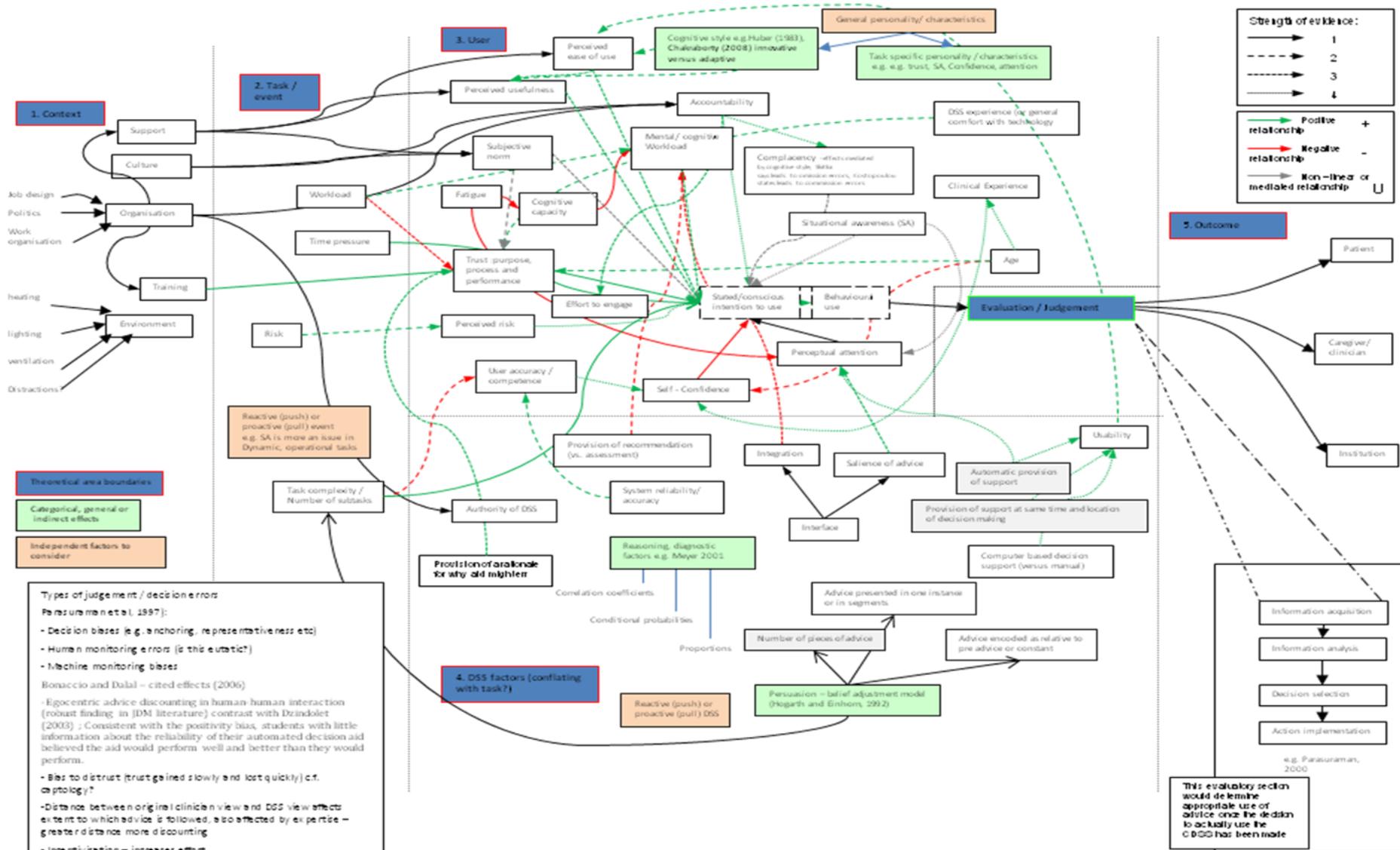
There is a common tacit assumption that that the use of a CDSS will improve decision quality. This study corroborates findings that though net accuracy tends to be improved, there is a penalty associated with correct decisions being switched to incorrect on the basis of incorrect advice – significantly lowering net performance. This can occur despite users feeling confident and satisfied with the DSS.

Factors that may increase the rate of decision switching, and potentially thus AB, include low decision confidence, high DSS trust, higher task difficulty and less user clinical experience. It should be assessed if changes to implementation and DSS design (for example providing information about the source of advice and how it was reasoned, and system level of confidence about accuracy) can improve reliance calibration, and it should be investigated whether the accidental errors outweigh the benefits of using a CDSS.

CDSS support software could have an important role in prescribing. It has been shown to improve overall decision quality and can offer promotion of cost-effective prescribing choices and can provide users with up-to-date prescribing information. People tend to project intelligence and objectivity onto computers which have in the past led to major errors being made e.g. the Therac-25 system accidents<sup>327</sup>, so CDSS should be implemented with caveats that users' own clinical judgement must always be maintained, and that systems should not be followed unquestioningly. The onus is also on the prescribing institution and software developers to regularly update and used to provide a balance between clinical and financial benefits.

It should also be ensured that CDSS are thoroughly tested before implementation to ensure benefits outweigh negatives, which include unforeseen errors due to automation.

# Appendix A: Conceptual model of reliance and AB



## **Appendix B: Quality assessment checklist for systematic review**

### **Quality Assessment Checklist**

A generic set of criteria have been adapted from relevant items taken from the CONSORT (for Randomised Trails, 2010) checklist, to assess for general paper quality. To more specifically assess the relevance and quality of papers found, a checklist was formulated for each specific question using the PICO categorisation to structure the criteria, taking into account the inclusion / exclusion criteria used in the previous extraction process.

#### **1. Generic criteria adapted from CONSORT items**

1 Mark if Yes, 0 if No

##### **Methods section**

- Trial design: Clear description of trial design (e.g. parallel, factorial)
- Participants: Setting and locations where the data was collected ;  
[Were the intervention and comparison groups treated identically (with the exception of the intervention)?]
- Outcomes: Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed
  
- Randomisation: Participants to experimental group
- Presentation of cases to participant]
  
- Blinding: Participants were unaware of purpose of experiment, or other experimental groups

##### **Results section**

- Numbers analysed: For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome

- Outcomes and Estimation: For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)
- Ancillary analyses: Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory

### **Discussion section**

- Limitations: Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses
- Generalisability: Applicability of the trial findings

Max score: 12

## **2. Relevance to study**

Aims:

- What is the rate of AB and what is the size of the problem
- Does it vary in different studies and settings? I.e. focus on causes, risk factors, barriers and facilitators and types of users.
- Is there a way to avoid AB? What is the impact of various methods to reduce AB?

PICO

*Context (external validity)*

1. Domain: Healthcare context > aviation (or other) or generic HCI (score 2, 1 respectively)
2. Definitions: AB (and synonyms) clearly defined a priori > AB effect reported post hoc > AB proxy measures mentioned e.g. cross verification > AB not mentioned by report (Score 3, 2, 1, 0 respectively)

*Population*

3. Participants: Physicians/healthcare professional as participants > experienced in the field vs naïve subjects (Score 2, 1, 0 respectively)

*Intervention*

4. Advice presentation: Text > visual or auditory or mixed (Score 2, 1 respectively)

5. DSS type: Passive (non interruptive) > active (interruptive) or adaptive (mixed level automation) (Score 2, 1 respectively)
6. Study assesses participant stated awareness of DSS: Yes > No (Score 2, 1 respectively)

*Control / design*

7. Design / controls: Experimental Studies: RCT > nRCT or before-after study > non controlled comparison study > Expert opinion/consensus/descriptive (Score 3, 2, 1, 0 respectively)

[Observational studies with control group: cohort or case > Observational studies with no controls: cross sectional, before-after, case series control – May not be relevant to the papers obtained, but need to double check for non-lab studies]

*Outcome*

8. Outcomes: Study reported omission and/or commission errors > no distinction made (Score 2, 1 respectively)
9. Outcomes: Rate for AB is rate of negative consultations (positive to negative) or opportunities to take bad advice taken > statistical difference / effect size > correlation between system inaccuracy and user inaccuracy > AB is strongly implied without statistical evidence (Score: 4, 3, 2, 1 respectively)

Max score: 22

## Appendix C: Final extracted papers from systematic review

Year	Title	Author	Journal	Objective	Design	Domain	Population	Intervention (DSS type)	Interruptive or non-interruptive	Presentation mode of advice	Control	Outcome	Other notes
2004	Effects of incorrect CAD output on human decision making in mammography	Alberdi	Academic Radiology	To investigate the effects of incorrect CAD output on the reliability of the decisions of human users.	Matched sample between subjects comparison (study 1 experimental group vs study 2 control)	Healthcare - diagnosis	Twenty film readers in experiment 1, 19 in experiment 2	CAD	Non interruptive	Visual	No CAD support	The average sensitivity of readers in Study 1 (with CAD) was significantly lower - 61% - than the average sensitivity of readers in Study 2 (without CAD) - 73%. Possible automation bias effects in CAD use.	
2008	Misuse of automated decision aids: Complacency, automation bias and the impact of training experience	Bahner	International Journal of Human-Computer Studies	To assess the impact of training on complacent behaviour and automation bias errors	Between subjects (AFIRA trained with faults vs no faults)	Aviation	24 engineering students	Automated aid supporting fault diagnosis and management (Automated Fault Identification and Recovery Agent, AFIRA).	Interruptive	Text	No non-DSS condition	Possible commission errors were assessed when the aid provided false recommendations. Only 5 out of 24 participants showed a "commission error" i.e. followed the wrong recommendation by initiating the suggested, but wrong, repair order. Because these participants were almost equally distributed across both experience and information groups, the effect was deemed to be not due to the type of training. The results provide evidence for complacency, reflected in an insufficient verification of the automation, while commission errors were associated with high levels of complacency. Participants of the "experience" group showed a significantly lower level of complacency i.e. sampled fewer parameters to verify the automated directives.	But there were differences between these 5 people and the other 19 with respect to fault identification times and their degree of complacency as measured by their verification behaviour: Participants who detected the failure of AFIRA needed almost twice as long for fault identification in the previous nine trials than participants who made a commission error. Both sub samples also differed with respect to their sampling behaviour. Participants who detected the false diagnosis of AFIRA for fault 10 were found to have spent significantly more effort in automation verification i.e. sampled a considerably higher portion of relevant information during the previous nine faults than those who committed a commission error

2005	The effects of operator trust, complacency potential, and task complexity on monitoring a highly reliable automated system.	Bailey	Dissertation Abstracts International: Section B: The Sciences and Engineering. US, ProQuest Information & Learning.	To assess the impact of system reliability, complacency potential, monitoring complexity, operator trust, and system experience on monitoring performance.	Exp 1: A 2 Reliability (high or low) X 3 Session X 3 Monitoring Complexity (gauge, mode, or digital readout) mixed design was used. /// Exp 2: System reliability and the degree of monitoring complexity were not manipulated - operator's ability to detect a single automation failure over several experimental sessions (more ecologically valid than first exp) /// Dependant variable was failure detection	Aviation	Exp 1 : 32 students, Exp 2: 9 students.	Automation alerting system	Interruptive	Visual	Multifactoria 1 - no non DSS control	Results indicated that realistic levels of system reliability severely impaired an operator's ability to monitor effectively. In addition, as system experience increased, operator performance for monitoring highly reliable systems continued to decline (high reliability mean detection rates of M= 51.7%, low reliability 66.7%). Further, operators who reported higher levels of trust, confidence, and more frequent usage of automation demonstrated poorer overall monitoring. The complexity of the monitoring task was also shown to be one of the most important factors influencing operator monitoring performance with poorer performance on more cognitively demanding tasks that continued to degrade as system experience increased (Correct detections for the gauge monitoring task were nearly three times higher than for the digital readout monitoring task). Results from both studies indicated that operator trust increased as a function of increasing system reliability and that as trust increased, monitoring performance decreased.	These results suggest that for highly reliable systems, increasing task complexity and extensive experience may severely impair an operator's ability to monitor for unanticipated system states. /// See Yerkes-Dodson type relationship between task complexity and monitoring performance.
------	---	--------	---	--	--	----------	---	----------------------------	--------------	--------	---	---	---

2003	Clinician Performance and Prominence of Diagnoses Displayed by a Clinical Diagnostic Decision Support System	Berner	AMIA 2003 Symposium Proceedings	To explore the extent to which consultations with DSSs improve clinicians' diagnostic hypotheses in a set of diagnostically challenging cases.	RCT, repeated measures	Healthcare - diagnosis	70 internal medicine residents	QMR	Non interruptive	Text	No QMR	The proportion of cases with correct diagnoses both prior to, and after using the CDSS, was 55%. The mean proportion of QMR screens where the correct diagnosis was prominently displayed was 17%. After using QMR, 130 of the 151 cases (86%) that were initially correct still contained the correct diagnosis (IE implication that 21 of these were incorrect after using QMR - negative consultations).. Similarly, after using QMR, in only 20 of the cases that initially failed to consider the correct diagnosis, was the correct diagnosis included on the final differential; 83% were unchanged from their unaided diagnosis.	Anchoring bias: After using QMR, 130 of the 151 cases (86%) that were initially correct still contained the correct diagnosis. Similarly, after using QMR, in only 20 of the cases that initially failed to consider the correct diagnosis, was the correct diagnosis included on the final differential; 83% were unchanged from their unaided diagnosis. When the unaided diagnosis was incorrect, but QMR displayed the correct diagnosis in a prominent position, in all but one of the cases (7/8 or 88%), subjects added the correct diagnosis to their final differential. The remaining correct diagnoses came from other positions, but most of the time if the correct diagnosis was not considered prior to using QMR and was not in the top ten diagnoses displayed, there was no change to a correct diagnosis after using QMR.
2004	The Influence of Task Load and Automation Trust on Deception Detection	Biros	Group Decision and Negotiation	To investigate the effects that user task load level has on the relationship between an individual's trust in and subsequent use of a system's automation.	Between subjects: Trust: groups 2,4 had IW manipulation , groups 1,3 no IW. Taskload: groups 1,2 low taskload, groups 3,4 high taskload	Military - aviation	40 military graduates	AWACS Weapons Director Trainer (AWDT)	Interruptive	Text	No non DSS control	Correlations: significant and strong positive correlation between ratings of trust in system automation and automation use at a significance level using post-treatment trust and automation measures. This suggests that as a user's perception of trust in system automation increases so will his use of that system's automation. /// No statistically significant difference in automation use between treatment groups 1 and 3, (high trust groups) but does show a statistically significant difference, although minor, between groups 2 and 4 (low trust groups).despite perceptions of low system automation trust, individuals tend to use automation more when task loads increased.	Information Warfare (IW). It has been shown that indications of IW may reduce the level of trust individuals have in the automated system they are using (Bisantz et al. 2000; Fields 2001). /// Talks about AB but doesn't really test it with respect to testing reliance on incorrect information - just usage in general and imply that higher trust and workload = more AB.

2003	The Effect of External Safeguards on Human-Information System Trust in an Information Warfare Environment	Biros	36th Annual Hawaii International Conference on System Sciences (HICSS'03)	To examine if the presence of incidents such as hacking and data manipulation would affect the decision-makers trusting behaviour. Also examined if the use of external safeguards, such as the Computer Emergency Response Teams (CERT), would affect the decision-maker	The first experiment manipulation was the construct called external safeguards. Treatment groups one and four were told by the experiment facilitator that the NSF was very effective (90%) at detecting enemy information attacks and defending the network against these attacks. Treatment groups two and three were told by the experiment facilitator that the NSF was not very effective (60%) in the same tasks. The second manipulation, Information Warfare (IW), was operationalized in the form of an information manipulation resulting in two spoofing events. Spoofing is a tactic whereby the enemy has covertly gained access to the system and manipulates the track identity, such that a friendly aircraft	Aviation	<b>Exp 1:</b> 56 of airborne warning and control system (AWACS) operators <b>Exp 2:</b> Thirty-eight military officers	Network Security Force (NSF) that indicated an attempted attack against the network had occurred	Interruptive	Text	No non DSS control	Correlational matrix analysis. The findings from both experiments suggest that the presence of information security incidents in a fast-paced C2 environment have no effect on the decision-makers trusting behaviour. Decision makers continued to trust information systems even though information security incidents occurred.	It was found that dispositional trust and situational trust were well correlated with each other. However, no evidence was found to suggest that external safeguards or an information warfare environment had any influence on the participants trusting behaviour. Post experiment interviews suggested that participants were so involved in the task domain that they lost focus of the external safeguards and IW present in the experiment. This task saturation seem so influential that a second experiment was designed to mitigate its effects. It also employed the command and control simulator. However, rather than use a three-dimensional aircraft tracking simulation, the second experiment used a 2 dimensional surf-to-air missile (SAM) simulation. This served to reduce the task load on the participants. Like the first experiment, the second found support for hypothesis 1, and it found support that disposition to trust will have a positive influence on trusting behaviour. As with the first experiment, no support was found to suggest that the presence of external safeguard would have a positive affect on participants trusting behaviour. Further, no support was found to suggest that an IW environment (i.e. computer attack) would have a negative influence on trusting behaviour.
------	---	-------	---	---	---	----------	--	--	--------------	------	--------------------	--	--

					appears on the display as an enemy and an enemy aircraft appears on the display as a friendly. Treatment groups three and four were subject to an information manipulation event during the simulation, while treatment groups one and two were not. Exp 2 essentially the same but with a decreased taskload.								
2004	Misdiagnosis of atrial fibrillation and its clinical consequences	Bogun	American Journal of Medicine	Computer algorithms are often used for cardiac rhythm interpretation and are subsequently corrected by an overreading physician. The purpose of this study was to assess the incidence and clinical consequences of misdiagnosis of atrial fibrillation based on a 12-lead electrocardiogram (ECG).	Retrospective	Healthcare - diagnosis	Ordering physicians, unknown number	CI of atrial fibrillation	Non interruptive	Text	No control	We found that 442 ECGs (19%) from 382 (35%) of the 1085 patients had been incorrectly interpreted as atrial fibrillation by the computer algorithm. In 92 patients (24%), the physician ordering the ECG had failed to correct the inaccurate interpretation, resulting in change in management and initiation of inappropriate treatment, including antiarrhythmic medications and anticoagulation in 39 patients (10%), as well as unnecessary additional diagnostic testing in 90 patients (24%). A final diagnosis of paroxysmal atrial fibrillation based on the initial incorrect interpretation of the ECGs was generated in 43 patients (11%).	FPs assessed. Not FNs. 19% (442 of 2298) of ECGs had incorrect computer interpretation. Of those, 10% had a change in clinical management due to misdiagnosis. Clinicians corrected the interpretation in 76% (290) patients and agreed with the incorrect diagnosis in 24% of patients (92). /// Level of expertise: those with higher levels of expertise were more able to correct interpretations than lower (e.g. internists, over specialities (non cardiology), p<0.05)

1996	The ameliorating effects of accountability on automation bias	Burdick	Human Interaction with Complex Systems	To investigate whether accountability might reduce automation bias two studies were conducted.		Aviation	Study 1 used 28 commercial airline pilots on a high-fidelity flight simulator while study 2 used 18 college students on a low-fidelity cockpit simulation.	Automated decision aid			UNSP ECIFI ED	Results from both studies indicate that perceived accountability for overall performance or accuracy significantly decreased both errors of omission and commission. In addition, subjects accountable for overall performance or accuracy were more likely to verify automated directives, indicating increased vigilance.	Recent research indicates that automated decision aids introduced into the workplace with the express purpose of reducing human error may have the paradoxical effect of simply changing the types of errors made.
1997	Effects of highlighting, validity, and feature type on air-to-ground target acquisition performance	Conejo	Published Masters Thesis - <a href="http://www.stormingmedia.us/74/7470/A747053.html">http://www.stormingmedia.us/74/7470/A747053.html</a>	To examine the differences in performance between natural and cultural (man made) feature types as targets and lead-in features under highlighted and non-highlighted conditions.	Within subjects. 1. Lead in feature type (natural or cultural), 2. target type (natural or cultural) 3. Highlighting condition (valid, incorrect highlighted, incorrect target highlighted and absent from view). Each pilot had 20 trials.	Aviation	18 aviation students	Target identification	Interruptive	Visual	Natural features and non-highlighting	Performance according to feature type was best under a target by lead-in interaction where the target feature type was different to that of the lead-in feature type. Valid highlighting did not provide significant benefit over nonhighlighted conditions; but invalid highlighting created performance costs which were not sufficiently improved by any feature type interaction or lead-in highlighted condition. Further analyses suggest that performance under target absent conditions results from different cognitive processes rather than when the target is visible. /// Valid trials: ANOVA for subjects score/ performance: No sig main effect or interaction effect of anything but the target variable (where cultural feature beat natural feature). Highly sig effect of highlighting validity on score. Overall average score (of 326 valid trials from 20 subjects) was 2.46; 2.32 on nonhighlighted trials and 2.63 for highlighted. See scoring matrix /// Highlighted trials analysis: The cost of invalid trials suggests that participants did get complacent - relying on automation even when they were unconfident of this choice and this effect appears stronger when incorrect highlighting occurs. From graph - valid trial was around 2.6, invalid was around 1.4, invalid and valid not in view was 1.6.	

2006	Automation Reliability in Unmanned Aerial Vehicle Control: A Reliance-Compliance Model of Automation Dependence in High Workload	Dixon	Human Factors	To highlight the qualitatively different effects of automation false alarms and misses as they relate to operator compliance and reliance, respectively.	Between subjects (control included - no automated aid) - 4 conditions with varying reliability FPs and FNs ; within subject high and low workload conditions	Aviation	Mix of pilots and non pilots. Experiment 1: 32 graduate and undergraduate students, 20 participants were licensed equally distributed across conditions. Experiment 2: 24 participants, same demographics as exp 1, same proportion of pilots to non pilots	Automated aid to alert for system failures	Interruptive	Auditory - warning tone	No automated aid	The 67% reliable conditions resulted in poorer detection rates than did the baseline condition, $t(19) = 1.97, p = .06$ . Detection rates were always poorer overall in the high workload condition. High workload condition: When automation was 100% reliable 88% of system failures (SFs) detected, and 68.8% when automation was 67% reliable (33% FPs), 92.9% when 67% reliable (33% FNs) versus 95.8% baseline (no alert). Low workload condition: When automation was 100% reliable 94.5% of SFs detected, and 68.8% when automation was 67% reliable (33% FPs), 97.9% when 67% reliable (33% FNs) versus 100% baseline (no alert)	Automation dependence emerges more if there is a high workload (costs if advice is incorrect, benefits if correct) /// An increase in false alarms is posited to reduce compliance, resulting in longer response times to automation alerts. In extreme cases, this results in a tendency to disregard those alerts entirely – the “cry wolf” effect. (Dixon & Wickens, 2006; Wickens, Dixon, Goh, & Hammer, 2005). An increase in the automation’s miss rate reduces reliance, causing the operator to allocate more attention to monitoring the raw data behind the automation in order to catch the possible automation misses.
2007	On the Independence of Compliance and Reliance: Are Automation False Alarms Worse Than Misses?	Dixon	Human Factors	Participants performed a tracking task and system monitoring task while aided by diagnostic automation. The goal of the study was to examine operator compliance and reliance as affected by automation failures and to clarify claims regarding independence of these two constructs.	Within subjects controlled	Generic HCI	Thirty-two undergraduate students performed the simulation that presented the visual display while dependent measures were collected.		Interruptive	Visual	No automated aid	The baseline condition produced performance better than the average of the two unreliable conditions, $t(14) = 2.43, p = .01$ . The FA60 condition ( $M = 2.04$ ) reduced performance worse than that of the M60 condition ( $M = 2.61$ ), $t(13) = 3.08, p < .01$ . Post hoc tests revealed that the baseline condition was performed better relative to the FA60 condition, $t(14) = 3.15, p < .01$ , but did not differ significantly from the M60 condition, $t(13) = 1.38, p > .10$ . were inclined to agree with the automation when it correctly detected an SF, the increased response times suggest that this agreement occurred was only after the participant doublechecked the raw data. When the automation presented an FA, operators incorrectly agreed only one third of the time. These two factors indicate low operator compliance.	When there was a signal, all groups tended to agree with automation but did so less with FA-prone automation ( $M = .93$ ) than with miss-prone automation ( $M = 1.00$ ), $t(14) = 3.75, p < .01$ . In contrast, when the automation was silent, the operator was less likely to agree in the miss-prone condition ( $M = .82$ ) than in the FA-prone condition ( $M = .92$ ), $t(13) = 2.14, p < .05$ . These findings are consistent with the postulation that FA-prone automation reduces compliance but that miss-prone automation reduces reliance.

2005	Do physicians value decision support? A look at the effect of decision support systems on physician opinion	Dreiseitl	Artificial Intelligence in Medicine	To investigate the question of how physicians react when faced with decision support suggestions that contradict their own diagnoses.	Repeated measures	Healthcare - diagnosis	52 dermatologists	Web-based CDSS to give advice on skin lesions	Non interruptive	Text	Before advice decision	- Based on the recommendation of a CDSS, physicians are willing to change a dichotomous decision in 24% of the cases. - The number of times a decision is reversed correlates negatively with the experience level of the physicians using the system. - Physicians are more willing to accept a CDSS recommendation when they are not confident of their diagnosis.	Does not assess for appropriateness of decision*
2000	A clinical decision support system for prevention of venous thromboembolism - Effect on physician behaviour	Durieux	JAMA	To determine whether presentation of venous thromboembolism prophylaxis guidelines using a CDSS increases the proportion of appropriate clinical practice decisions made.	CDSS - The study had an alternating timeseries design, with three 10-week intervention periods, four 10-week control periods, and a 4-week washout between each period.	Healthcare	Hospital physicians	CDSS	Non interruptive	Text	No CDSS condition	191 inappropriate prescriptions ordered during the control periods and the 44 inappropriate prescriptions that were not changed during the intervention periods. The system did not allow for analysis of the 69 initial prescriptions that were changed according to the recommendation.	Does not give data for negative consultations
2002	The Perceived Utility of Human and Automated Aids in a Visual Detection Task	Dzindolet	Human Factors			Generic HCI	Study 1: 68 Cameron University students					Study 1: The results of Study 1 indicate that a bias toward automation exists; the participants predicted that the automated aids would perform better than human aids. The bias toward automation found in this study should not be confused with Mosier and Skitka's (1996) automation bias, which occurs when human operators rely on a decision heuristically; the decision does not necessarily have to have been made by an automated aid.	

2003	The role of trust in automation reliance.	Dzindolet	International Journal of Human-Computer Studies	To explore the relationship among automation reliability, trust, and reliance.	<p><b>Exp 1:</b> Repeated measures (pre and post DSS advice),</p> <p><b>Exp 2:</b> 2 (Aid's Performance Level: Superior or Inferior) x 2 (Provision of Aid's Decision: Present or Absent) x 3 (Type of Feedback: No Feedback, Cumulative Feedback, or Continuous Feedback) between subjects design.</p> <p><b>Exp 3:</b> Between subjects; 2 (aid's relative performance level: inferior or superior) x 2 (provision of a rationale for the aid's errors) 4 (aid's decision) design</p>	Generic HCI	Study 1: Fifteen students // Study 2: One hundred eighty students // Study 3: Twenty-four students	"Contrast detector" decision support	Interruptive	Visual	Exp 1 only: Before condition without DSS	<p><b>Exp 1:</b> Consistent with the positivity bias, students with little information about the reliability of their automated decision aid believed the aid would perform well and better than they would perform. Knowing little about the automated aid, participants deemed the aid trustworthy. <b>Exp 2:</b> Participants who were prevented from viewing decisions but received continuous feedback regarding the aid's performance seemed much more willing to trust a superior aid than those in other conditions. Eliminating operators' awareness of an automated decision aid's obvious errors (through blinding the participants to the decisions of the aid) was useful in promoting appropriate automation reliance if participants were continually reminded of their and their aid's performance. <b>Exp 3:</b> Participants who were given a reason why the aid might err trusted the aid's decisions more and were more likely to rely on the aid than those not provided with this information. Although superior aids were deemed more trustworthy than inferior aids, the two were equally likely to be relied upon. Participants paired with an inferior aid were just as likely to rely on the aid as were those paired with a superior aid - thus creating potential for misuse.</p>
2007	Influence of Computer-Aided Detection on Performance of Screening Mammography	Fenton	The New England Journal of Medicine	To assess the effect of computer-aided detection on the performance of screening mammography in community-based settings	Retrospective	Healthcare - diagnosis	159 radiologists who interpreted mammograms, of whom 122 (77%) provided complete responses and written informed consent for linkage to mammography	CAD	Non interruptive	Visual	No CAD	Use of computer software designed to improve the interpretation of mammograms was associated with significantly higher false positive rates. Diagnostic specificity decreased from 90.2% to 87.2% after implementation of CAD (p<0.001). Positive predictive value decreased from 4.1% to 3.2% (p=0.01) biopsy rates increased 19.7% (p<0.001). Any improvements were non significant.

1999	Enhancement of Clinicians' Diagnostic Reasoning by Computer-Based Consultation A Multisite Study of 2 Systems	Friedman	JAMA	To explore the extent to which consultations with DSSs improve clinicians' diagnostic hypotheses in a set of diagnostically challenging cases.	RCT	Healthcare - diagnosis	216 physicians in total: 72 at each site - including 24 medical internal medicine faculty members, 24 senior residents, 24 4th year medical students	Two DSSs, ILIAD (version 4.2) and Quick Medical Reference (QMR; version 3.7.1)	Non interruptive	Text	No DSS ("before" condition)	Correct diagnoses appeared in subjects' hypothesis lists for 39.5% of cases prior to consultation and 45.4% of cases after consultation. Positive consultations, where the correct diagnosis was present after consultation but not before, were observed for 232 cases (12.0%); negative consultations, where the correct diagnosis was present before consultation but not after, were observed in 117 cases (6.0%). The overall consultation effect (net gain) is 115 cases (5.9%).	Preconsultation performance, based on subjects' personal knowledge only, increased with experience level. The largest consultation effects were observed for the students, with smaller effects for residents and faculty. Larger consultation effects were observed in subjects using QMR.
2002	The efficacy of a computerized caries detector in intraoral digital radiography	Gakenheimer	Journal of the American Dental Association	Assessing the effectiveness of the software	Repeated measures - before-after intervention design	Dentistry	Eighteen dentists	Caries detector	Non interruptive	Visual	Visual diagnosis (before)	Overall improvement in detection rates (sensitivity (TPs)) - 70.3 to 90.5%, accuracy from 75 but there was a small (not significant) decrease in specificity (TNs) (from 88.6 to 88.3%, drop in 0.3%)	
2004	An examination of complex human-machine system performance under multiple levels and stages of automation	Galster	Dissertation Abstracts International: Section B: The Sciences and Engineering	With advances in technology increasing, it is no longer applicable to look at single automated tools but rather at how several automated tools fit together and affect system performance.		Generic HCI						The first 3 experiments utilized a visual search paradigm and varied the stage the automation was present and the reliability of the automation that was used. For these studies, the automation that helped the operator locate the potential target demonstrated a clear advantage over automation that recommended a course of action when the automation was perfectly reliable. The 4th study examined all of the possible combinations of manual and automated aiding for the 4 stages in an air-to-ground search and destroy mission that was carried out in a high fidelity combat flight simulator. By utilizing separate stage metrics, it was demonstrated that the automation in 1 stage influenced performance in subsequent stages and throughout the entire mission.	
1995	Effects of Multiple-Signal Discrimination on Vigilance Performance and Perceived Workload	Grubb	Human Factors and Ergonomics Society Annual Meeting Proceedings, Visual Performance	To explore performance on sustained attention tasks with more complex displays. successive monitoring tasks involving absolute judgments are more capacity-demanding than simultaneous tasks which are comparative in nature.	Observers monitored either one (0-bits display uncertainty), two (1-bit display uncertainty), or four (2-bits display uncertainty) indicators on a simulated aircraft display for the occurrence of critical signals presented in	Aviation						Results indicated that correct detections declined as display uncertainty increased, and that this effect was more pronounced in the simultaneous format. Moreover, workload scores increased with display uncertainty, particularly in the simultaneous condition. These findings suggest that in more complex monitoring situations in which there is a scanning imperative successive tasks may have an advantage over their simultaneous counterparts.	

					either a simultaneous or a successive format.								
1995	The antibody identification assistant (AIDA), an example of a cooperative computer support system	Guerlain	Systems, Man and Cybernetics, 1995. Intelligent Systems for the 21st Century., IEEE International Conference on	To compare performance when using a critiquing expert system to performance with no decision support for two groups of medical technologists solving antibody identification cases.	Between subjects controlled	Healthcare - diagnosis	32 certified medical technologists.	AIDA - antibody identification assistant	Non interruptive	Text	No AIDA	Clear evidence that critiquing system reduced errors, even when the criticism was not appropriate. There was 46% misdiagnosis in the control group, but only 5% misdiagnosis in the treatment group. However this 5% was found in the the weak antibody case for which AIDA's knowledge was not fully competent. IMPLICATION OF AB?	Study which implies that DSS is worth it despite a small risk of AB? /// there was a trend for improved performance even on a case for which the computer's knowledge was not fully competent. This is in contrast to the usual problems with people not being able to recover from faulty reasoning exhibited by a brittle, partially automated decision support system. Users of critiquing systems are doing the task themselves and given feedback in the context of what they are doing. Thus, the computer can monitor for errors in the human's reasoning, and the human has a basis for judging the computer's reasoning, resulting in cooperative problem-solving between the two decision makers

1993	Factors influencing the cooperative problem-solving of people and computers	Guerlain	Proceedings of the human Factors and Ergonomics Society	Study the influence of different computer system designs on cooperative problem solving performance.	Randomised. Two between subjects comparison groups - first to a critiquing system and the other to a partially automated system. Solve 2 practice cases followed by 5 test cases. In the critiquing system, participants were able to rule out antibodies one by one (as is currently done without the computer), but if a computer disagreed with any of their conclusions about what should be ruled out a brief error message was displayed. In the partially automated system, an automatic rule-out function was available, which could cause the computer to rule out all antibodies possible, given the current data for the case.	Healthcare - antibody identification	Thirty two certified blood bankers	Critiquing and partially automated systems	Non interruptive	Text	No non DSS control	Overall performance on cases where computer was competent was better for subjects using the partially automated system than those using the critiquing system (5.6% versus 11.9% mean misdiagnosis rate) - but non sig. For computer-incompetent cases, performance was significantly worse for the subjects using the partially automated system than for the critiquing system (76% versus 43%, p<0.05)	Suggests that partially automated systems can cause the practitioners to make more errors on cases where the computer's knowledge is inappropriate than if using a critiquing system. This was uniformly true regardless of practitioner level.
------	---	----------	---	--	---	--------------------------------------	------------------------------------	--	------------------	------	--------------------	---	---

2004	Improvement in Radiologists' Characterization of Malignant and Benign Breast Masses on Serial Mammograms with Computer-aided Diagnosis: An ROC Study	Hadjiski	Radiology	To evaluate the effects of computer-aided diagnosis (CAD) on radiologists' characterization of masses on serial mammograms.	Repeated measures - before-after intervention design	Healthcare - diagnosis	Eight radiologists, two breast imaging fellows	CAD	Non interruptive	Visual	No CAD	The average Az for radiologists' estimates of the likelihood of malignancy was 0.79 without CAD and improved to 0.84 with CAD. When the radiologists evaluated the temporal pairs in the sequential mode with CAD, an average (per radiologist) of 2.3% (3.2 of 138) of additional malignant masses were correctly recommended for callback and 0.6% (0.7 of 115) of additional benign masses were incorrectly recommended for callback compared with the evaluation in the independent mode. The reading in sequential mode with CAD compared with reading in sequential mode without CAD resulted in an average of 1.4% (1.9 of 138) of additional correct callbacks for malignant masses and 2.1% (2.4 of 115) of additional incorrect callbacks for benign masses.	Generally, when the radiologists used CAD, they correctly recommended additional callbacks for malignant masses but also increased the callbacks for benign masses. This indicates that the radiologists would increase their sensitivity but might also reduce their specificity when they used CAD
2004	Sensitivity of Noncommercial Computer-aided Detection System for Mammographic Breast Cancer Detection: Pilot Clinical Trial	Helvie	Radiology	To evaluate a noncommercial computer-aided detection (CAD) program for breast cancer detection with screening mammography.	Repeated measures - before-after intervention design	Healthcare - diagnosis	Thirteen radiologists	CAD	Non interruptive	Visual	No CAD	Our high 91% sensitivity was achieved with some negative consequences. These consequences included a higher recall rate, which was increased by 1.4%, and a higher biopsy rate, which was increased by 8%. These negative risks were associated with a 9% improvement in cancer detection (i.e. FPs mislead the radiologist).	
2008	Providers Do Not Verify Patient Identity during Computer Order Entry	Henneman	Academic emergency medicine	To determine the frequency of verifying patient ID during computerized provider order entry (CPOE).	Prospective, investigative (eye tracking and error recording), no a priori conditions	Healthcare - identification	Nine attending physicians, 5 Physician assistants, and 11 emergency medicine residents (Postgraduate Year 2 and 3).	CPOE	Non interruptive	Text	No control	Two of 25 (8%; 95% CI = 1% to 26%) noted the DOB error; the remaining 23 ordered tests on an incorrect patient. One of 25 (4%, 95% CI = 0% to 20%) noted the last name error; 12 ordered tests on an incorrect patient. No participant (0%, 0/107; 95% CI = 0% to 3%) verified patient ID by looking at MRN prior to selecting a patient from the alphabetical list. Twenty-three percent (45/200; 95% CI = 17% to 29%) verified patient ID prior to ordering tests.	Medical providers were asked to review 10 charts (scenarios), select the patient from a computer alphabetical list, and order tests. Two scenarios had embedded ID errors compared to the computer (incorrect DOB or misspelled last name), and a third had a potential error (second patient on alphabetical list with same last name).

1995	The Effects of Computer-Assisted Electrocardiographic Interpretation on Physicians' Diagnostic Decisions	Hillson	Medical Decision Making	To evaluate the effect of computer-assisted interpretation of electrocardiograms (ECGs) on diagnostic decision making by primary care physicians.	RCT	Healthcare - diagnosis	Forty family physicians and general internists	Clinical vignettes accompanied by ECGs and reported their diagnostic impressions.	Non interruptive	Text	Vignette (no CATI)	Overall agreement with CATI was 52% for those receiving versus 34% for those not receiving CATI. First erroneous case: Agreed with by 16/21 (with CATI) versus 8/19 (sig at $p<0.05$ ) /// Second case: 19/21 (with CATI) versus 15/19 (non CATI) (not sig, $p>0.3$ ) /// Third case: 20/21 (CATI) versus 17/19 (not sig $p>0.4$ Fisher's Exact Test)	Ss receiving CATI reports more than twice as likely to arrive at the correct clinical diagnosis BUT in one of the three misleading CATI vignettes, Ss who received erroneous reports were more likely to make a corresponding error.
2005	Age differences in trust and reliance of a medication management system	Ho	Interacting with Computers	The present study examined age differences in trust and reliance of an automated decision aid. Differences in omission and commission errors were examined.	Exp 1 and 2: Counterbalanced 2 repeated measures for high versus low AMM reliability x 2 between subjects age (old versus young)	Healthcare - medication management	Exp 1: Thirteen younger and 12 older participants (naïve students) Exp 2: Twelve younger and 12 older participants	AMM -automated medication manager	Non interruptive	Text	No non DSS control	Exp1: The results indicated that older adults had greater trust in the aid and were less confident in their performance, but they did not calibrate trust differently than younger adults. Exp 2: The results indicated that older adults were more reliant on the decision aid and committed more automation related errors. A signal detection analyses indicated that older adults were less sensitive to automation failures.	The results of Experiment 2 suggest that both age groups were susceptible to automation reliance effects. More commission and omission errors were made in the high reliability relative to the low reliability condition. Older adults were more susceptible to automation failures than younger adults. In general they made more commission and omission errors relative to younger adults. Chen and Sun (2003) and Johnson (1990) (without DSS, not relevant for this SR) have reported that older adults use simpler heuristics when engaged in cognitively demanding decision-making and this may influence their reliance on automation as well.
2000	A Signal-detection Experiment Measuring the Effect of Computer-aided Detection on Radiologists' Performance	Ikeda	Medical Decision Making	To evaluate how the specificity and sensitivity of computer-aided detection (CADe) algorithm outputs affected radiologists' diagnostic performances.	Repeated measures - control (no CADe vs 25 simulated CADe algorithms with various sensitivities and specificities (from 60% to 100%).	Healthcare - diagnosis	Six novice radiologists	CADe	Non interruptive	Text	No CADe	Found an approximately linear relationship between the sensitivity (specificity) of the CADe output and the reader's sensitivity (specificity), and the slope of reader sensitivity (specificity) as a linear function of CADe sensitivity (specificity) can be considered to be a positive number less than unity. /// A comparison of the Az value changes due to the CADe output overall accuracy between results and a predicted "ideal" reader amenable to the CADe outputs indicated that the influence of the CADe outputs on the readers' performances in this study was different from what would be expected based on the ideal reader's performance, and that the study readers did not uncritically obey the CADe outputs.	The overall accuracy of CADe outputs is the most significant factor affecting radiologists' performances in image interpretation.

1996	Effect of a Computer-aided Diagnosis Scheme on Radiologists' Performance in Detection of Lung Nodules on Radiographs	Kobayashi	Radiology	To evaluate the effect of a computer-aided diagnosis (CAD) scheme on radiologists' performance in the detection of lung nodules, and to examine a new method of receiver operating characteristic (ROC) analysis.	Observer study	Healthcare - diagnosis	Sixteen radiologists (two thoracic, six general, and eight residents)	CAD Digitised radiographs	Non interruptive	Visual	Conventional radiograph	ROC analysis. In the cases of false-positive detections with CAD, observers were not detrimentally influenced. Even when lung nodules were missed with the CAD scheme, we found that observers were generally able to find the nodules despite the potentially detrimental effect of the CAD result in this situation. This seems to indicate that observers can use CAD effectively as a second opinion and detect some nodules on chest images that were missed with CAD output.	Explicit no AB finding for FP detection - authors suggest this is probably because false-positive nodules detected with CAD output were generally different from those detected by human observers, thus, it was not very difficult for observers to disregard false-positive nodules demonstrated on digitized images.
------	--	-----------	-----------	---	----------------	------------------------	---	---------------------------	------------------	--------	-------------------------	--	---

1994	Design of a Cooperative Problem-Solving System for En-Route Flight Planning: An Empirical Evaluation	Layton	Human Factors	Case 3 was designed to present the pilots with a difficult planning problem and to put the various system designs to a demanding test.	Each pilot was randomly assigned to one of the three alternative system designs. Three different en-route flight-planning support systems were designed that represented variations on the levels and timing of support provided by the computer. 1. The sketching-only system allowed the human planner to sketch proposed flight paths on a map display while the computer filled in lower-level details 2. Route constraints and sketching system. The route constraints and sketching system retained all of the capabilities of the sketching only system and added another capability: Operators could specify higher-level constraints	Aviation	Thirty male commercial airline pilots	Computerised flight route advisor	Non interruptive	Visual	Sketching only (minimal automation contribution)	The effects of providing automatic suggestions by the computer can be quite pronounced. Subjects in Cases 1 and 3 who were presented with the computer's suggestion clearly reasoned less (or not at all) about the uncertainty associated with the forecast, leading them to accept a poor flight plan in Case 3: 40% versus 10% in control accepted a poor route plan. A number of subjects in all three conditions exhibited poor performance in Case 3. Although more subjects appeared to be biased toward a poor solution when it was suggested by the computer, this bias cannot be explained simply by overreliance in the sense of blindly accepting the computer's recommendations. These subjects showed clear evidence of generating and evaluating alternatives. Thus much deeper explanations had to be developed to account for their acceptance of the computer's poor suggestion. In addition, forcing the pilots to be more involved by making them sketch their own solutions resulted in the selection of fewer poor plans. Nevertheless, because of the use of an elimination-by-aspects strategy by one subject, he generated and selected the poor eastern deviation without any suggestions from the computer.
------	--	--------	---------------	--	---	----------	---------------------------------------	-----------------------------------	------------------	--------	--	--

					<p>on the solution they desired and then ask the computer to find the shortest route that satisfied those constraints.</p> <p>3. Automatic route constraints, route constraints, and sketching system. This version took the computer's involvement one step further: The computer automatically suggested a deviation (based on default constraints of no turbulence no precipitation and the originally planned destination) as soon as it detected a problem with the original routes.</p>								
2006	Improving radiologists' recommendations with computer-aided diagnosis for management of small nodules detected by CT	Li	Academic Radiology	To evaluate how computer-aided diagnosis (CAD) can improve radiologists' recommendations for management of possible early lung cancers on CT	Observer study	Healthcare - diagnosis	Sixteen radiologists	CAD	Non interruptive	Visual	no CAD	The number of recommendations changed by radiologists by use of CAD was 163 (18%) among all 896 observations. Among these changed recommendations, the fraction showing a beneficial effect from CAD was 68% (111/163), and the fraction showing a beneficial effect regarding biopsy recommendations was 69% (48/70).	The overall accuracy of CADe outputs is the most significant factor affecting radiologists' performances in image interpretation. Possible implication that upto 32% and 31% showed detrimental changes/ switches in decision.

2007	Situation awareness and driving performance in a simulated navigation task	Ma	Ergonomics	The objective of this study was to identify task and vehicle factors that may affect driver situation awareness (SA) and its relationship to performance, particularly in strategic (navigation) tasks.	Between subjects automated aid versus human aid (via mobile) x 3 within subjects reliability (100%, 80% and 60%)	Driving - navigation	20 students	Automated aid	Interruptive	Text	control condition was used in which each aid presented a telemarketing survey and participants navigated using a map - all subjects "before"	Results revealed perfect navigation information generally improved driver SA (SAGAT questionnaire) and performance compared to unreliable navigation information and the control condition (task-irrelevant information).	These findings suggest that perfect navigation aid reliability could negatively influence Level 1 SA in navigation driving tasks by motivating driver complacency with regard to perceiving changing states of the roadway environment. " levels of SA defined by Endsley, including perception (Level 1 SA), comprehension (Level 2 SA) and projection (Level 3 SA) /// Further investigation (described in a separate paper) - results demonstrated drivers had a higher initial trust expectation for the automated aid than the human aid. However, once the automated aid was used, trust declined dramatically across degrading aid performance conditions. In general, there was a trend for trust to decrease more sharply with the automated aid than the human aid, but there was no statistical difference among the aids. Trust declined as both aids degraded in performance. Driving errors also increased as the navigation aid performance decreased, and the control condition produced the highest number of errors. This study demonstrated the role of driver trust in in-vehicle navigation aid use and has implications for designing navigation systems that support trust and performance.
2005	Cognitive anchoring on self-generated decisions reduces operator reliance on automated diagnostic aids	Madhavan	Human Factors	The extent to which users' agreements with an aid are anchored to their personal, self-generated diagnoses was explored.	Between subjects (anchor group versus non anchor group)	Pump operation	75 participants	Automated diagnostic aid	NS	NS	Before DSS in forced anchor group	Within the nonforced anchor group, participants' self-reported tendency to prediagnose system failures significantly predicted their tendency to disagree with the aid, revealing a cognitive anchoring effect. Agreement rates of participants in the forced anchor group indicated that public commitment to a diagnosis did not strengthen this effect. Potential applications include the development of methods for reducing cognitive anchoring effects and improving automation utilization in high-risk domains.	One group (nonforced anchor, n = 50) provided diagnoses only after consulting the aid. Another group (forced anchor, n = 25) provided diagnoses both before and after receiving feedback from the aid.

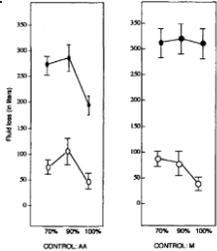
2007	Effects of information source, pedigree, and reliability on operator interaction with decision support systems	Madhavan	Human Factors	Tot examine operators' perceptions of decision aids.	<b>Exp 1:</b> 2 (source: human vs. automated adviser) × 2 (pedigree: expert vs. novice) within-subjects // <b>Exp 2:</b> information source (human or automated), pedigree (novice or expert), and reliability (low or high) of the adviser varied between subjects	Luggage-screening task	180 undergraduate and graduate students	Human or automated advisers	Interruptive	Text	None without DSS, comparison between 4 advisor types	<b>Exp 1:</b> measures of perceived reliability indicated that automation was perceived as more reliable than humans across pedigrees. Measures of trust indicated that automated "novices" were trusted more than human "novices"; human "experts" were trusted more than automated "experts." <b>Exp 2:</b> perceived reliability varied as a function of pedigree, whereas subjective trust was always higher for automation than for humans. Advice acceptance from novice automation was always higher than from novice humans. However, when advisers were 70% reliable, errors generated by expert automation led to a drop in compliance/reliance on expert automation relative to expert humans. // When advice was 70% reliable, results suggest that merely portraying the human adviser as an expert led users to agree more with the advice, regardless of accuracy. When automated aids were portrayed as experts, initial expectations were likely very high (see Study 1). When participants saw their adviser generating errors on 30% of occasions, it led to a rapid breakdown in initial expectations, generating a negative trend in dependence.	"Compliance refers to the probability of agreeing with advice when a DSS generates a diagnosis of "target present"; Reliance refers to the probability of agreeing with a diagnosis of "target absent." It is important to represent agreement as compliance versus reliance because the compliance reliance trade-off determines the types of errors generated by operators during a task, thereby providing the opportunity for the direct mapping of advice use with performance efficiency."
------	--	----------	---------------	--	---	------------------------	---	-----------------------------	--------------	------	--	---	--

2003	New alternative methods of analyzing human behaviour in cued target acquisition	Maltz	Human Factors	To assess target acquisition performance under different levels of task complexity and cuing system reliability.	Between subjects: 1 control (no automated aid) vs 9 different reliability groups (combinations of the three levels of cue-Pd (.5, .7, and .9) and the three levels of cue-FAR (0, 1, and 3).) Within subjects - task difficulty (colour vs infrared)	Military	132 undergraduate students	Automated cuing	Interruptive	Visual	Unaided control group	Observer reliance on the cue correlated with task difficulty and the perceived reliability of the cue. Cuing was generally helpful in complex tasks, whereas cuing reduced performance with the infrared (hard) pictures was a positive outcome of reliant behaviour, whereas the lowered performance with colour (easy) pictures illustrated the disadvantage of the reliant mode of behaviour. // Automated aids should only be introduced in tasks complex enough to warrant the intrusion of the cue. // Findings indicate that false cues are more detrimental to performance than are cue misses. False cues can lead to observer false alarms, whereas cue misses do not seem to decrease detection significantly, especially when the picture is "easy."	To determine the level of observer reliance on the cue, we used the cue dependency measure (CD). First, they measured the difference between d' values under conditions of correct cuing and under conditions of incorrect cuing to see if the CD was high, which would mean that the observers were relying on the cue. Once significant reliance on the cue was established, they used the measured CD values to see if reliance on the cue was further affected by other conditions. Compared with the control group's d' = 1.90, the cued groups averaged d' = 0.70 under conditions of incorrect cuing and d' = 3.39 when the cuing system was correct, F(1, 117) = 566.19, p < .0001, showing high reliance on the cues.
2004	Computer-assisted detection of pulmonary nodules: performance evaluation of an expert knowledge-based detection system in consensus reading with experienced and inexperienced chest radiologists	Marten	European Radiology	To evaluate the performance of experienced versus inexperienced radiologists in comparison and in consensus with an interactive computer-aided detection (CAD) system for detection of pulmonary nodules.	Repeated measures	Healthcare - diagnosis	Four blinded radiologists.	CAD	Non interruptive	Visual	no CAD	CAD and experienced readers outperformed inexperienced readers. Performance of reader 1+CAD was superior to single reader and reader 1+3 performances. Reader 3+CAD did not perform superiorly to experienced readers or CAD. Consensus of reader 1+CAD significantly outperformed all other readings, demonstrating a benefit in using CAD as an inexperienced reader replacement.	Authors suggest it is questionable whether inexperienced readers can be regarded as adequate for interpretation of pulmonary nodules in consensus with CAD, replacing an experienced radiologist.
2003	Effects of training operators on situation-specific automation reliability.	Masalonis	Systems, Man and Cybernetics, 2003. IEEE International Conference on		Free flight versus normal flight /// Between subjects trained versus untrained	Aviation	Air traffic controllers	Automated aircraft conflict detection aid	Interruptive		No non DSS control	"Subjective trust was lower in FF for the trained participants. Overall performance did not differ, but the trained group were more likely to detect both real and perceived conflicts (bias shift). Also, they were more likely in general to unquestioningly accept the automation's judgments (which in general was appropriate), as assessed by a new experimental scale of self-reported use-of-information. Only the non-trained group showed a relationship between subjective trust and unquestioning acceptance of the automation's judgments on the new scale."	

2000	Strict reliance on a computer algorithm or measurable ST segment criteria may lead to errors in thrombolytic therapy eligibility	Massel	Am Heart J	There is accumulating evidence that thrombolytic therapy is underused among eligible patients with acute myocardial infarction. We sought to determine whether potential errors in electrocardiographic diagnosis might be a contributing factor.	Prospective - observer	Healthcare - diagnosis	3 cardiologists	CAD	Non interruptive	Visual	No control - compared against independent gold standard	Raw agreement and agreement corrected for chance between raters for both criteria were excellent and tended to be better for interpretive compared with measured criteria (kappa = 0.89 vs 0.78, respectively). Strict reliance on measured electrocardiographic criteria alone would have resulted in overuse of thrombolysis among all 3 raters. Based on the consensus opinion, the absolute overuse of thrombolysis would have been approximately 15% (P <.0034). The computer algorithm had a specificity of 100% and a sensitivity of 61.5%. Reliance on the computerized interpretation alone would have lead to underuse of thrombolytic therapy compared with consensus opinion (21.3% vs 34. 6%; P <.005).
2004	Factors affecting performance on a target monitoring task employing an automatic tracker	McFadden	Ergonomics	To examine the extent to which performance on a task employing an automatic tracker was similar to performance on tasks employing other types of automation.	Experimental - non controlled comparative - Exp 1 : Repeated measures multifactorial (one within (task difficulty) one between measures(training type)). Exp 2: within participants design (2 variables - AT reliability and time on task)	Generic HCI	Twenty four participants	Automated tracker (94-95% reliable)	Interruptive	Visual	No proper control - comparison of different levels of reliability, training versus no AT training, comparison of varying task difficulties.	Exp 1: Rate of misassociations was 5.3-11.9% in the AT trained group and 7.2-14.8% in the manually trained group - trained had (non sig) less AT misassociation errors, misassociations increased in difficulty. Exp 2: In Moderate reliability 6.3 - 9.8% misassociations occurred, in high reliability 15.2-24.4% misassociations occurred. /// Perceived workload: correlated best with time on task (0.85) and misses (0.45) and moderately with targets tracked (-0.45) and misassociations (-0.36)

2006	Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information	McGuirl	Human Factors	To examine whether continually updated information about a system's confidence in its ability to perform assigned tasks improves operators' trust calibration in, and use of, an automated decision support system (DSS).	INDEPENDENT FACTORS: > Between subjects - Information type - fixed (one mention of system reliability) or updated (continuous updates). > Within subjects- DSS type (command or status), DSS performance (accurate or inaccurate), Task load (high or low), Ice location (tail or wing), Updated group only; (confidence level - high (89%), variable (50%) or low (25%)), Information availability - constant or on demand). DEPENDENT MEASURES: Initial and subsequent responses to buffeting, Compliance with DSS recommendation, Stall incidence, Sampling behaviour, Performance on failure detection and tracking, Pilot estimates of system accuracy.	Aviation	Two groups of 15 instructor pilots	Neural net-based decision aid that assists pilots with detecting and handling in-flight icing encounters.	Interruptive	Visual	No control condition for effect of DSS	The fixed group showed a tendency to follow system advice more often than was justified by its overall accuracy. Opposite trend for updated condition. > Fixed condition: 88% compliance rate for 70% system accuracy. > Updated group: High confidence condition: 80% compliance for 89% system accuracy; Variable confidence condition: 38% compliance for 50% system accuracy; Low confidence condition: 31% compliance for 25% system accuracy. >> People in the fixed group were more likely to initially comply with the DSS, and also more likely to remain anchored in the initial response even if the outcome suggested that the system advice was inaccurate. Odds ratios: > Initial compliance with system recommendation : If Fixed = 1, then in Updated, High = 0.51, Variable = 0.08, Low = 0.05 . > Switching to alternative recovery technique: If Fixed = 1, then in Updated, High=1.52, Variable= 4.49, Low= 2.54.
------	---	---------	---------------	---	--	----------	------------------------------------	---	--------------	--------	--	---

2006	Effectiveness of clinician-selected electronic information resources for answering primary care physicians' information needs.	McKibbin	JAMIA	To determine if clinician-selected electronic information resources improve primary care physicians' abilities to answer simulated clinical questions.	Within subjects before-after study	Healthcare	23 physicians	Online information retrieval system	Non interruptive	Text	Before - no DSS	On average 43.5% of the answers to the original 23 questions were correct. For the questions that were searched, 18 (39.1%) of the 46 answers were correct before searching. After searching, the number of correct answers was 19 (42.1%). This difference of 1 correct answer was attributed to 6 questions (13.0%) going from an incorrect to correct answer and 5 (10.9%) questions going from a correct to incorrect answer.
2005	Automation in Future Air Traffic Management: Effects of Decision Aid Reliability on Controller Performance and Mental Workload	Metzger	Human Factors	To examine the effects of an aircraft-to-aircraft conflict decision aid on performance and mental workload of experienced, full-performance level controllers in a simulated Free Flight environment.	Exp 2: Repeated measures design with control : Automation condition (levels: 1. reliable automation 2. automation failure with 2 mins to recover 3. automation failure with 4 mins to recover) and 4. manual condition	Aviation	Exp 2: Twenty active full-performance level controllers	Aircraft-to-aircraft conflict decision aid	Interruptive	Visual	Manual condition	Under reliable automation: More conflicts ( $F(1, 19)=8.14, p=0.01$ ) and more self separations ( $F(1, 19)=13.11, p<0.01$ ) were detected under automated conditions than under the manual condition. // Under unreliable automation there was a trend for better detection under manual than under automated conditions $F(1, 19)=2.40, p=0.14$ .
2001	Computed assisted detection of interval breast cancers	Moberg	Eur J Radiol	To examine interval cancer detection rate for a system of computer assisted detection (CAD) and its influence on radiologists' sensitivity/specificity in a screen-like retrospective review situation.	Repeated measures within subjects	Healthcare - diagnosis	Three screening radiologists	CAD	Non interruptive	Visual	No CAD	Although CAD specificity was low (38%) no reduction in radiologists' specificity occurred using CAD (73%, 82% and 89% without and 78%, 90% and 92% with CAD). Non-mixed reading increased radiologists' detection rate to 21, 17 and 19 interval cancers respectively. CONCLUSION: Despite sufficiently high sensitivity for CAD alone no increase in radiologist sensitivity (or decrease in specificity) occurred with CAD. Improving CAD specificity, with unaffectedly high sensitivity, should make radiologists more inclined to revise interpretations according to CAD.

2000	Adaptive automation, trust, and self-confidence in fault management of time-critical tasks	Moray	Journal of Experimental Psychology-Applied	To assess the effect of adaptive automation on performance	2 Mode of Control (manual or adaptive automated) X 2 Type of Fault (leak versus break) X 3 Level of Reliability X 3 Order of Condition /// Between subjects: level of reliability, within subjects: remaining variables	Central heating management	30 graduate and undergraduate students	Automated fault-management system	Interruptive	Text	Manual condition	<p>Root mean square error measured: At 90% reliability, not sig difference between automated and manual, at 70% reliability - When the plant is very reliable, RMSE is actually lower under M than under AA; there is little difference between the two modes of control when the reliability is 90%, and only when reliability falls to 70% does the RMSE in M become substantially greater than the mean in AA. /// When AA was only 70% reliable, more false shutdowns occurred in AA mode than in M (p = .33 vs. p = .20). The conclusion is that if the payoff structure of the task makes it important to avoid false shutdowns, then human operators should be retained, particularly if there is any unreliability in the automated fault-diagnosis systems. There is little effect of unreliability if reliability is at least 90%. Somewhere below that level, the effects of unreliability become widespread and important. Inverted "U" effect.</p>	 <p>Figure 3. Mean and standard error of fluid loss in the three-way interaction between reliability of automated fault management system, mode of control (adaptive automation (AA) and manual (M)), and type of fault. Filled circles represent breaks and open circles represent leaks.</p>
2008	Computer-aided detection in computed tomography colonography: current status and problems with detection of early colorectal cancer	Morimoto	Radiat Med	The aim of this study was to evaluate the usefulness of computer-aided detection (CAD) in diagnosing early colorectal cancer using computed tomography colonography (CTC).	Repeated measures within subjects	Healthcare - diagnosis	Three radiologists	CAD	Non interruptive	Visual	No CAD	<p>CAD decreased specificity in all three readers. CAD detected 100% of protruding lesions but only 69.2% of flat lesions. On ROC analysis, the diagnostic performance of all three readers was decreased by use of CAD. /// Currently available CAD with CTC does not improve diagnostic performance for detecting early colorectal cancer. An improved CAD algorithm is required for detecting flat lesions and reducing the FP rate. /// In terms of the diagnostic accuracy for identifying 150 negative segments correctly as negative, the average specificity was decreased from without CAD (96.0%) to with CAD (93.3%) (not sig). The average AUC for the three readers was also decreased by using CAD, from 0.944 to 0.918, and there was a statistically significant difference between AUCs obtained without and with CAD (P = 0.02). There were more detrimental effects after using CAD than beneficial effects (not sig effect though).</p>	<p>CAD had an average false-positive number of 17.1, which might contribute to the inferior diagnostic performance. Fenton et al.31 reported that CAD increased the number of false-positive diagnoses on mammograms, leading to increased recall and biopsy rates. Therefore, reducing CAD false-positive rates would be required for improving the diagnostic performance.</p>

1997	Automation bias - Decision making and performance in high-tech cockpits	Mosier	International Journal of Aviation Psychology	Investigate automation bias, a recently documented factor in the use of automated aids and decision support systems. The term refers to omission and commission errors resulting from the use of automated cues as a heuristic replacement for vigilant information seeking and processing.	Between subjects comparative	Aviation	25 glass-cockpit pilots	EICAS (Engine Indicating and Crew Altering System) with experimentally manipulated accountability	Interruptive	Text	EICAS (Engine Indicating and Crew Altering System) without experimentally manipulated accountability	Replicating Skitka et al (1996) 55% for omission rates (55% of opportunities to detect go undetected). Omission errors did not vary significantly as a function of experimentally manipulated accountability. However, subjects who internally felt more accountable were less likely to make omission errors than those who didn't. 0% commission error rate. All 21 pilots who had a false alarm (engine fire) message ultimately shut down the engine, contrary to instructions to say that this was not sufficient (other cues were necessary)	Although experimentally manipulated accountability demands did not significantly impact performance, post hoc analyses revealed that those pilots who reported an internalized perception of accountability for their performance and strategies of interaction with the automation were significantly more likely to double-check automated functioning against other cues and less likely to commit errors than those who did not share this perception.
2001	Aircrews and automation bias: The advantages of teamwork?	Mosier	International Journal of Aviation Psychology		<b>Between subjects</b> (a) 2 crew size (single or two people) (b) 3 levels of training (systems-only training, training that emphasized that they must verify automated actions and directives, or training that incorporated information about automation bias, errors people tend to make in automated contexts, and how these errors can be avoided); and <b>within subjects</b> : (c) 2 whether or not participants received a prompt to verify automated functioning each time a clearance was autoloading and each time the	Aviation	48 commercial glass cockpit pilots	Engine Indicating and Crew Alerting System [EICAS]	Interruptive	Text	Each crew received basic systems training on the mini-ACFS (mini-Advanced Concepts Flight Simulator) and served as its own control.	Overall omission error rates were slightly, but not significantly, better for crews than for solo pilots (43% vs. 52%), $F(1, 26) = .89$ , ns. Data for the solo pilots were comparable to the baseline error rate found in the previous single-pilot study (55%; Mosier et al., 1998). No further statistical analyses were performed on solo-pilot data. // No significant effects on errors were found for training type or display presence, $F(1, 16) = .14$ and $.09$ , ns, respectively, for crews. Omission error performance on the experimental legs was best predicted by performance on the control leg, $r(18) = .47$ , $p < .05$ . A significant effect was found for event, $F(5, 95) = 4.04$ , $p < .01$ , with altitude and runway being corrected more often than frequency, arrival, heading, or nav frequency. Unexpectedly, in 21% of the total of frequency, altitude, arrival waypoint, and heading events, automation discrepancies were caught and verbally acknowledged by the crews, but no corrective action was taken. With respect to the single opportunity for a commission error, all but two of the two-person crews (and all of the solo crewmembers) responded to the false engine fire EICAS event by shutting down the supposedly affected engine on go-around. On the debriefing questionnaire, pilots responded that the presence of an EICAS message by itself would not be sufficient to ensure that an engine fire was definitely present ( $M = 4.2$ , $SD = 2.2$ ). Pilots did not agree with the statement that it would be safer, in the event of only an EICAS message while performing a go-around, to shut down the supposedly affected engine rather than to retard the throttle and leave it running ( $M = 3.3$ , $SD = 2.15$ ). In 43% of the solo pilots and 74% of the two-person crews, one or both members erroneously remembered at least one other diagnostic cue as being present during the event. It is interesting to note that none of the four crewmembers (two crews) who left the engine running at idle recalled	Performance during the control leg was more predictive of later performance than any external manipulation suggesting the nature of pilot interaction with automation is, in part, a product of individual difference characteristics. Individual differences among pilots in attitudes toward automation and in automation use have been found in previous research, and they have been shown to be related to performance with automated systems. These differences are associated with the interaction between personal factors, such as trust and self-confidence, and more objective characteristics, workload, and cognitive overhead associated with automation use

					EICAS (Engine Indicating and Crew Alerting System) displayed a warning message.							any extra indicators.	
--	--	--	--	--	---	--	--	--	--	--	--	-----------------------	--

1996	Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation	Muir	Ergonomics	To examine operators' trust in and use of the automation in a simulated supervisory process control task.		Generic HCI		Automation in a simulated supervisory process control task				Results showed that operators' subjective ratings of trust in the automation were based mainly upon their perception of its competence. Trust was significantly reduced by any sign of incompetence in the automation, even one which had no effect on overall system performance. Operators' trust changed very little with experience, with a few notable exceptions. Distrust in one function of an automatic component spread to reduce trust in another function of the same component, but did not generalize to another independent automatic component in the same system, or to other systems. There was high positive correlation between operators' trust in and use of the automation; operators used automation they trusted and rejected automation they distrusted, preferring to do the control task manually. There was an inverse relationship between trust and monitoring of the automation. These results suggest that operators' subjective ratings of trust and the properties of the automation which determine their trust, can be used to predict and optimize the dynamic allocation of functions in automated systems.	
1996	Effects of adaptive task allocation on monitoring of automated systems	Parasuraman	Human Factors	Effects of adaptive task allocation on monitoring for automation failure	Before and after. Before, all three tasks were manual. Also 2 nonadaptive control groups used.	Aviation	27 non pilots	Engine status (with automation) and Manual conditions (tracking and fuel management), 2 adaptive groups	Interruptive	Visual	Before condition (no automation), and 2 nonadaptive groups	Detection rates: Automation failure was constant across all conditions (7 malfunctions per 10 minute block) Mean detection rates were 71.4%, 68.4% and 73% for the model based, performance based and nonadaptive control groups (Across groups average was 70.9%, comparable to rate of 72% found in Parasuraman (1993). See Figure 3 - automation failure detection rate was around 31-35%, blocks 1-4 (manual condition), blocks 5-9 had adaptive groups at detection rates of 53-55%, non adaptive was around 28%	Adaptive task allocation (temporary returning control to the human operator) - model based ("human control" in middle of second session) and performance based (same, but only if the participant performance did not meet the criterion)

1993	Performance consequences of automation-induced 'complacency.'	Parasuraman	International Journal of Aviation Psychology	To examine the effect of varying the reliability of a piece of automation on the ability of the human to detect automation failures.	Between subjects: automation reliability (constant versus variable)	Aviation	Exp 1 : 24 participants, Exp 2 : 16 participants (non pilots)	Automated system monitoring	Interruptive	Visual	Manual condition	From Parasuraman 1996: Participants detected more than 72% of malfunctions on the engine status task when they did the task manually while simultaneously carrying out tracking and fuel management. However, when the engine status task was under automation control, participants detected only 32% of malfunctions. // Exp 1 : Variable reliability group significantly more likely to detect automation failure (149.% difference in detection probability). Variable condition also showed the best detection improvement over blocks. No real difference between detection rates in constant reliability group (high versus low). Exp 2: Single task detection was easy and did not vary between conditions (probability of around .97 of detecting).
------	---	-------------	--	--	---	----------	---	-----------------------------	--------------	--------	------------------	--

2008	CT Colonography with Computer-aided Detection as a Second Reader: Observer Performance Study	Petrick	Radiology	To evaluate the effect of computer-aided detection (CAD) as second reader on radiologists' diagnostic performance in interpreting computed tomographic (CT) colonographic examinations	Sequential read design (i.e. before-after CAD. Plus confidence scores for decisions	Healthcare - diagnosis	Four radiologists	CAD	Non interruptive	Visual	No CAD	CAD increased sensitivity and decreased specificity in small and medium polyps. CAD increased the average sensitivity by 15% (3.2 of 21), 16% (2.5 of 16), and 14% (0.7 of 5) (smaller, medium and larger polyps respectively). With this increase came a 14% decrease in specificity associated with CAD reading.	Level of experience: The least experienced readers tended to have the strongest improvement in ROC performance. The more experienced readers showed strong improvement for the 6-9-mm group, but a reduction in performance for the 10 mm or larger range, owing to their perfect sensitivity in this group so that any false-positive result prompted by CAD, even one smaller than 10 mm, could only hurt performance. It is interesting to note that reader 2 experienced the largest increase in sensitivity (smallest polyp: 12/21 to 18 /21 = +6 (29% increase), medium polyp: 7/16 to 13/16, = +6 (38% increase) p<0.05 , as well as the largest decrease in specificity (33/39 to 23/39, = -10 (-26%)). This suggests that reader 2 was more willing to utilize the CAD information than were the other readers. CAD may assist less experienced readers in detecting larger adenomatous polyps, but the overall benefits of CAD are likely also tied to how an individual reader interacts with the CAD program.
------	--	---------	-----------	--	---	------------------------	-------------------	-----	------------------	--------	--------	--	---

2005	Individual Differences in Complacency and Monitoring for Automation Failures	Prinzel	Individual Differences Research	To examine the relationship between the individual differences of complacency potential, boredom proneness, and cognitive failure with automation-induced complacency.	Comparison of high versus low complacency potential: Between subjects : A 2 (constant or variable automation reliability) X, Within subjects: 2 (sessions) X 4 (10 min blocks) X 2 (median split of CPRS - Complacency potential) mixed factorial design was employed for these analyses.	Aviation	Forty undergraduate students	Automated system monitoring task	Interruptive	Visual	No non DSS control	Significant main effect found for reliability. Participants performing the monitoring task under the variable-reliability condition did significantly better than those participants in the constant-reliability condition. Participants with high complacency potential (HCP) in the constant reliability condition did significantly worse than participants in the other three conditions // Perceived Workload: Parasuraman, Molloy, and Singh (1993) noted that automation-induced complacency only arises under conditions of high workload. The present study suggests that perception of workload and automation-induced complacency was determined largely on the basis of whether the participant was classified as high or low in complacency potential. HCP participants in the variable reliability condition rated workload significantly higher than the LCP participants, in both the variable and constant reliability conditions, and the HCP participants in the constant reliability conditions. Those low in complacency potential did not ever trust the automation and therefore, relatively speaking didn't statistically report a difference in workload between the two reliability conditions. HCP participants, on the other hand, have a predisposition toward trusting the automation and it requires a great deal of "cognitive overhead" to decide not to trust and monitor the automation.	Theoretically, automation-induced complacent behaviours may actually improve performance on other tasks because of the "automation trust" that therein allows the automation to perform that task and frees up cognitive resources to manage other tasks. But this was not found
2003	Radiologists' detection of mammographic abnormalities with and without a computer-aided detection system	Quek	Australas Radiol	The aim of this study was to evaluate the role of a computer-aided program (CAD) in assisting detection of mammographic lesions by radiologists not specifically trained in mammography and its potential utility in breast screening.	Before, without CAD (before and after condition)	Healthcare - diagnosis	Experienced mammographers	CAD	Non interruptive	Visual	No CAD	There is overall increased sensitivity in detecting mammographic abnormalities with the aid of the CAD system from 74.4 to 87.2%, which is statistically significant. However, it failed to detect suspicious abnormalities in 71 breasts (24.1%).	The CAD system improved detection of suspicious mammographic abnormalities by radiologists who are not specifically trained in mammography. However, there is also a substantial failure to detect suspicious mammographic features that cautions against over-reliance on the system, emphasizing its role as a second reader at best.

2007	Effects of imperfect automation on decision making in a simulated command and control task	Rovira	Human Factors	To examine the differential impact of information and decision automation and to investigate the costs of automation unreliability	A4 (type of automation) × 2 (overall automation reliability) × 2 (trial reliability) within-subjects design was used. The four automation support conditions included information automation and three different forms of decision automation: low, medium, and high. Overall automation reliability was varied across two values (60% and 80%). Trial reliability referred to a correct automated assessment (reliable) versus an incorrect automated assessment (unreliable).	Military	Eighteen undergraduate students	Automated decision aid - low, med and high levels of support for sensor to shooter task	Interruptive	Text	Manual condition	A paired samples t test of engagement selection rates showed that there was no difference in decision accuracy between manual (M = 89.4%) and reliable automation (M = 88.4%), $t(17) = 0.62, p = .541$ . However, there was a significant difference in decision accuracy between the manual and the unreliable automation support conditions, $t(17) = 6.9, p < .001$ , with accuracy declining to 70% under unreliable automation. In general, there was no difference in accuracy performance between manual and reliable automation, but accuracy declined under unreliable automation. Mean accuracy rates for reliable and unreliable trials were 88.5% and 70.0%, respectively.	When automation provided an incorrect assessment the accuracy of target engagement decisions declined. When overall automation reliability was 80%, this cost of automation reliability was greater for the three decision automation support tools than for the information automation condition. Information automation was information presentation only, whereas decision automation involved different levels of recommending a decision.
2001	Supporting Decision Making and Action Selection under Time Pressure and Uncertainty: The Case of In-Flight Icing	Sarter	Human Factors	To examine the effectiveness of two different DSS implementations: status and command displays.	Independent variables. A 3 (display type) × 2 (autopilot) × 3 (location of ice accretion) × 2 (accuracy of decision aid) mixed-factorial design was used.	Aviation	Twenty-seven pilots (9 pilots each in a baseline, status, and command group)	Command or status display type DSS	Interruptive	Visual	Baseline condition with no DSS	When inaccurate information was presented, performance dropped below that of the baseline condition. The cost of inaccurate information was particularly high for command displays and in the case of unfamiliar icing conditions.	Findings suggested that unless perfect reliability of a decision aid can be assumed, status displays may be preferable to command displays in high-risk domains

1997	Automation-induced monitoring inefficiency : role of display location	Singh	International Journal of Human – Computer Studies	Can inefficient monitoring be overcome by locating the automated task centrally rather than off to the side.	Before - after	Aviation	Twelve naïve subjects	Automation routine	Interruptive	Visual	Manual condition	Manual practice: automation failure detection was around 78.54% (SD 7.81) or 74.02% (SD 9.58) (comparison of constant or variable reliability group baseline detection rates before application of automated condition) // Mean number of false alarms detected was 2-3.5 // Automated condition: In this study 80% of automation failures were detected in the variable reliability condition, and 41% in the constant reliability condition (in Parasuraman, 1993 it was 82% and 33% respectively).	Monitoring performance under automation was inferior to performance of the same task under manual conditions . Contrary to our expectations , centrally locating the monitoring display did not affect the pattern of results appreciably , indicating that the automation “complacency” effect discovered by Parasuraman et al . (1993) is a relatively robust phenomenon .
1993	Individual-differences in monitoring failures of automation	Singh	Journal of General Psychology	To investigate whether personality differences affect monitoring and automation related complacency	Between subjects: automation reliability (constant versus variable)	Aviation	24 non pilots	Automated monitoring aid	Interruptive	Visual	No non DSS control	The three personality measures were not correlated. Complacency was inversely related to detection of automation failure. No relation to intro - extraversion. High arousal subjects are better at detecting in constant reliability conditions. Suggest a modest relationship between complacency potential, energetic-arousal and automation related monitoring energy efficiency.	

2000	Automation Bias and Errors: Are Crews Better Than Individuals?	Skitka	The International Journal of Aviation Psychology	This study examined automation bias in two-person crews versus solo performers under varying instruction conditions	The study represented a $2 \times 3 \times 2$ (Crew (single versus 2 person crews) $\times$ Training (training that instructed participants that they could verify automated directives, training that emphasized that they must verify automated directives, or training that included instruction about errors people tend to make in automated contexts and how they can be avoided, as well as instructions that they could verify automated directives) $\times$ Prompt to Verify (prompt or no prompt)) three-way between-subjects experimental design. The dependent variables of interest were the number of omission and commission errors participants made across these conditions	Aviation	One-hundred forty-four students from a large Midwestern university received partial course credit for their participation in the study, yielding 48 two-person crews and 48 one-person crews.	An Automated Monitoring Aid (AMA) detected and announced all but 6 of 100 events that required responses, creating six opportunities for participants to make omission errors (i.e., failing to detect an event if not explicitly prompted about it by the AMA). Similarly, the AMA gave an inappropriate directive six times (e.g., indicating that a gauge was in a red zone when in fact it was not), providing six opportunities for commission errors - thus 88% reliable	Interruptive	Text	No prompts. Multifactorial (see Design column)	Omission errors: Descriptively, 51% of the participants made one or more omission errors, and almost 30% made three or more. On average, participants made 1.85 omission errors out of a total of six possible errors, regardless of experimental condition. /// Commission errors: On average, participants made 3.25 commission errors out of a possible 6 (SD = 1.88), and almost 80% made 2 or more commission errors. An examination of the number of commission errors as a function of crew size, prompts to verify, training, and trial order indicated that only training affected the number of commission errors participants made, $F(2, 84) = 3.64, p < .05, \eta^2 = .08$ . In other words, 8% of the variance in commission errors could be accounted for by the training manipulation (an effect size that Cohen, 1977, would categorize as above a medium effect size). Tukey tests indicated that the group that was explicitly trained about automation bias and resultant omission and commission errors made fewer commission errors ( $M = 2.59, SD = 1.72$ ) than either the could-verify training group ( $M = 3.84, SD = 1.61$ ) or the must-verify group ( $M = 3.31, SD = 2.12$ ).	Training that focused on automation bias and associated errors successfully reduced commission, but not omission, errors. Teams and solo performers were equally likely to fail to respond to system irregularities or events when automated devices failed to indicate them, and to incorrectly follow automated directives when they contradicted other system information.
------	--	--------	--	---	---	----------	---	--	--------------	------	--	---	---

1999	Does automation bias decision-making?	Skitka	International Journal of Human-Computer Studies	To compare error rates in a simulated flight task with and without a computer that monitored system states and made decision recommendations.	Between subjects: AMA versus no AMA	Aviation	Eighty undergraduate students	AMA (automated monitoring aid)	Interruptive	Text	Manual condition	Participants in the manual condition outperformed counterparts with a very but not perfectly reliable automated aid on a monitoring task. Participants with an aid made omission and commission errors /// Respondents in the automated condition on the whole underestimated the reliability of the AMA which was 94% reliable Participants on average believed the AMA was only 81.35% reliable. /// Omission errors: participants in the automated condition missed more of these events (M=2.44 or a 59% accuracy rate) than those in the non-automated condition (M=0.18 or a 97% accuracy rate), /// Commission errors: Analysis of participant responses across the six commission error events indicated that on average, participants made 3.92 / 6; an average accuracy rate of 35%. Only one participant made no commission errors; 23.1% of the participants made commission errors on all six events. Results indicated that omission errors a likely occurrence in automated contexts, and that commission errors are highly probable events as well.	
2000	Accountability and automation bias	Skitka	International Journal of Human-Computer Studies	To explore the extent to which errors of omission and commission can be reduced under conditions of social accountability.	Between subjects: 2 x accountability (high versus low) i.e. high: had to justify their performance, low: performance not analysed or justified	Aviation	181 undergraduates	AMA (automated monitoring aid)	Interruptive	Text	No non DSS control	Participants in the high accountability condition committed significantly fewer commission and omission errors and a higher rate of advice verification. This did not come at a price in response time or tracking performance (no significant differences). Results indicated that making participants accountable for either their overall performance or their decision accuracy led to lower rates of "automation bias". <b>Questionnaire for subjective opinions</b> results: Errors of <b>omission</b> : Result of cognitive vigilance decrements; errors of <b>commission</b> proved to be the result of a failure to take into account information and a belief in the superior judgement of automated aids.	Despite participants being explicitly aware that their gauges always provided 100% reliable and accurate information, they still sometimes contradicted this information in favour of AMA advice.

2009	The Effect of Erroneous Computer Interpretation of ECGs on Resident Decision Making	Southern	Medical Decision Making	To examine the effect a computer mis-interpretation might have on resident physician ECG interpretation and decision-making.	RCT	Healthcare - diagnosis	110 Physicians	CI of atrial fibrillation	Non interruptive	Text	No CI support	The overall reading of the ECGs (Diagnostic vs. Non-Diagnostic or Normal) did not differ significantly between the two groups ( $p = 0.62$ ). The 56 residents with the erroneous CI reading recommended urgent revascularization more frequently than the 49 residents without the erroneous CI reading (30% vs. 10% $p = 0.01$ )
2006	A Comparison of Medication Administrations Errors Using CPOE Orders vs. Handwritten Orders for Pediatric Continuous Drug Infusions	Sowan	AMIA 2006 Symposium Proceedings	To test nurse's ability to detect medication administration errors by comparing CPOE orders with handwritten orders for pediatric continuous drug infusions. To compare the time required to detect errors using each method, and to assess user satisfaction with each method.	Two way between subjects - controlled	Healthcare - ordering	Pediatric ICU nurses	CPOE	Non interruptive	Text	Handwriting	Nurses checked a total of 108 infusions of which 72 were programmed with an error and 36 were correct. Of the incorrect infusions, nurses failed to identify the errors in 38 of 72 infusions (53%) using CPOE orders, compared to 29 of 72 infusions (40%) using handwritten orders, $p = .07$ . Of the 36 correct infusions, nurses correctly identified all using the CPOE orders and 35 of 36 using the handwritten orders. Nurses required less time to check the infusions using the CPOE orders (6 minutes + 2.5 minutes) as compared to the handwritten orders (9 minutes + 3 minutes), $p = .0001$ . Nurses who used their calculator rather than the dose-rate table in the CPOE orders tended to commit more dosage checking errors, $p = .06$ . The user-satisfaction survey indicated more satisfaction with the CPOE orders compared to the handwritten orders, $p = .0001$ .
2003	Computer Decision Support as a Source of Interpretation Error: The Case of Electrocardiograms	Tsai	JAMIA	To determine the effect that the computer interpretation (CI) of electrocardiograms (EKGs) has on the accuracy of resident (noncardiologist) physicians reading EKGs.	RCT	Healthcare - diagnosis	Thirty internal medicine residents who were either in their second or third years of training.	Electrocardiogram (EKG) expert system.	Non interruptive	Text	No CI support	Subjects erroneously agreed with the incorrect CI more often when it was presented with the EKG 67.7% (57.2% to 76.7%) than when it was not 34.6% (23.8% to 47.3%; $p, 0.0001$ ). // The CI was Incorrect in 12 of 54 findings. Without the CI, 102 of 180 (56.7%; 48.5% to 64.5%) subject findings were interpreted correctly; this decreased to 87 of 180 (48.3%; 40.4% to 56.4%) subject findings when the CI was included

2008	The use of computer-aided detection for the assessment of pulmonary arterial filling defects at computed tomographic angiography	Walsham	J Comput Assist Tomogr	To validate a computer-aided detection (CAD) tool for the detection of pulmonary arterial filling defects at computed tomographic pulmonary angiography (CTPA) and to assess its benefit for readers of different levels of experience	Within subjects controlled	Healthcare - diagnosis	Three readers	CAD	Non interruptive	Visual	No CAD	Computer-aided detection did not influence the most experienced reader (a chest fellow). Although CAD improved the subjective confidence of the second-year resident in some cases, it had no influence on overall interpretation or accuracy. Computer-aided detection improved accuracy only for the most inexperienced reader, helping this reader to identify 9 emboli not initially appreciated.	Twenty-one studies (21%) were positive for pulmonary embolism. Of these, 18 were true positive on a case basis, and 3 were false negative. Of the 79 negative studies, 16 were true negative with no CAD marks, and the remaining 63 were FP. On a case basis, CAD sensitivity was 86%, specificity was 20%, negative predictive value was 84%, and positive predictive value (PPV) was 22%.
2005	Do Online Information Retrieval Systems Help Experienced Clinicians Answer Clinical Questions?	Westbrook	J Am Med Inform Assoc	To assess the impact of clinicians' use of an online information retrieval system on their performance in answering clinical questions.	Within subjects before-after study	Healthcare	75 clinicians	Online information retrieval system	Non interruptive	Text	Before - no DSS	System use resulted in a 21% improvement in clinicians' answers, from 29% (95% confidence interval [CI] 25.4–32.6) correct pre- to 50% (95% CI 46.0–54.0) post-system use. In 33% (95% CI 29.1–36.9) answers were changed from incorrect to correct. In 21% (95% CI 17.1–23.9) correct pre-test answers were supported by evidence found using the system, and in 7% (95% CI 4.9–9.1) correct pre-test answers were changed incorrectly. For 40% (35.4–43.6) of scenarios, incorrect pre-test answers were not rectified following system use. Despite significant differences in professional groups' pre-test scores [family practitioners: 41% (95% CI 33.0–49.0), hospital doctors: 35% (95% CI 28.5–41.2), and clinical nurse consultants: 17% (95% CI 12.3–21.7; $\chi^2 = 29.0$ , $df = 2$ , $p < 0.01$ )], there was no difference in post-test scores. ( $\chi^2 = 2.6$ , $df = 2$ , $p = 0.73$ ).	To examine changes in the direction of answers pre- and post-test, scenario answers were categorized using the classification below. Wrong Wrong (WW): Wrong answer before online information retrieval system use and wrong answer after system use [system did not help] Wrong Right (RW): Wrong answer before but right answer after [system helped] Right Wrong (RW): Right answer before but wrong after [system leads to error] Right Right (RR): Right answer before and right after use [system possibly helped to confirm answer]

2007	Dual-task performance consequences of imperfect alerting associated with a cockpit display of traffic information	Wickens	Human Factors	To examine performance consequences related to integrating an imperfect alert within a complex task domain	Exp 1: Tracking difficulty (stable vs. unstable), alert modality (visual vs. auditory), and alert type (binary vs. likelihood) were manipulated within subjects. Exp 2: The method for Experiment 2 was identical to that of Experiment 1 except that the ratio of automation FAs to misses was 4:1 (16 FAs out of 40 nonconflict trials, and 4 misses out of 40 conflict trials) instead of 1:1 (higher FA rate than Exp 1)	Aviation	Exp 1: Twelve student pilots Exp 2: new 12 student pilots	Traffic Alert and Collision Avoidance System [TCAS]	Interruptive	Visual and auditory	No non DSS control	As the alerting system became more prone to false alerts, pilot compliance decreased and concurrent performance improved. Results indicate that with high-reliability automation (reliability that is greater than that of the human alone), total system performance is improved above the capabilities of the human alone (but less than total dependence on the automated system would dictate). Then, as reliability degrades, humans also become less dependent, but even as reliability drops below a threshold at around $r = .75$ ( $d' = 1.35$ ), humans may continue to depend on the imperfect diagnostic automation, even if their performance would be better if this automated advice were ignored. Auditory alerts are more attention grabbing - improving performance in exp 1 (less FAs) but decreasing performance when FAs were more frequent. With the higher threshold setting in Experiment 1, the likelihood alert appeared to engender more dependence on automation (increased reliance and compliance) - not good with imperfect automation.	There is justification for increased false alarm rates, as miss-prone systems appear to be costly. The 4:1 false alarm to miss ratio employed here improved accuracy and concurrent task performance. More research needs to address the potential benefits of likelihood alerting.
------	---	---------	---------------	--	---	----------	--	---	--------------	---------------------	--------------------	--	---

2000	Workload and Reliability of Predictor Displays in Aircraft Traffic Avoidance	Wickens	Transportation Human Factors	To examine the effect of imperfect automation with an overall reliability of 83% on pilots using a CDTI	Exp 1: A 3 × 3 × 3 factorial, within-subjects design was used. The factors of interest were display type (BL, IP, TV), vertical traffic geometry (ascending, level, descending (level is easiest)), and longitudinal geometry (45°, 90°, and 135°). The order in which pilots saw the three display types was counterbalanced across Sessions 1 and 2. The order of the different conflict geometries was randomized. Exp 2: A 2 × 3 × 3 × 2 factorial mixed design was used. Display type (straight line or W) was varied between subjects, and vertical traffic geometry (ascending, level, descending) longitudinal geometry (45°, 90°, 135°) and trial predictor accuracy (correct, error), were varied	Aviation	Exp 1: 15 licensed flight instructors Exp 2: 20 licensed pilots	CDTI (cockpit display of traffic information)	Interruptive	Visual	No non DSS control	In Experiment 2 the consequences to performance and visual attention if prediction is occasionally in error was examined. Hypothesis: trust is related to the relative allocation of attention between the predictor symbol and the raw data of actual aircraft state. Such unreliability damages performance to some extent, particularly when the unreliable predictor forecasts more complex conflict geometry. This cost reveals the substantial allocation of attention to the predictor symbol. However, pilots, knowing the level of unreliability, appear to be relatively well calibrated in their allocation of attention between the 2 information sources. // They found that pilots appeared to be able to partially compensate for the imperfect automation, but procedural differences between the imperfect (experiment 2) and perfect (experiment 1) conditions prevented direct comparisons of performance between the two groups. They also found that the cost of erroneous automation (time spent in a predicted loss of separation, as well as deviation from the prescribed flight path) relative to correct prediction trials was more pronounced on difficult than on easy trials.
------	--	---------	------------------------------	---	---	----------	--	---	--------------	--------	--------------------	---

					within subjects.								
1999	Unreliable automated attention cueing for air-ground targeting and traffic maneuvering.	Wickens	Proceedings of the 43rd Annual Meeting of the Human Factors & Ergonomics Society.	To examine performance when pilots' attention is occasionally directed to inappropriate or inaccurate locations in space, replicating the effects of imperfect automation	Exp 1: Air-ground targeting Exp 2: pilots in a free flight simulation are engaged in a series of traffic conflict avoidance maneuvers, using a cockpit display of traffic information (CDTI). On rare trials the CDTI knowledge of the traffic intruder's intentions, reflected in a predictor	Aviation	Pilots	Target identification	Interruptive			Exp 1: Target cueing, based upon semi-reliable sensor information, sometimes directs attention away from the true target. Yet pilots follow such guidance, even knowing its unreliability, a result of the difficulty of the unaided task. Exp 2: Pilots' avoidance behaviour is governed by the predictor symbol (despite occasional unreliability), and a display manipulation that calls attention to the inaccuracy of the predictor does little to influence pilots' reliance upon the predictor symbol although it does reduce visual workload. The data are interpreted in terms of appropriate trust calibration.	

					symbol, is unreliable and does not correspond with the actual aircraft behaviour.								
2002	Agreeing with automated diagnostic aids: A study of users' concurrence strategies	Wiegmann	Human Factors	In the present study, users' tendencies to either agree or disagree with automated diagnostic aids were examined under conditions in which (a) the aids were less than perfectly reliable but aided-diagnosis was still more accurate than unaided diagnosis; and (b) the system was completely opaque, affording users no additional information upon which to base a diagnosis.	Repeated measures - all participants carried out the same task and questionnaires	Pump operation	50 undergraduate students	Pump failure diagnostic aid	Non interruptive	Text	No non DSS control	The results revealed that some users adopted a strategy of always agreeing with these aids. Only four (8%) of the 50 participants in this study concurred with the aids on every trial. Exactly half of the participants (50%) disagreed with an aid before either aid had provided a wrong diagnosis. However, only a small number of participants (7, or 14%) disagreed with an automated diagnostic aid on the very first testing trial. These findings suggested that participants differed in the type of automation utilization strategy that they adopted in this experiment. /// The distribution of agreement scores of participants in the high-concurrence group was consistent with a maximization strategy. Specifically, average agreement scores ( $M = 95.93\%$ , $SD = 2.86$ ) were relatively high and stable across testing trials. As a result, the frequency of correct diagnoses of system failures ( $M = 78.06\%$ , $SD = 4.09$ ) approached the maximum accuracy score obtainable, given the 80% reliability of the diagnostic aid and the lack of any additional information upon which to base a diagnosis. /// In contrast, agreement scores in the low-concurrence group generally reflected the use of a probability matching strategy. Agreement scores of participants in the low-concurrence group averaged roughly 65% during the first 10 trials and then gradually reached a plateau of about 84% during the latter half of testing. Across all testing trials, average agreement scores of participants in the low-concurrence group ( $M = 81.65\%$ , $SD = .11$ ) were similar to, and did not differ significantly from, the 80% reliability level of the diagnostic aids.	Those participants who adopted the maximization strategy generally agreed with the aids across most of the trials, which optimized their number of correct diagnoses. In contrast, participants who adopted the probability-matching strategy agreed with the aids on roughly 65% of the trials early during testing and on 80% of all testing trials. The probability-matching strategy, therefore, resulted in lower accuracy scores than was maximally possible. Apparently, participants who adopted the maximization strategy initially trusted the aids and were less affected by aid failures. In contrast, those who adopted the matching strategy may have had lower initial levels of trust, which they then adjusted to match actual aid reliabilities.

2005	Expert decision support system use, disuse, and misuse: a study using the theory of planned behaviour	Workman	Computers in Human behaviour	This empirical study used theory of planned behaviour to formulate hypotheses about the use, disuse, and misuse of an expert system decision support (EDSS) technology.	Retrospective and questionnaire	Finance - telecommunications networking	209 randomly selected network engineers	Expert system decision support (EDSS) technology	Non interruptive	Text	No non DSS control	Hypothesised that errors will be positively associated with EDSS misuse. H4b stated that employees who adhered to the EDSS recommendations with more frequency would have fewer induced errors than employees who more frequently disregarded the recommendations. Again, this hypothesis was supported. Because social influence may elevate furtive behaviour, for hypotheses H5 and H6, interactions were proposed. Hypothesis 5 (H5) suggested that attitude and subjective norms would be associated in such a way that more negative attitudes and greater subjective norms would correspond with greater incidence of EDSS misuse. The interaction was significant /// Hypothesis 6 (H6) made a similar assertion as H5 only with regard to perceived control. It suggested that perceived control and subjective norms would be associated in such a way that more perceived control and greater subjective norms would correspond with EDSS misuse. The interaction term was not significant	The theory of planned behaviour framework further asserts that beliefs predicate intentions, which predicate behaviours, and while some attenuation is expected, intentions are immediate precursors of behaviour and thus are highly predictive of whether or not people will perform a task
1989	Lessons learned from the field trial of ACORN, an expert system to advise on chest pain.	Wyatt	In: Barber B, Cao D, Qin D, eds. Proc. Sixth World Conference on Medical Informatics, Singapore. Amsterdam: North Holland 1989: 111-115	ACORN (management of A&E chest pain) field trial to see how feedback influenced doctors' decisions	RCT	Healthcare - diagnosis	Doctors	ACORN - cardiac support device	Non interruptive	Text	No DSS	Strict criteria (ACORN vs control): FN (8/25, 32% vs 5/17, 29%), FP (10/54, 19% vs 6/50, 12%), Crude Accuracy (61/79, 77% vs 56/67, 84%)	Patient management is negatively affected by the use of ACORN compared to the control (not statistically significant however) /// strong circumstantial/implicating evidence for AB, but the fall in performance is not directly analysed within the context of ACORN accuracy /// Comparison of assessors advice and ACORN advice revealed ACORN was around 81% accurate (crude); FP - 8/54=15%, FN 7/25=28%

2007	Effects of conflict alerting system reliability and task difficulty on pilots' conflict detection with cockpit display of traffic information	Xu	Ergonomics	To investigate the effects of conflict alerting system reliability and task difficulty on pilots' conflict detection with cockpit display of traffic information	Matched (from a previous "baseline" experiment), repeated measures	Aviation	Twenty four pilots	Automated alerting system	Interruptive	Visual	No alerting system (previous baseline trial)	Roughly half the pilots depended on automation to improve estimation of miss distance relative to the baseline pilots, who viewed identical trials without the aid of automated alerts. Moreover, they did so more on the more difficult traffic trials resulting in improved performance on the 83% correct automation trials without causing harm on the 17% automation-error trials, compared to the baseline group. The automated alerts appeared to lead pilots to inspect the raw data more closely. While assisting the accurate prediction of miss distance, the automation led to an underestimate of the time remaining until the point of closest approach. The results point to the benefits of even imperfect automation in the strategic alerts characteristic of the CDTI, at least as long as this reliability remains high (above 80%). ///	However, the results were a little surprising in that even on the automation error trials performance was no worse than its level had been in the baseline experiment and sometimes showed a hint of being better. That is, unlike other findings, erroneous automation did not yield a 'complacency cost' of over-dependence, corresponding to an automation-induced beta shift (e.g. Yeh and Wickens 2001, Maltz and Shinar 2003, Metzger and Parasuraman 2005). //// The current results did reveal three important departures from the anticipated findings. First, in contrast to the predictions of hypothesis 4, it was found that 'bad errors' were no worse than 'modest errors'. As noted above, the authors believe that the potential cost of higher automation error magnitude was mitigated by pilot strategy, whereby the sounding of an alert led to a closer scrutiny of the raw data, rather than a simple dependence on the automation's advice to dictate the pilot's response (beta shift). Indeed, it is possible that the most urgent (level 3) level of alert led to an even closer inspection than the modest (level 2) alert.
------	---	----	------------	--	--	----------	--------------------	---------------------------	--------------	--------	--	--	---

2001	Display Signaling in Augmented Reality: Effects of Cue Reliability and Image Realism on Attention Allocation and Trust Calibration	Yeh	Human Factors	To examine the relationships among three advanced technology features (presentation of target cuing, reliability of target cuing, and level of image reality and the attention) and trust given to that information.	The manipulations of cue reliability (100% vs. 75%) and interactivity (active vs. passive viewing) were imposed between participants. The manipulations of scene detail (high vs. low) and target type (cued vs. uncued targets, high vs. low expectancy) were imposed within participants. The detection distance and accuracy data were analyzed using a 2 (reliability: 100% vs. 75%) × 2 (interactivity: active vs. passive) between-subjects × 2 (scene detail: high vs. low) × 2 (cuing: cued vs. uncued) × 3 (target type) within-subject ANOVA.	Military - aviation	16 military personnel	Automated target cuing	Interruptive	Visual	No target cuing	Participants were generally less sensitive when cuing symbology was available to aid them in the target detection task than when it was not. More important, as participants' sensitivity decreased, their response criterion shifted so that responses were riskier when they believed that the cuing information was reliable. However, when participants were presented with repeated instances of the automation failure (following Block 4), their sensitivity (and trust in the system) was recalibrated. Sensitivity improved, but not to the level originally seen with no cuing whatsoever. Their response criterion was also adjusted to show a reduced willingness to report a target. However, they were still somewhat guided by the advice of the cue, as witnessed by the lower and therefore riskier $\beta$ (0.77) setting in the cued than in the uncued condition ///	The effect of scene realism on reliance on cuing information, specifically when the cuing information failed the first time; the response criterion changed little with realism when targets were cued reliably; however, when the cuing information became less reliable (75%) or unavailable (uncued), the data reveal a progressive trend toward a more conservative bias, particularly with a highly realistic scene. That is, as the attentional guidance became less informative, participants were more likely to examine the raw data underlying the cue in the high-detail scene than in the lowdetail scene. This trend toward a more conservative response criterion in the high-detail scene as the cue became less informative was confirmed by subjective ratings, which revealed that participants trusted the unreliable cuing information less (and hence were less likely to rely on it) when it occurred in the high-detail scene than in the low-detail one.
------	--	-----	---------------	--	---	---------------------	-----------------------	------------------------	--------------	--------	-----------------	--	--

2001	Soft-Copy Mammographic Readings with Different Computer-assisted Detection Cuing Environments: Preliminary Findings	Zheng	Radiology	The purpose of our study was to assess the performance of radiologists in the detection of masses and microcalcification clusters on digitized mammograms in a CAD environment after modulating cuing sensitivity levels and false-positive rates.	Repeated measures - within subjects	Healthcare - diagnosis	Seven board-certified radiologists with a minimum of 3 years experience	CAD	Non interruptive	Visual	No CAD	The expectation that observers can readily and easily discard most false positive cues regardless of their presentation or prevalence was not found. Both true- and false-positive cues affected the results. Highly accurate cuing (ie, 90% sensitivity and 0.5 false-positive cue per image) helped the observers to improve their performance, compared with the noncued environment. As the accuracy of the cuing decreased, so did the performance of the typical observer. The study results clearly indicate that poorly performing CAD can result in significant degradation of observer performance. As CAD cuing sensitivity was reduced to 50%, the average number of missed abnormalities in noncued areas increased significantly. More important, approximately 30% of these regions were detected by the radiologists in mode 1 (no CAD).
------	---	-------	-----------	--	-------------------------------------	------------------------	---	-----	------------------	--------	--------	--

## Appendix D: Invitation email



### PRIMARY CARE PRESCRIBING - CLINICAL DECISION SUPPORT SYSTEM STUDY

The Centre for Health informatics at City University, London would like to invite you to take part in our research study. This study aims as to find out the usefulness of decision support for Primary Care prescribers. The study should take no more than 30 minutes overall, and will take place in one sitting. **All participants will have a chance to win a 32GB iPod Touch and £100 to donate to a charity of their choice.**

You will be asked to view 20 clinical prescribing scenarios, for which you will be asked to provide your opinion on the appropriate prescription before and after the Decision Support System provides you with prescribing advice.

If you are interested, please register at [chivm soi.city.ac.uk/dsssolution/register.aspx](http://chivm soi.city.ac.uk/dsssolution/register.aspx).

You will provided with a unique password, with which you will be able to sign in at [chivm soi.city.ac.uk/dsssolution](http://chivm soi.city.ac.uk/dsssolution). Here you can view the Participant Information Sheet and Consent Form, and then view and give your opinion on the 20 prescribing scenarios.

We hope you find this study interesting and useful. Please contact our researcher Kate Goddard at [kate.goddard.1@city.ac.uk](mailto:kate.goddard.1@city.ac.uk) with any further questions or comments you may have.

All information given will be totally confidential, anonymised and only given for the purposes of this study. This study has been granted a favourable ethical opinion by the London Bentham Committee.

Best regards

Kate Goddard

PhD Researcher  
Centre for Health Informatics  
City University, London  
<http://www.city.ac.uk/informatics/school-organisation/centre-for-health-informatics>  
<http://www.city.ac.uk/informatics>

## Appendix E: Information sheet

**Title: Electronic prescribing decision support systems** (sponsored by, and carried out at City University, London).

We would like to invite you to take part in our research study. This study aims as to find out the usefulness of decision support systems in helping prescribing decisions for Primary Care prescribers. We hope in doing so to help better inform the design of electronic prescribing decision aids.

The study should take no more than 30 minutes overall, and will take place in one sitting - you cannot return to previous pages, and once you log out you will not be able to log in again.

It is up to you to decide to join the study which is described in the information sheet below. If you agree to take part, we will then ask you to sign a consent form. You are free to withdraw at any time, without giving a reason.

At the end of the study you will be asked if you would like to take in a *prize draw for £100* to donate to a charity of your choice and a *32GB iPod*. The draw will take place once the study has been completed.

### Procedure

You will be asked to view 20 clinical prescribing scenarios. Once you have read the scenario, please give a prescription (and/or appropriate management) for the condition – as far as possible please include the *drug, dosage, and frequency*. *[Those participants under time conditions will have the addition of “You will have 30 seconds to make each decision”].*

The decision support system will then give advice – you can then choose to follow or ignore this advice. You will also be asked for your confidence in these decisions. You have the option to revise your prescription on viewing the advice given.

Only relevant information about the patient is given, and you can assume the preliminary diagnosis is correct. For any information that is not mentioned you may assume that the findings are not divergent. For example: if you want to know the temperature of the patient and it is not mentioned you may assume that it is normal.

DISCLAIMER: As always, some advice given by the decision support system may be incorrect.

### Research Ethics

All data will be stored will be anonymous and confidential and will be used solely for the purpose of this study, with only the researchers having access to data. Overall results may form sections of submitted papers to peer reviewed journals. All email addresses will be destroyed after the prize draw.

This study has been reviewed and given a favourable opinion by the London Bentham Research Ethics Committee.

Participation is entirely voluntary and participants can withdraw at any stage without giving a reason for doing so. Data collected may still be used, but you have the right to ask any data given be withdrawn. We will contact you further only to debrief you about the study.

If you would like to be informed of the results of the study, please contact the researcher below.

If you have any questions or comments about the research, please contact Kate Goddard, the researcher at:

Email: [kate.goddard.1@city.ac.uk](mailto:kate.goddard.1@city.ac.uk)

Telephone: 0207 040 8435 (please ask for Kate Goddard)

Centre for Health Informatics

City University London

Northampton Square

London  
EC1V 0HB

If there is any concern about the project, you may also contact the BMA support line on 08459 200 169 ([http://www.bma.org.uk/doctors\\_health](http://www.bma.org.uk/doctors_health)).

**The City University London complaints clause:**

If you would like to complain about any aspect of the study, City University London has established a complaints procedure via the Secretary to the Senate Research Ethics Committee. To complain about the study, you need to phone 020 7040 3040. You can then ask to speak to the Secretary of the Senate Research Ethics Committee and inform them that the name of the project is: Electronic prescribing decision support systems

.....

You could also write to the Secretary at:

Anna Ramberg  
Secretary to Senate Research Ethics Committee  
CRIDO  
City University London  
Northampton Square  
London  
EC1V 0HB  
Email: [Anna.Ramberg.1@city.ac.uk](mailto:Anna.Ramberg.1@city.ac.uk)

[Button to proceed to next page]

## Appendix F: Consent form

### 1. Consent Form

**Project Title:** Electronic prescribing decision support systems

**Researcher:** Kate Goddard, City University, London (kate.goddard.1@city.ac.uk)

I confirm that I have read the Information Sheet (at [chivm.soi.city.ac.uk/dsssolution](http://chivm.soi.city.ac.uk/dsssolution)).

I have had the opportunity to consider the information, ask any questions I may have and have had these answered satisfactorily.

I agree that I will check my email for the study debrief immediately after the study

#### Data Protection

I understand that any information I provide is confidential, and that no information that could lead to the identification of any individual will be disclosed in any reports on the project, or to any other party. No identifiable personal data will be published. The identifiable data will not be shared with any other organisation.

#### Withdrawal from study

I understand that my participation is voluntary, that I can choose not to participate in part or all of the project, and that I can withdraw at any stage of the project without being penalised or disadvantaged in any way.

I agree to read my emails immediately after this study finishes; the researchers will be sending debriefing information which is important to read.

[Button to return to Information Sheet page]

**[ALL BOXES MUST BE TICKED TO PROCEED]**

**I have read and understand the information and consent pages and agree to take part in this study**

**[Proceed to Demographics]**

**I do not wish to participate in this study**

**[Take to a page to thank the participant for their interest, and please come back at any time, or refer people you might think may be interested]**

## Appendix G: Debrief information

**Thank you for taking part in this study.**

### **Purpose of the Study**

This study is being carried out to investigate the effect of automation bias in healthcare. Automation bias is the tendency to over-rely on automated advice, even if the advice is incorrect. In this study hypothetical scenarios were accompanied by pieces of advice from a simulated clinical decision support tool.

Through the scenarios you have seen, occasional pieces of incorrect advice from the simulated decision support tool were given to assess the effect incorrect advice has on prescribing decisions. **PLEASE ENSURE YOU ARE AWARE OF THE INCORRECT ADVICE GIVEN DURING THE COURSE OF THIS STUDY.**

The scenarios with the incorrect advice are given below:

[List of scenarios with incorrect advice]

[Please tick this box to confirm that you have read and understand the purpose of the study.

**End screen only, not in debrief email**

This is part of ongoing research into the effect of automated interventions on clinical decision making. We would like to ask you to forward the initial study invitation email on to FIVE colleagues who may be interested in this research (please do not reveal the full nature of the study). This will encourage increasingly valuable results.

**If you have any questions or comments, please contact Kate Goddard at**

**[kate.goddard.1@city.ac.uk](mailto:kate.goddard.1@city.ac.uk).**

**Recent paper:** Goddard K, Roudsari A, Wyatt JC (2011) Automation Bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association.

Online at: <http://jamia.bmj.com/content/early/2011/06/16/amiajnl-2011-000089.abstract>

## Appendix H: City University ethical approval



**CITY UNIVERSITY  
LONDON**

City Research Development & International Relations Office

Northampton Square  
London, EC1V 0HB  
United Kingdom  
T +44 (0)20 7040 5060  
F +44 (0)20 7040 3803  
[www.city.ac.uk](http://www.city.ac.uk)

Professor Abdul Roudsari  
Ms Kate Goddard  
Centre for Health Informatics  
Department of Computing  
School of Informatics  
City University London  
London  
EC1V 0HB

17 February 2011

Dear Professor Roudsari and Ms Goddard

**Reference:** 10/03/K  
**Project Title:** Electronic prescribing decision support system  
**Start Date:** 01 March 2011  
**End Date:** 31 January 2012  
**Approval Date:** 15 February 2011

I am writing to you to confirm that the research proposal detailed above has been granted formal approval from the City University Senate Research Ethics Committee following Chair's action to approve the proposal.

Please note that you are required to submit an end of study report within 90 days of the conclusion of the study or within 15 days of early termination. The end of study report form can be found on the research ethics website. You are also required to notify the Committee of any amendments made to this study. If there are significant alterations to the protocol you may need to reapply.

Should you have any further queries relating to this matter then please do not hesitate to contact me. On behalf of Senate Research Ethics Committee I do hope that the project meets with success and many thanks for your patience.

Kind regards

A handwritten signature in black ink, appearing to read 'Anna Ramberg', written over a white background.

Anna Ramberg  
Research Development Manager  
Secretary to Research Ethics Committee

Email: [Anna.Ramberg.1@city.ac.uk](mailto:Anna.Ramberg.1@city.ac.uk)  
Tel: 020 7040 3040

## Appendix I: NHS REC ethical approval



### **National Research Ethics Service** **NRES Committee London - Bentham**

Research Ethics Committee Offices  
Room 4W/12, 4th Floor  
Charing Cross Hospital  
Fulham Palace Road  
London  
W6 8RF

Telephone: 020 7794 0500 x34847

Miss Kate Goddard  
Centre for Health Informatics  
City University  
London  
EC1V 0HB

05 September 2011

Dear Miss Goddard

**Study title:** Automation bias in clinical decision support systems:  
over-reliance on automated advice - rates, effect  
mediators and mitigators  
**REC reference:** 11/LO/0701  
**Protocol number:** NA

Thank you for your letter of 08 August 2011, responding to the Committee's request for further information on the above research and submitting revised documentation.

The further information was considered in correspondence by a sub-committee of the REC 01<sup>st</sup> September 2011. A list of the sub-committee members is attached.

#### **Confirmation of ethical opinion**

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation as revised, subject to the conditions specified below.

#### **Ethical review of research sites**

##### **NHS sites**

The favourable opinion applies to all NHS sites taking part in the study, subject to management permission being obtained from the NHS/HSC R&D office prior to the start of the study (see "Conditions of the favourable opinion" below).

##### **Non-NHS sites**

#### **Conditions of the favourable opinion**

The favourable opinion is subject to the following conditions being met prior to the start of the study.

Management permission or approval must be obtained from each host organisation prior to the start of the study at the site concerned.

*Management permission ("R&D approval") should be sought from all NHS organisations involved in the study in accordance with NHS research governance arrangements.*

Guidance on applying for NHS permission for research is available in the Integrated Research Application System or at <http://www.rdforum.nhs.uk>.

*Where a NHS organisation's role in the study is limited to identifying and referring potential participants to research sites ("participant identification centre"), guidance should be sought from the R&D office on the information it requires to give permission for this activity.*

*For non-NHS sites, site management permission should be obtained in accordance with the procedures of the relevant host organisation.*

*Sponsors are not required to notify the Committee of approvals from host organisations*

**It is the responsibility of the sponsor to ensure that all the conditions are complied with before the start of the study or its initiation at a particular site (as applicable).**

#### **Approved documents**

The final list of documents reviewed and approved by the Committee is as follows:

Document	Version	Date
Advertisement		03 March 2011
Covering Letter		03 March 2011
Evidence of insurance or indemnity		03 March 2011
Investigator CV	Kate Goddard	
Investigator CV	Abdul V Roudsari	
Letter of invitation to participant		03 March 2011
Other: PhD Transfer peer review		03 March 2011
Other: City University Senate Research Ethics Committee letter of ethical approval		03 March 2011
Other: Three Questions		03 March 2011
Other: Letter of support from Dr Bainbridge		
Other: Validated primary care prescribing scenarios		
Other: E-mail to de-brief participants		
Other: E-mail to advertise the study		
Other: Content of debrief page at end of study		
Participant Consent Form	2	
Participant Information Sheet	2	
Protocol	1	03 March 2011
Questionnaire: Twenty Scenarios		
REC application		21 March 2011
Response to Request for Further Information		08 August 2011

### Statement of compliance

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees (July 2001) and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

### After ethical review

#### Reporting requirements

The attached document *"After ethical review – guidance for researchers"* gives detailed guidance on reporting requirements for studies with a favourable opinion, including:

- Notifying substantial amendments
- Adding new sites and investigators
- Notification of serious breaches of the protocol
- Progress and safety reports
- Notifying the end of the study

The NRES website also provides guidance on these topics, which is updated in the light of changes in reporting requirements or procedures.

#### Feedback

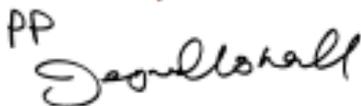
You are invited to give your view of the service that you have received from the National Research Ethics Service and the application procedure. If you wish to make your views known please use the feedback form available on the website.

Further information is available at National Research Ethics Service website > After Review

11/LO/0701	Please quote this number on all correspondence
------------	--

With the Committee's best wishes for the success of this project

Yours sincerely

PP  


**Professor David Katz**  
Chair

Email: georgina.marshall@nhs.net

*Enclosures: List of names and professions of members who were present at the meeting and those who submitted written comments [if final opinion was confirmed was given at a meeting]*

*"After ethical review – guidance for researchers" [SL-AR2]*

*Copy to: Dr Peter Weller  
Ms. Maggie Elliott, SWL Primary Care Community Services*

## Appendix J: Twenty final prescribing scenarios

The following patient comes to see you:

### 1. Depression

A 33-year-old man visits to report a generally low mood; during your consultation you notice that he is tearful, and when asked he described feelings of worthlessness. He lives with his wife and two young children in a nearby council house.

On enquiry, you elicit the following symptoms: his appetite is poorer than usual and he is irritable, he is moderately hypertensive, he has lost weight, he has a persistent mood of sadness and a tendency to wake up briefly in the middle of the night for the past 5 weeks. He does not have early morning waking or diurnal variation of mood but his concentration is disturbed and he is lethargic. He denies any suicidal thoughts and there is no past history of self-harm. He asks if he could be put on medication.

Preliminary diagnosis: Depression

- a) Would you give a rx?: Y/N - Y
- b) What would you prescribe?

#### CORRECT ANSWER

- Fluoxetine, 20mg, 1 per day, 14 capsules (NHS cost £0.70)
- Citalopram, 20mg, 1 per day, 14 capsules (NHS cost £0.64)
- Paroxetine, 20mg, 1 per day, 14 capsules (NHS cost £1.40)
- Sertraline, 50mg, 1 per day, 14 capsules (NHS cost £0.69)
- 

#### INCORRECT ANSWER [contraindicated for people with hypertension]

- Venlafaxine, 37.5mg, 1 tablet twice per day, 28 tablets (NHS cost £11.71)
- Duloxetine, 20mg, 1 tablet twice per day, 28 capsules (NHS cost £15.40)

## 2. Acute gastro-enteritis

A mother comes to you with her 6-month-old daughter. Her daughter has had diarrhoea for the last three days. It is a watery diarrhoea several times per day, without blood or mucus. The baby cries a lot, hardly drinks milk anymore and the mother thinks she has stomach cramps. She has lost 0.5 kg weight (from 8 to 7.5 kg). Her temperature is 38.4 °C.

On physical examination you find no signs of dehydration, and increased bowel sounds of the intestines. No abnormal findings are revealed by further history and physical examination.

**Preliminary diagnosis:** Acute gastro-enteritis

- a. Would you give a rx?: Y/N – Y or NO
- b. What would you prescribe?

### CORRECT ANSWER

- Rehydration - BNF for Children states antimotility drugs not suitable for children under 12 years. Use ORAL REHYDRATION SALTS e.g. Dioralyte. Amount: 1–1½ times usual feed volume

INCORRECT ANSWER [Adult antimotility drugs e.g. Loperamide, not recommended for children under 12 years]

- Loperamide, 2mg, 3 per day, 30 capsules (NHS cost £1.06)

### 3. Lyme disease

A 28-year-old woman comes in with a 1-day history of a number of oval, diffusely erythematous patches scattered across her chest, back, and extremities. She mentions she has recently been on a camping trip and having received numerous insect bites. The rash, which became more prominent after a hot shower or exposure to warmth, was also described as "burning". She reported a one week history of flu-like symptom; headache, neck pain, generalized body aches, fever and chills. She also complained that her cheeks felt unusually hot and were bright red. She denied having any associated respiratory or gastrointestinal symptoms.

A complete blood count and a comprehensive metabolic panel were performed. All results were normal except the liver function tests, which were mildly elevated with an alanine aminotransferase of 76 U/L, aspartate aminotransferase of 52 U/L, and alkaline phosphatase of 150 U/L. She states that she is 5 months pregnant.

**Preliminary diagnosis:** Lyme disease

- a. Would you give a rx?: Y/N – Y
- b. What would you prescribe?

**CORRECT ANSWER**

- Amoxicillin, 500mg, 3 times per day, 42 capsules (NHS cost £3.10)

**INCORRECT ANSWER [unsuitable for pregnant women]**

- Doxycycline, 100mg, 2 capsules first day then one per day for next 9 days, 11 capsules (NHS cost £0.77)

#### 4. Ankylosing spondylitis

A 65-year-old man presented with acute pain and swelling of one knee; the joint was tender and restricted in movement. X-ray of the knee showed periarticular osteoporosis. On investigation, he had a raised erythrocyte sedimentation rate of 100mm/h, a mild anaemia (Hb 104g/l) with a detectable serum rheumatoid factor. The knee effusion was aspirated; the fluid contained a polymorphonuclear leucocytosis. X-rays of his pelvis showed the classic changes of ankylosing spondylitis and tissue typing revealed that he was HLA-B27 positive. He has had intermittent backache over the last 5 years, although daily exercises have limited the stiffness. He has developed bony ankylosis between the lumbar vertebrae. He is asthmatic and has found that taking ibuprofen for the joint pain exacerbated his asthma.

**Preliminary diagnosis:** Ankylosing spondylitis

- a. Would you give a rx?: Y/N – Y
- b. What would you prescribe?

CORRECT ANSWER [Asthmatic – NSAID may not be appropriate therefore prescribe paracetamol or codeine]

- Paracetamol, 500mg, 2 tablets every 4-6 hours, 200 tablets (NHS cost £3.30)
- Codeine, 30mg, 1-2 tablets every 4-6 tablets, 84 tablets (NHS cost £3.57)

INCORRECT ANSWER [NSAIDs worsen asthma]

- Diclofenac sodium, 25mg, 3 times a day, 84 tablets (NHS cost £1.27)
- Diclofenac sodium, 50mg, 3 times a day, 84 tablets (NHS cost £1.43)
- Naproxen, 250mg, 1 tablet twice a day, 56 tablets (NHS cost £2.84)
- Naproxen, 500mg, 1 tablet twice a day, 56 tablets (NHS cost £3.80)

## 5. Acute otitis media

A 4 month old boy is brought in by his mother (he weighs 6kg). She mentions that for the past 2 days he awakes during the night and appears fussy and in discomfort. He has had increased nasal discharge, diarrhoea and has been vomiting. On inspection of both his ears the tympanic membrane shows a diminished light reflex and the bony landmarks are obscured.

**Preliminary diagnosis:** Bilateral acute otitis media

- a. Would you give a rx?: Y/N – Y
- b. What would you prescribe?

### CORRECT ANSWER

- Amoxicillin, 125mg suspension, 5ml 3 times a day, 5 days, 100ml (NHS cost £2.46)
- Ibuprofen, 100mg suspension, 2.5ml three times a day, 50ml (NHS cost £0.82)
- Paracetamol, 60-120mg suspension, 2.5ml to 5ml every 4 to 6 hours, upto 4 times a day, 150ml (NHS cost £0.84)

### INCORRECT ANSWER [overdose, below is for 5-12 year olds]

- Amoxicillin, 500mg suspension, 5ml 3 times a day, 5 days, 200ml (NHS cost £2.96)

## 6. Septic Olecranon bursitis

A 23 year old man present with pain at olecranon and down posterior arm which started 6 weeks ago after a fall in which the patient banged their elbow. He feels mild intermittent and increasing pain. The patient has begun to experience a mild fever with chilling and some sweating. On inspecting the skin around the olecranon, there is redness and swelling, the patient reports tenderness. Aspiration of the bursa reveals a leukocyte count above 100,000/mL.

**Preliminary diagnosis:** Olecranon bursitis

- a. Would you give a rx?: Y/N – Y
- b. What would you prescribe?

**CORRECT ANSWER**

- Flucloxacillin, 500mg, 1 capsule 4 times a day, supply 28 capsules (NHS cost £3.21)
- Plus appropriate analgesic

**INCORRECT ANSWER** [there is sepsis, need antibiotic as well]

- Ibuprofen, 400mg, 3-4 times a day, supply 84 tablets (NHS cost £1.72)
- Ibuprofen, 600mg, 3 times a day, supply 84 tablets (NHS cost £4.06)
- Ibuprofen, 2x400mg, 3 times a day, supply 168 tablets (NHS cost £3.74)
- Diclofenac sodium, 25mg, 3 times a day, supply 84 tablets (NHS cost £1.14)
- Diclofenac sodium, 50mg, 3 times a day, supply 84 tablets (NHS cost £1.31)
- Naproxen, 250mg, 2 times a day, supply 56 tablets (NHS cost £2.70)
- Naproxen, 500mg, 2 times a day, supply 56 tablets (NHS cost £3.44)

## 7. Renal Colic

A 53 year old man presents complaining of severe pain originating in the loin area which seems to spread into his groin. Pain is severe and episodic with the patient unable to find a comfortable position. Vomiting has occurred as a result of the pain. He has had renal colic twice in the past ten years. Urine dipstick testing tests positive for haematuria. He is currently on a course of naproxen for arthritis. He asks for something to reduce the pain in the first instance.

**Preliminary diagnosis:** Renal colic

- a. Would you give a rx?: Y/N – Y
- b. What would you prescribe?

**CORRECT ANSWER**

- Diamorphine (intramuscular), 5mg, 1 ampoule (NHS cost £2.69)

**INCORRECT ANSWER [contraindicated with naproxen]**

- Diclofenac (intramuscular), 75mg/3ml, 1 ampoule (NHS cost £0.83)

## 8. Acne - moderate

A 17 year old male comes in complaining of extensive and frequent papules and pustules on the face and trunk. He has tried several over the counter products that have not helped.

**Preliminary diagnosis:** Moderate acne

- a. Would you give a rx?: Y/N – Y
- b. What would you prescribe?

### CORRECT ANSWER

- Oxytetracycline, 250mg, 2 per day, 112 tablets (NHS cost £36.36).
- Tetracycline, 500mg, 2 per day, 112 tablets (NHS cost £36.36)
- Doxycycline, 50mg, 1 per day, 28 capsules (NHS cost £1.76)
- Lymecycline, 408mg, 1 per day, 28 capsules (NHS cost £7.77)
- Topical. Benzoyl peroxide

INCORRECT ANSWER [Oxy: Overdose for acne; Eryth: If tetracyclines are contraindicated]

- Oxytetracycline 1.5 g twice per day
- Erythromycin, 500mg, (2x250mg) 2 per day, 112 tablets (NHS cost £7.12)

## 9. Hypertension in pregnancy (Mild/moderate)

A 16 year old patient visits. She is 16 weeks pregnant and has attended the antenatal clinic three times. All findings were within normal limits until her last antenatal visit 1 week ago. At that visit, it was found that her blood pressure was 140/90 mm Hg. Her urine was negative for protein. The foetal heart sounds were normal, the foetus was active and uterine size was consistent with dates. A repeat visit has shown that her blood pressure remains at 140/90mmHg. She states she generally has a slightly high blood pressure.

She has no adverse symptoms (headache, visual disturbance, upper abdominal pain, convulsions or loss of consciousness).

The foetus is active and foetal heart sounds are normal. Uterine size is consistent with dates.

**Preliminary diagnosis:** Hypertension

- a. Would you give a rx?: Y/N – Y
- b. What would you prescribe?

**CORRECT ANSWER**

- Aspirin, 75mg, 1 per day, 28 tablets (NHS cost £0.83)

**INCORRECT ANSWER** [Nifedipine should only be used after 20 weeks' gestation if other treatment options are not indicated or have failed.]

- Nifedipine, 20mg, 1 per day, supply 28 capsules (NHS cost £5.06)

## 10. Meniere's disease

The patient is a 60-year old man suffering from hearing loss. He has noticed that his symptoms would increase when he was exposed to perfume smells used by women in his office.

He described symptoms of dizziness, tinnitus, and hearing loss for the past 6 years. The dizziness would occur several times per week. During these attacks, he feels lightheaded, nauseous and imbalanced. In addition, he complained of numbness in his arms and legs, severe pain inside his head, and intolerance of certain smells.

**Preliminary diagnosis:** Meniere's disease

- a. Would you give a rx?: Y/N – Y
- b. What would you prescribe?

CORRECT ANSWER [for recurrent Meniere's]

- Betahistine, 16mg, 3 per day, 84 tablets (NHS cost £3.97)
- Vestibular sedative (eg prochlorperazine or cinnarizine). This helps to control sickness and vertigo

INCORRECT ANSWER [wrong dose and strength]

- Betahistine, 24mg, 6 per day, 84 tablets (NHS cost £7.23)

## 11. Pericarditis (idiopathic)

A 56 year old patient visits, having already been diagnosed with pericarditis. He was previously prescribed ibuprofen to reduce fever, chest pain, and inflammation and has taken this alongside a proton pump inhibitor for the past 6 weeks. He complains that he still experiences recurrent chest pains.

**Preliminary diagnosis:** Idiopathic pericarditis

- a. Would you give a rx?: Y/N – Y
- b. What would you prescribe?

**CORRECT ANSWER** [For recurrent pericarditis, patients are reloaded with **colchicine** and continue colchicine therapy for at least 3 additional months with an NSAID]

- Colchicine, 500 micrograms, 2 per day, 12 tablets (£4.65)

**INCORRECT ANSWER**

- Azathioprine, 25mg, 1 per day, 28 tablets (NHS cost £7.49)

[In rare situations in which patients are refractory to corticosteroids, use of azathioprine may be considered, either in combination with colchicine and NSAIDs or as monotherapy.]

## 12. Pharyngitis

A mother comes in with her 2 year old child with a fever and a cough. Physical evaluation with a penlight reveals pharyngeal exudates and cervical adenopathy. Rapid antigen testing tests positive for Streptococcus. The child has demonstrated a penicillin allergy in the past.

**Preliminary diagnosis:** Pharyngitis

- a. Would you give a rx?: Y/N – Y
- b. What would you prescribe?

### CORRECT ANSWER

- Clarithromycin, 125mg/5ml oral suspension, one spoonful twice a day, 10 days (140ml) (NHS cost £13.64)
- Clarithromycin, 125mg/5ml oral suspension, 2.5ml twice a day, 10 days (70ml) (NHS cost £6.82)
- Erythromycin ethyl succinate, 250mg/5ml, one 5ml spoonful 4 per day, 10 days (100ml) (NHS cost £5.52)
- Ibuprofen, 100mg/5ml, one 5ml spoonful 3 per day, (100ml) (NHS cost £1.49)
- Paracetamol, 120mg/5ml, 1-2 spoonfuls every 4-6 hours, (300ml) (NHS cost £1.30)

### INCORRECT ANSWER [Too high a dose, this is for 8-11 year olds]

- Erythromycin ethyl succinate, 500mg/5ml, 1 spoonful 4 times a day, 10 days (200ml), (NHS cost £8.62)
- Aspirin, 300mg, 1 tablet every 4-6 hours, 64 tablets (NHS cost £0.62)

[Analgesics and local anaesthetics can be used for symptoms of sore throat, headache, and fever, although aspirin should be avoided in children because of its association with Reye's syndrome.]

### 13. Carpal tunnel syndrome

A 26-year-old male described a constant from pain and tingling from his right hand, up his arm, to his neck. His job involves working on a computer all day. When questioned about the timing of the onset of his symptoms, he recalled a bad fall skiing, which preceded the onset of his symptoms. Positive Phalen's test — flexing the wrist for 60 seconds causes pain or paraesthesia in the median nerve distribution. The symptoms are aggravating but mild and do not appear to be progressing further.

**Preliminary diagnosis:** Carpal tunnel syndrome

- a. Would you give a rx?: Y/N – N
- b. What would you prescribe?

#### CORRECT ANSWER

No drugs. Overnight splint. Any improvement should be apparent within 8 weeks of use.

[CKS recommendation]

#### INCORRECT ANSWER

Recommend the use of nonsteroidal anti-inflammatory drugs or diuretic medication.

- Aspirin, 75mg, 1 per day, 28 tablets (NHS cost £0.29)
- Naproxen, 250mg, 2 per day, 56 tablets, (NHS cost £2.84)

[The use of nonsteroidal anti-inflammatory drugs or diuretic medication is not recommended e.g. aspirin, ibuprofen, and naproxen. Diuretic: hydrochlorothiazide]

#### 14. Glandular fever/ infectious mononucleosis

A 25-year-old woman presented with a 10 day history of extreme malaise, loss of appetite, sore throat, and stiffness and tenderness of her neck. On examination, she had a mild fever with enlarged lymph nodes, palatal petechiae and pharyngeal inflammation without an exudate. Abdominal examination showed a mildly enlarged spleen.

Her white cell count was  $12 \times 10^9/l$  (NR  $4-10 \times 10^9/l$ ) with over 50% of the lymphocytes showing atypical morphology. Blood serum contained IgM antibodies to Epstein-Barr viral capsid antigen. Liver function tests were normal.

**Preliminary diagnosis:** Glandular fever

- a. Would you give a rx?: Y/N – Y
- b. What would you prescribe?

#### CORRECT ANSWER

- Paracetamol: adults: 500-1000 mg orally every 4-6 hours when required, maximum 4000 mg/day

*and/or*

- Ibuprofen: children 5-10 mg/kg orally every 6-8 hours when required, maximum 40 mg/kg/day; adults: 200-400 mg orally every 4-6 hours when required, maximum 1200 mg/day

[The goal of treatment is supportive care, including good hydration, anti-pyretics and analgesics, such as paracetamol and nonsteroidal anti-inflammatory drugs. Aspirin should not be given to children because of the possibility of Reye's syndrome.]

#### INCORRECT ANSWER

- Paracetamol: children: 10-15 mg/kg orally every 4-6 hours when required, maximum 90 mg/kg/day;
- Ibuprofen: children 5-10 mg/kg orally every 6-8 hours when required, maximum 40 mg/kg/day;

[these are children's doses]

## 15. Rubella

A 32 year old woman presents with a fever, enlarged lymph nodes and cold like symptoms. She also has a rash over her face and neck and appears to be spreading to her trunk and extremities. The rash is non-confluent and maculopapular. She has not had an MMR vaccination.

**Preliminary diagnosis:** Rubella

- a. Would you give a rx?: Y/N – Y
- b. What would you prescribe?

**CORRECT ANSWER**

- Ibuprofen, 200-400mg, 3-4 times per day, 56 tablets (NHS cost £1.19)
- Paracetamol, 500mg-1g, every 4-6 hours, 50 tablets (NHS cost £0.81)

**INCORRECT ANSWER** [inappropriate to prescribe antivirals]

- Zanamivir, 10mg inhaled, 2 per day, for 5 days, 100ml (NHS cost £24)
- Oseltamivir, 75mg, 2 per day, 10 capsules, 5 days (NHS cost £16.74)

## 16. Epistaxis

A 32 year old male complains of recurring nosebleeds over the past couple of months; they occur almost daily and can last for around 20 minutes. It is currently not bleeding. He mentions he is asthmatic and also has an allergy to peanuts.

**Preliminary diagnosis:** Epistaxis

- a. Would you give a rx?: Y/N – Y
- b. What would you prescribe?

### CORRECT ANSWER

[if peanut allergy]

- Silver nitrate cautery
- Mupirocin, 2% nasal ointment, small amount of cream to inside of nose, 2-3 times per day, for 5-7 days, supply 3g (NHS cost £5.80)

### INCORRECT ANSWER

[Not with people with peanut allergies]

- Naseptin, four times a day, 10 days, supply 15g (NHS cost £1.90)

## 17. Hirsutism

A 23 year old woman presents complaining of a gradual increase in body hair (over 6-8 months) over her face, chest, abdomen and legs. On inspection the hair is thick and coarse. Otherwise, her skin is clear and she is not overweight. The Ferriman-Gallaway score indicates she has moderate hirsutism. She confirms there is some personal and family history of deep vein thrombosis.

### **Preliminary diagnosis:** Hirsutism

- a. Would you give a rx?: Y/N – Y
- b. What would you prescribe?

**CORRECT ANSWER** [the oral contraceptives are contraindicated with DVT]

- Consider referring to secondary care, for systemic treatments such as spironactolone.
- Topical Eflornithine 11.5% cream: apply twice a day to face, supply 60 grams) (NHS cost £52.08)
- Advise cosmetic treatments such as laser treatment/ waxing/ electrolysis

**INCORRECT ANSWER**

[not for people with history of DVT]

- Cyproterone acetate, 2mg + ethinylestradiol 35mcg, 1 per day for 21 days, supply 63 tablets (NHS cost £6.51)  
Yasmin, drospirenone 3mg + ethinylestradiol 30mcg, 1 per day for 21 days, supply 63 tablets (NHS cost £14.70 [better than Dianette (history of thrombosis)])

## 18. Insomnia

A 34-year-old woman has difficulty falling and staying asleep. She needs two hours to fall asleep most nights, and often wakes very early in the morning and is unable to get back to sleep, leading her to feel drowsy during the day. The problem began after childbirth 5 years earlier, in association with mild postpartum depression. She has not had a recurrence of the depression and reports no significant psychosocial stressors. She always tries to get to bed at 11pm; going to bed later does not allow her to fall asleep more easily. She has no symptoms of sleep-disordered breathing, restless legs syndrome or sleepwalking and is otherwise well.

**Preliminary diagnosis:** Insomnia

- a. Would you give a rx?: Y/N – Y
- b. What would you prescribe?

**CORRECT ANSWER**

[NICE recommendations]

“Z” drugs:

[CKS advice]

- Zopiclone, 7.5mg, 1 per night, supply 7 tablets (NHS cost £0.37)
- Zolpidem, 10mg, 1 per night, supply 7 tablets (NHS cost £0.42)
- Zaleplon, 10mg, 1 per night, supply 7 tablets (NHS cost £1.43)

[BZs – may be prescribed with caveat of risk of addiction over longer term use]

- Temazepam, 10mg, 1 per night, supply 7 tablets (NHS cost £1.11)
- Loprazolam, 1mg, 1 per night, supply 7 tablets (NHS cost £4.50)
- Lormetazepam, 100micrograms, 1 per night, supply 7 tablets (NHS cost £14.75)]

-CBT is a possibility

**INCORRECT ANSWER**

[CKS recommends NOT to prescribe]

- Long-term hypnotics.
- Diazepam, 2mg, 3 per day, supply 21 tablets (NHS cost £1.13)
- Chloral hydrate, mixture 1g/10ml, 10ml per night, supply 200ml (NHS cost £10.67)
- Hydroxyzine, 25mg, 1 per night, supply 14 tablets, (NHS cost £0.55)
- Complementary and alternative therapies (such as acupuncture, acupressure).
- Valerian

## 19. Central Vertigo

A 60 year old woman comes in complaining of recurrent dizzy spells which have been occurring for 3 months. She has no other significant medical problems or on any current medications. The dizziness was vertiginous in nature (“rooms turning”). The frequency of the episodes has increased from once per month to 3 or 4 times per week. The vertigo lasts for 5 to 10 minutes each time. These episodes were associated with nausea and vomiting. She has suffered right ear tinnitus and loss of hearing for 3 months. There is no headache. Both eyes maintain usual visual acuity. The dizziness is not aggravated by change of position, head turning or motion. Gastrointestinal review is negative. No history of recent head trauma is reported. There is also no report of recent flu.

**Preliminary diagnosis:** Central vertigo

- a. Would you give a rx?: Y/N – Y
- b. What would you prescribe?

### CORRECT ANSWER

[CKS recommendation]

- Short course antihistamine
  - o Cyclizine, 50mg, upto 3 times a day, supply 21 tablets, (NHS cost £1.56)
  - o Promethazine teoclate, 25mg, upto 3 times a day, supply 21 tablets (NHS cost £2.35)
  - o Cinnarizine, 30mg, 3 times a day, supply 42 tablets (NHS cost £8.35)
- Short course prochlorperazine
  - o Prochlorperazine, 5mg, 1 upto 3 times today, supply 42 tablets, (NHS cost £1.97)
  - o Prochlorperazine, 3mg, 1-2 tablets upto 2 times per day, 30 buccal tablets (NHS cost £2.95)

### INCORRECT ANSWER

[too much for mild symptoms, unnecessary. This is for rapid relief of severe symptoms]

- Prochlorperazine, 12.5mg/1ml solution, supply 1x1ml ampoule (injection), (NHS cost £0.52)
- Prochlorperazine, 6mg stat, 2x3mg buccal tablets, supply 2 tablets, (NHS cost £0.20)

## 20. Paroxysmal Atrial Fibrillation

A 45 year old male with presents complaining of irregular pulse and heart palpitations. He has mild idiopathic, hypertrophic, subaortic stenosis. The episodes are recurrent and last 1-2 days appearing to self terminate. Each episode last no more than one day at a time. He has no history of heart failure, diabetes mellitus, stroke or transient ischemic attack.

**Preliminary diagnosis:** Paroxysmal atrial fibrillation

- a. Would you give a rx?: Y/N – Y
- b. What would you prescribe?

### CORRECT ANSWER

[low risk of stroke (looking at age/history), while waiting for specialist, beta blocker if there are frequent paroxysms without need for admission]

- Aspirin, 300mg, 1 per day, supply 28 tablets, (NHS cost £0.27)
- Aspirin, 75mg, 1 per day, supply 28 tablets, (NHS cost £0.82)
- Atenolol, 25mg, 1 per day, supply 28 tablets, (NHS cost £0.82)

### INCORRECT ANSWER

[CKS does not recommend the use of clopidogrel or a combination of aspirin and clopidogrel for AF in primary care]

- Warfarin, 1mg, take as directed in yellow anticoagulant booklet, supply 28 tablets, (NHS £1.10)
- Warfarin, 3mg, take as directed in yellow anticoagulant booklet, supply 28 tablets, (NHS £1.15)
- Warfarin, 5mg, take as directed in yellow anticoagulant booklet, supply 28 tablets, (NHS £1.21)
- Clopidogrel, 75mg, 1 per day, supply 30 tablets, (NHS cost £36.35)

## References

- <sup>1</sup> Protti D. The Application of Computer Technology in GP Surgeries is Beginning To Have Positive Effects on Chronic Disease Management. World View Reports. London: National Health Service; 2005 [cited Oct 2010] Available from: <http://www.connectingforhealth.nhs.uk/worldview/protti6>
- <sup>2</sup> Coiera E, Westbrook J, Wyatt JC. The safety and quality of decision support systems. Commentary in: Haux R, Kulikowski C (eds). Yearbook of Medical Informatics 2006. Stuttgart: Schattauer Verlag, 2006: p. 20-5
- <sup>3</sup> Dean B, Schachter M, Vincent C, Barber N. Causes of prescribing errors in hospital inpatients: a prospective study. *Lancet* 2002; 359(9315):1373–8
- <sup>4</sup> Bates DW, Gawande AA. Improving safety with information technology. *N Engl J Med*. 2003; 348(25):2526–2534
- <sup>5</sup> Bobb A, Gleason K, Hush M et al. The epidemiology of prescribing errors. *Arch Intern Med*. 2004; 164: 785-792.
- <sup>6</sup> Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, Sam J, Haynes RB. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *J Am Med Assoc*. 2005; 293(10):1223-38
- <sup>7</sup> Hunt D, Haynes B, Hanna S, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: A systematic review. *J Am Med Assoc*. 1998; 280(15):1339-46.
- <sup>8</sup> Field TS, Rochon P, Lee M, et al. Computerized clinical decision support during medication ordering for long-term care residents with renal insufficiency. *J Am Med Inform Assoc*. 2009; M2981.
- <sup>9</sup> Sullivan F, Wyatt JC. ABC of Health Informatics 4: How decision support tools help define clinical problems. *Brit Med J*. 2005; 331: 831-3
- <sup>10</sup> Skitka L. Does automation bias decision-making? *Int J Hum-Comput St*. 1999; 51: 991-1006
- <sup>11</sup> Weingart SN, Toth M, Sands DZ, Aronson MD, Davis RB, Phillips RS. Physicians' decisions to override computerized drug alerts in primary care. *Arch Intern Med*. 2003;163(21):2625-31
- <sup>12</sup> Scott GPT, Shah P, Wyatt JC, Makubate B, Cross FW. Making electronic prescribing alerts more effective: scenario-based experimental study in junior doctors. *Journal of the American Medical Informatics Association* [Internet]. 2011 [cited 2011 Sep 26]; Available from: <http://jamia.bmj.com/content/early/2011/08/11/amiainl-2011-000199.abstract>
- <sup>13</sup> Goddard K, Roudsari A, Wyatt J. Decision Support and Automation Bias: Methodology and Preliminary Results of a Systematic Review. *Stud Health Technol Inform*. 2011; 164
- <sup>14</sup> Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inf Assoc*, 2011. Available at: <http://jamia.bmj.com/content/early/2011/06/16/amiainl-2011-000089.abstract>
- <sup>15</sup> Department of Health. Building a safer NHS for patients 2001. [cited Sept 2009] Available from: [http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH\\_4006525](http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4006525)

- 
- <sup>16</sup> Thimbleby H. Errors + Bugs Needn't Mean Death. Public Service Review: UK Science & Technology. 2011; 2:18-19. [cited Jan 2012] Available from: <http://www.cs.swan.ac.uk/~csharold/cv/files/STUK2.pdf>
- <sup>17</sup> Williams S, Osborn S. National Patient Safety Agency: an introduction. Clin Gov Int J. 2004; 9(2):130-131
- <sup>18</sup> National Patient Safety Agency, Rapid Response Report NPSA/2010/RRR015: Prevention of over infusion of intravenous fluid and medicines in neonates, 2010 [cited Jan 2012] Available from: <http://www.nrls.npsa.nhs.uk/resources/?EntryId45=75519>
- <sup>19</sup> Holbrook A, Pullenayegum E, Thabane L, Troyan S, Foster G, Keshavjee K, Chan D, Dolovich L, Gerstein H, Demers C, Curnew G. Shared Electronic Vascular Risk Decision Support in Primary Care: Computerization of Medical Practices for the Enhancement of Therapeutic Effectiveness (COMPETE III) Randomized Trial. Arch Intern Med. 2011;171(19):1736-1744
- <sup>20</sup> Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. Brit Med J. 2005; 330(7494): 765
- <sup>21</sup> Parasuraman R, Molloy R, Singh IL. Performance consequences of automation-induced complacency. Int J Aviat Psychol; 1993, 3, 1-23
- <sup>22</sup> Parasuraman R, Riley V. Humans and Automation: Use, Misuse, Disuse, Abuse Hum Factors. 1997; 39(2): 230-253
- <sup>23</sup> Friedman CP, Elstein AS, Wolf FM, Murphy GC, Franz TM, Heckerling PS, Fine PL, Miller TM, Abraham V. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems. J Am Med Assoc. 1999; 282(19):1851-6
- <sup>24</sup> Skitka LJ, Mosier K, Burdick. Does Automation bias decision making? Int J Hum-Comput St.1999, 51(5): 991–1005
- <sup>25</sup> Skitka LJ, Mosier KL, Burdick M, Rosenblatt B. Automation bias and errors: are crews better than individuals? Int J Aviat Psychol. 2000;10(1):85-97
- <sup>26</sup> Reason JT. Human Error, Cambridge University Press, New York, 1990
- <sup>27</sup> Lee JD, See KA. Trust in automation: Designing for appropriate reliance. Hum Factors. 2004; 46(1): 50-80
- <sup>28</sup> Singh IL, Molloy R, Parasuraman R. Automation induced “complacency”: development of the complacency-potential rating scale. Int J Aviation Psychol. 1993; 3, 111-122
- <sup>29</sup> Cummings ML. Automation bias in intelligent time critical decision support systems. In: AIAA 1<sup>st</sup> Intelligent Systems Technical Conference, AIAA 2004
- <sup>30</sup> Over-reliance on computers ‘caused Flash Crash’. [cited June 2011] Available at: <http://www.information-age.com/channels/information-management/news/1275048/overreliance-on-computers-caused-flash-crash.shtml>
- <sup>31</sup> Turkish Airlines Flight 1951 [cited June 2010] Available at: [http://en.wikipedia.org/wiki/Turkish\\_Airlines\\_Flight\\_1951](http://en.wikipedia.org/wiki/Turkish_Airlines_Flight_1951)
- <sup>32</sup> Learmount, D. Flight Safety Foundation's European Aviation Safety Seminar. Flight

- <sup>33</sup> Vicente KJ. Less is (sometimes) more in cognitive engineering: the role of automation technology in improving patient safety. *Qual Saf Health Care* 2003;12:291-294
- <sup>34</sup> Azen R and Budescu, DV. The dominance analysis approach for comparing predictors in multiple regression. *Psychol Methods*. 2003; 8(2): 129–148
- <sup>35</sup> Skitka L, Mosier, K, Burdick K. Accountability and automation bias. *Int. J. Hum-Comput St.* 2000a; 52: 701-717
- <sup>36</sup> Humphrey, SE, Hollenbeck, JR, Meyer CJ, Ilgen, DR. Hierarchical team decision making. In: Ferris GR and Martocchio JJ, (Eds), *Research in personnel and human resources management*, JAI Press, Stamford, CT; 2002, pp. 175–213
- <sup>37</sup> Friedman CP, Wyatt JC. *Evaluation methods in biomedical informatics*. 2nd ed. New York: Springer-Publishing, October; 2005
- <sup>38</sup> Des Roches CM, Campbell EG, Rao SR,, et al. Electronic health records in ambulatory care—a national survey of physicians. *N Engl J Med* 2008 Jul;359(1):50-60
- <sup>39</sup> Hsaio C, Burt C, Rechtsteiner E, et al. Preliminary estimates of electronic medical records use by office-based physicians, 2008. *Health E-Stat National Center for Health Statistics*; 2008. [cited March, 2010] Available at: <http://www.cdc.gov/nchs/products/pubs/pubd/hestats/physicians08/physicians08.htm>
- <sup>40</sup> Electronic Prescription Service [cited June 2011] Available from: <http://www.connectingforhealth.nhs.uk/systemsandservices/eps>
- <sup>41</sup> Coiera E. *The Guide to Health Informatics*. 2nd Edition. Arnold, London, October; 2003
- <sup>42</sup> Campbell NC, Murray E, Darbyshire J, Emery J, Farmer A, Griffiths F, Guthrie B, Lester H, Wilson P, Kinmonth AL. Designing and evaluating complex interventions to improve health care. *Brit Med J*. 2007; 334:455-459
- <sup>43</sup> Wyatt J, Spiegelhalter D. Field trials of medical decision-aids: potential problems and solutions. In Clayton P (Ed). *Proc. 15th Symposium on Computer Applications in Medical Care*, Washington. New York: McGraw Hill Inc. 1991; 3-7
- <sup>44</sup> O’Cathain A, Webber E, Nicholl J, Munro J, Knowles E. NHS Direct: consistency of triage outcomes. *Emerg Med J* 2003;20:289-292
- <sup>45</sup> Blackmore CC, Mecklenburg RS, Kaplan GS. Effectiveness of Clinical Decision Support in Controlling Inappropriate Imaging. *J American Coll Radiol*, 2011; 8(1): 19-25
- <sup>46</sup> Kong G, Xu DL, Yang JB. Clinical decision support systems: A review on knowledge presentation and inference under uncertainties. *Int J Comput Int Sys*. 2008; 1(2); 159-167
- <sup>47</sup> Harvey K, Dartnell J, Hemming M. Improving antibiotic use: 25 years of antibiotic guidelines and related initiatives. *Comm Dis Intell*. 2003; 27: S9-11
- <sup>48</sup> Hecker MT, Aron DC, Patel NP, Lehmann MK, Donskey CJ. Unnecessary use of antimicrobials in hospitalized patients: Current patterns of misuse with an emphasis on the anaerobic spectrum of activity. *Arch Int Med*. 2003; 163: 972-978

- 
- <sup>49</sup> National Patient Safety Agency. Patient Safety Incident Reports in the NHS National Reporting and Learning System Quarterly Data Summary England. Issue 14. 2009. [cited Sept 2011]. Available at: <http://www.nrls.npsa.nhs.uk/resources/collections/quarterly-data-summaries/>
- <sup>50</sup> Barber N, Rawlins M, Franklin BD. Reducing prescribing error: competence, control, and culture. *Qual Saf Health Care*. 2003;12:29-32
- <sup>51</sup> Williams ML, Dennis AR, Stam A, Aronson JE. The Impact of DSS Use and Information Load on Errors and Decision Quality. *Eur J Operational Res*. 2007; 176(1): 468–481
- <sup>52</sup> Sayers YM, Armstrong P, Hanley K. Prescribing errors in general practice: A prospective study. *Eur J Gen Pract*. 2009; 15(2): 81 - 83
- <sup>53</sup> Gandhi TK, Weingart SN, Seger AC, Borus J, Burdick, E, Poon EG, Leape LL, Bates DW. Outpatient Prescribing Errors and the Impact of Computerized Prescribing. *J Gen Intern Med*. 2005 September; 20(9): 837–841
- <sup>54</sup> An in depth investigation into causes of prescribing errors by foundation trainees in relation to their medical education - EQUIP study, published Dec 2009 [cited Oct 2011] Available from [http://www.gmc-uk.org/about/research/research\\_commissioned\\_4.asp](http://www.gmc-uk.org/about/research/research_commissioned_4.asp).
- <sup>55</sup> Avery AJ, Barber N, Maisoon G et al. Investigating the prevalence and causes of prescribing errors in general practice: The PRACtICe Study (PREvalence And Causes of prescribing errors in general practiCe) A report for the GMC. Available at: <http://www.gmc-uk.org/about/research/12996.asp>
- <sup>56</sup> Avery AJ, Rodgers S, Cantrill JA at al. A pharmacist-led information technology intervention for medication errors (PINCER): a multicentre, cluster randomised, controlled trial and cost-effectiveness analysis. *Lancet*. 2012; 379; 1320-1319
- <sup>57</sup> Pearson SA, Moxey A, Robertson J, Hains I, Williamson N, Reeve J, Newby D. Do computerised clinical decision support systems for prescribing change practice? A systematic review of the literature (1990-2007). *BMC Health Serv Res*. 2009; 28;9:154
- <sup>58</sup> Durieux P, Trinquart L, Colombet I, , Niès J, Walton R, Rajeswaran A, Rège Walther M, Harvey E, Burnand B. Computerized advice on drug dosage to improve prescribing practice. *Cochrane Database of Systematic Rev* 2008; (3): CD002894.
- <sup>59</sup> Judge J, Field TS, DeFlorio M et al. Prescribers' response to alerts during medication ordering in the long term care setting. *J Am Med Inform Assoc*. 2006; 13(4): 385-390
- <sup>60</sup> van Doormal JE, van den Bernt PMLA, Zaal RJ. Et al. The influence that electronic prescribing has on medication errors and preventable adverse drug events. *J Am Med Assoc*. 2009; 16; 816-825
- <sup>61</sup> Miller RA, Gardner RM, Johnson KB, Hripcsak G. Clinical decision support and electronic prescribing systems: a time for responsible thought and action. *J Am Med Inform Assoc*. 2005 Jul-Aug;12(4):403-9.
- <sup>62</sup> McTaggart S. Decision Support Systems. In: Bond C (Ed) *Using Medicines Information*. Oxford: Radcliffe Press; 2008
- <sup>63</sup> Avery AJ. Why ePrescribing should be supporting patient care. Online presentation. Available at: <http://www.connectingforhealth.nhs.uk/systemsandservices/eprescribing/systemimp/tonyavery.pdf> [cited June 2012]
- <sup>64</sup> Sintchenko V, Westbrook J, Tipper S, et al. Electronic Decision Support Activities in Different Healthcare Settings in Australia. Appendix A in: National Electronic Decision Support Taskforce.

---

Electronic Decision Support for Australia's Health Sector. Canberra, Department of Health and Aging, November 2002.

<sup>65</sup> Electronic Prescribing in Hospitals: Challenges and Lessons Learned. Report commissioned by NHS Connecting for Health (NHS CFH) [cited June 2011] Available from: [http://www2.lse.ac.uk/LSEHealthAndSocialCare/pdf/information%20systems/eprescribing\\_report.pdf](http://www2.lse.ac.uk/LSEHealthAndSocialCare/pdf/information%20systems/eprescribing_report.pdf)

<sup>66</sup> ScriptSwitch [cited June 2010] Available at: <http://www.scriptswitch.com/>

<sup>67</sup> British Medical Association. Briefing Paper – General Practitioners [cited June 2012] Available at: [http://www.bma.org.uk/press\\_centre/pressgps.jsp#.T8izDNW0yuk](http://www.bma.org.uk/press_centre/pressgps.jsp#.T8izDNW0yuk)

<sup>68</sup> Moxey A, Robsertson J, Newby D et al. . Computerized clinical decision support for prescribing: provision does not guarantee uptake. *J Am Med Inform Assoc* 2010;17:25-33

<sup>69</sup> European Centre for Disease Control and Prevention/European Medicines Agency Joint Working Group. The Bacterial Challenge: Time to React. Stockholm, 2009. [cited Nov 2011] Available from: [http://www.ema.europa.eu/docs/en\\_GB/document\\_library/Report/2009/11/WC500008770.pdf](http://www.ema.europa.eu/docs/en_GB/document_library/Report/2009/11/WC500008770.pdf)

<sup>70</sup> Jolleys JV, Oleson F. A comparative study of prescribing of hormone replacement therapy in USA and Europe. *Maturitas*. 1996; 23(1), 47-53

<sup>71</sup> Butler CC, Hood K, Verheij T, Little P, Melbye H, Nuttall J, Kelly MJ, Mölsted S, Godycki-Cwirko M, Almirall J, Torres A, Gillespie D, Rautakorpi U, Coenen S, Goossens H. Variation in antibiotic prescribing and its impact on recovery in patients with acute cough in primary care: prospective study in 13 countries. *Br Med J*. 2009; 23;338:b2242

<sup>72</sup> Fretheim A, Oxman AD. International variation in prescribing antihypertensive drugs: Its extent and possible explanations. *BMC Health Serv Res*. 2005; 5:21

<sup>73</sup> Stolk P, Van Wijk BL, Leufkens HG, Heerdink ER., 2006. Between-country variation in the utilization of antihypertensive agents: guidelines and clinical practice. *J Hum Hypertens*. 2006; 20: 917–922

<sup>74</sup> Lawson DH, Jick D. Drug prescribing in hospitals: an international comparison. *Am J Pub Health*. 1976; 66(7): 644-648

<sup>75</sup> Handelsman D. Trends and regional differences in testosterone prescribing in Australia, 1991–2001. *Med J Aust*. 2004; 181(8):419-22

<sup>76</sup> McClean P, Hughes C, Tunney M, Goossens H, Jans B. Antimicrobial prescribing in European nursing homes. *J Antimicrob Chemother*. 2011;66(7):1609-1616

<sup>77</sup> Kotseva K, Wood D, De Backer G, De Bacquer D, Pyorala K, Keil U et al. Clinical reality of coronary prevention guidelines: a comparison of EUROASPIRE I and II in nine countries. *Lancet*. 2001;357:995-1001

<sup>78</sup> Sturm HB, van Gilst WH, Veeger N, Haaijer-Ruskamp FM. Prescribing for chronic heart failure in Europe: does the country make the difference? A European survey. *Pharmacoepidemiol Drug Saf*. 2007; 16: 96-103

<sup>79</sup> Cleland JG, Cohen-Solal A, Aguilar JC, Dietz R, Eastaugh J, Follath F et al. Management of heart failure in primary care (the IMPROVEMENT of Heart Failure Programme): an international survey. *Lancet* 2002;360:1631-9.

- 
- <sup>80</sup> Clavenna A, Bonati M. Differences in antibiotic prescribing in paediatric outpatients. *Arch Dis Child* 2011;96:590-595
- <sup>81</sup> Anderson S. Researching and writing the history of pharmacy. In: Anderson S, ed. Making medicines: a brief history of pharmacy and pharmaceuticals. London : Pharmaceutical Press, 2005: 10-18
- <sup>82</sup> Schoen C, Osborn R, Doty MM, Bishop M, Peugh J, Murukutla N. Toward higher-performance health systems : adults' health care experiences in seven countries. *Health Affairs*. 2007; 6: w717-w734
- <sup>83</sup> Berbatis C. Community pharmacy practice: 3. international reports show opportunities for Australia. *i2P Issue 63*. 2007 [cited June 12, 2011] Available from: <http://i2P.com.au/?page=site/article&id=745>
- <sup>84</sup> Sturm HB, van Gilst WH, Swedberg K, Hobbs R, Haaijer-Ruskamp F. Heart failure guidelines and prescribing in primary care across Europe. *BMC Health Serv Res*. 2005; 5: 57
- <sup>85</sup> Reggi V, Balocco-Mattavelli R, Bonati M, Breton I, Figueras E, Jambert C. Prescribing information in 26 countries: a comparative study. *Eur J Clin Pharmacol*. 2003; 59(4): 263-270
- <sup>86</sup> Helfand M, Mahon SM, Eden KB, Frame PS, Orleans CT. Screening for skin cancer. *Am J Prev Med*. 2001; 20(3 Suppl):47–58.
- <sup>87</sup> Van Duijn, Kuyvenhoven MM, Butler CC, Coenen S, Verheij TJM. Variation in outpatient antibiotic use in three European countries: exploration of possible – determinants. *Eur J Gen Pract*. 2005;11(3-4):139-40.
- <sup>88</sup> Freidson E. *Doctoring together*. New York/Oxford/Amsterdam: Elsevier, 1975
- <sup>89</sup> Steinke DT, Bain DJ, MacDonald Tm, Davey PG. Practice factors that influence antibiotic prescribing in general practice in Tayside. *J Antimicrob Chemother*. 2000; 46(3): 509-512.
- <sup>90</sup> Patten S, Cipriani A, Brambilla P, Nose M, Barbui M. International Dosage Differences in Fluoxetine Clinical Trials. *Can J Psychiatry*. 2005; 50(1): 31-38
- <sup>91</sup> Akici A, Kalaça S, Ugurlu MU, *et al*. Prescribing habits of general practitioners in the treatment of childhood respiratory-tract infections. *Eur J Clin Pharmacol*. 2004; 60: 211 – 16
- <sup>92</sup> Bocquet A, Chalumeau M, Bollotte D, *et al*. Comparison of prescriptions by pediatricians and general practitioners: a population-based study in Franche- Comté from the database of Regional Health Insurance Fund. *Arch Pediatr*. 2005 ; 12: 1688 – 96
- <sup>93</sup> Quach C, Collet JP, LeLorier J. Acute otitis media in children: a retrospective analysis of physician prescribing patterns. *Br J Clin Pharmacol*. 2004; 57: 500 – 5
- <sup>94</sup> De Las Cuevas C, Sanz EJ, De La Fuente JA. Variations in antidepressant prescribing practice: clinical need or market influences. *Pharmacoepidemiol Drug Saf*. 2002; 11(6): 515–522
- <sup>95</sup> Peyrot M, Rubin PR, Lauritzen T, Skovlund SE, Snoek FJ, Matthews DR, Landgraf R, Kleinebreil L. Resistance to insulin therapy among patients and providers: results of the cross-national Diabetes Attitudes, Wishes, and Needs (DAWN) study. *Diabetes Care*. 2005; 28(11): 2673-9
- <sup>96</sup> Castelo-Branco C, Ferrer J, Palacios S, Cornago S. The prescription of hormone replacement therapy in Spain: Differences between general practitioners and gynaecologists. *Maturitas*. 2006; 55(4):308-16

---

<sup>97</sup> Barclay S, Todd C, Finlay I, Grande G, Wyatt P. Not another questionnaire! Maximizing the response rate, predicting non-response and assessing non-response bias in postal questionnaire studies of GPs. *Fam Pract.* 2002; 19(1): 105-111

<sup>98</sup> McAvoy BR, Kaner EFS. General Practice postal surveys: a questionnaire too far? *Br Med J.* 1996; 313: 732 - 733

<sup>99</sup> Bonevski B, Horton G, Foster M, Girgis A. Response rates in GP surveys: Trialling two recruitment strategies. *Aust Fam Physician.* 2011; 40(6): 427-430

<sup>100</sup> Britt H, Miller GC, Charles J, et al. General practice activity in Australia 2008–09. BEACH: Bettering the Evaluation and Care of Health. General Practice Series No. 25. Cat No. GEP 25. Canberra: AIHW; 2009

<sup>101</sup> Mealing NM, Banks E, Jorm LR, et al. Investigation of relative risk estimates from studies of the same population with contrasting response rates and designs. *BMC Med Res Methodol* 2010; 10:26

<sup>102</sup> Morris CJ, Cantrill JA, Weiss MC. GP survey response rate: a miscellany of influencing factors. *Fam Pract.* 2001; 18(4): 454-456

<sup>103</sup> Krueger JL. Return of the ego—self-referent information as a filter for social prediction: comment on Karniol. *Psychological Review.* 2003; 110: 585–590

<sup>104</sup> Tversky A, Kahneman D. Judgment under uncertainty: Heuristics and biases. *Science.* 1974; 185: 1124-1130

<sup>105</sup> Cain DM, Loewenstein G, Moore DA. The Dirt on Coming Clean: Perverse Effects of Disclosing Conflicts of Interest. *J Legal St.* 2005; 34(1): 1-25

<sup>106</sup> Malmandier U, Shanthikumar D. Are small investors naïve about incentives? *J Financ Econ.* 2007; 85(2): 457-489

<sup>107</sup> Dzindolet MT, Pierce LG, Beck HP, Dawe LA, Anderson BW. Predicting misuse and disuse of combat identification systems. *Mil Psychol.* 2001; 13(3): 147-164

<sup>108</sup> Dzindolet MT, Pierce LG, Beck HP, Dawe LA. The perceived utility of human and automated aids in a visual detection task. *Hum Factors.* 2002; 44: 79-94

<sup>109</sup> Ostermann J, Weissenfeld A. Face Animation for Human Computer Interfaces. *Proc. Workshop on Image Analysis For Multimedia Interactive Services (WIAMIS) Lisboa, Portugal.* 2004

<sup>110</sup> Hedlund, J, Ilgen, DR. Hollenbeck, JR. Decision accuracy in computer-mediated versus face-to-face decision-making teams. *Org Behav Hum Dec.* 1998; 76: 30–47.

<sup>111</sup> Madhavan P, Wiegmann DA. Effects of information source, pedigree and reliability on operators' utilization of diagnostic advice; *Proceedings of the 49h Annual Meeting of the Human Factors and Ergonomics Society*; pp. 487- 491. Santa Monica: CA; 2005

<sup>112</sup> Parkes A. Persuasive decision support: Improving reliance on decision support systems. *Pacific Asia Conference on Information Systems.* Hyderabad, India: Association of Information Systems; 2009

<sup>113</sup> Yaniv I. Receiving other people's advice: influence and benefit. *Org Behav Hum Dec.* 2004b; 93: 1–13

- 
- <sup>114</sup> Harries C, Yaniv I, Harvey N. Combining advice: the weight of a dissenting opinion in the consensus. *J Behav Dec Making*. 2004; 17: 333–348
- <sup>115</sup> Hogarth RM, Einhorn HJ. Order effects in belief updating: the belief-adjustment model. *Cog Psychol*. 1992; 24: 1–55
- <sup>116</sup> Price PC, Stone ER. Intuitive evaluation of likelihood judgment producers: evidence for a confidence heuristic. *J Behav Dec Making*. 2004; 17: 39–57
- <sup>117</sup> Mosier KL, Skitka LJ. Human decision makers and automated decision aids: made for each other? In: Parasuraman, R., Mouloua, M. (Eds.), *Automation and Human Performance: Theory and Application*. Erlbaum, Mahwah, NJ, 1996; p. 201–220
- <sup>118</sup> Simon HA. Invariants of human behaviour. *Annu Rev Psychol*. 1990; 41: 1-19
- <sup>119</sup> Gigerenzer G, Goldstein DG. Reasoning the fast and frugal way: Models of bounded rationality. *Psychol Rev*. 1996; 103(4): 650-669
- <sup>120</sup> Shafir E, LeBoeuf RA. Rationality. *Annu Rev Psychol*. 2002; 53: 491-517
- <sup>121</sup> Hammond KR. Functionalism and illusion: Can integration be usefully achieved? In RM Hogarth (Ed), *Insights in decision-making*; Chicago: University of Chicago Press; 1990; p 227-261
- <sup>122</sup> Anderson JR. *The adaptive character of thought*. Hillsdale, NJ: Erlbaum; 1990
- <sup>123</sup> Zajonc, RB. Attitudinal Effects of Mere Exposure. *J Pers Soc Psychol*. 1968; 9(2): 1-27
- <sup>124</sup> McCloy R, Beaman CP, Goddard K. Rich and famous: Recognition-based judgment in the Sunday Times rich list. In: R. Sun & N. Miyake (Eds.) *Proceedings of the 28th Annual Conference of the Cognitive Science Society*; pp. 1801-1805; Austin, TX: Cognitive Science Society; 2006
- <sup>125</sup> Bown NJ. The relevance of judgment and decision making research for marketing: Introduction to the special issue. *Marketing Theory*. 2007; 7(1): 5-11
- <sup>126</sup> Agosto DE. Bounded rationality and satisficing in youngpeople' web-based decision making. *J Am Soc Inform Sci*. 2002; 53(1): 16-27
- <sup>127</sup> Spring B. Health decision making: lynchpin of evidence-based practice. *Med Decis Mak*. 2008; 28(6): 866-874
- <sup>128</sup> Zachary WW, Ryder JM. Decision support systems: integrating decision aiding and decision training. In M Helander, TK Landauer, P Prabhu (Eds). *Handbook of human computer interaction*. 2<sup>nd</sup> ed. Elsevier Science, 1997: 1235-1258
- <sup>129</sup> Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. *Arch Intern Med*. 2005; 165(13):1493-9.
- <sup>130</sup> Beck HP, Dzindolet MT, Pierce LG. Applying a decision-making model to understand misuse, disuse, and appropriate automation use. In E. Salas, C. A. Bower, N. Cooke, J. Driskell, & D. Stone (Eds.) *Advances in Human Factors and Cognitive Engineering, Vol 2*, Boston: JAI Press; 2002.
- <sup>131</sup> Wang L, Jamieson GA, Hollands JG. Selecting methods for the analysis of reliance for automation. *Hum Factors*. 2008; 52(4): 287-291
- <sup>132</sup> Baba, ML, Falkenburg, DR, Hill, DH. Technology management and American culture: Implications for business process redesign. *Res Technol Manage*. 1996; 39(6): 44–54

- 
- <sup>133</sup> Doney, PM, Cannon, JP, Mullen, MR. Understanding the influence of national culture on the development of trust. *Acad Manage Rev.* 1998; 23: 601–620
- <sup>134</sup> Lapointe L, Rivard S. Getting physicians to accept new information technology: insights from case studies. *Can Med Assoc J.* 2006; 174(11): 1573-78
- <sup>135</sup> Masalonis AJ. Effects of situation-specific reliability on trust and usage of automated decision aids. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting.* 2003; 47: 3533-537
- <sup>136</sup> Bahner JE, Huper AD, Manzey D. Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *Int. J. Human-Computer Studies.* 2008; 66: 688-699
- <sup>137</sup> Mosier KL, Sethi N, McCauley S, Khoo L, Orasanu JM. What you don't know can hurt you: factors impacting diagnosis in the automated cockpit. *Hum Factors.* 2007; 49(2):300-10.
- <sup>138</sup> Gomaa MI. The Effect of Time Pressure, Task Complexity and Litigation Risk on Auditors'. *Reliance on Decision Aids*, Published PhD Dissertation. 2005 [cited Jan 2010] Available from: <http://arno.unimaas.nl/show.cgi?fid=6461>
- <sup>139</sup> Payne JW. Task complexity and contingent processing in decision making: an information search and protocol analysis. *Org Behav Hum Perform.* 1976; 16: 366-387
- <sup>140</sup> Wright P. The harassed decision maker: time pressures, distractions, and the use of evidence. *J App Psychol.* 1974; 59(5): 555-561
- <sup>141</sup> Gomaa MI, Hunton JE, Rose JM. The effects of control risk and litigation risk on decision aid reliance. 2008. *Int J App Dec Sci.* 2008; 1(1): 80-106
- <sup>142</sup> Graham GN. Perceived versus actual risk for hypertension and diabetes in the African American community. *Health Prom Prac.* 2006; 7(1): 34-46
- <sup>143</sup> Powers BJ. Perceived and actual stroke risk among men with hypertension. *J Clin Hypertens.* 2008; 10(4): 287-294
- <sup>144</sup> Kontos AP. Perceived Risk, Risk Taking, Estimation of Ability and Injury Among Adolescent Sport Participants. *J Ped Psychol.* 2004; 29(6): 447-455
- <sup>145</sup> Blais AR, Weber EU. A Domain-Specific Risk-Taking (DOSPERT) scale for adult populations. *Judg Decis Mak.* 2006; 1(1): 33-47
- <sup>146</sup> Bin G. Moderating effects of task characteristics on information source use: An individual level analysis of R&D professionals in new product development. *J Inf Sci.* 2009; 35(5): 527-547
- <sup>147</sup> Newell A, Simon HA. *Human Problem Solving*, Prentice Hall. Englewood Cliffs, NJ; 1972.
- <sup>148</sup> Jacoby J, Speller DE, Kohn CA. Brand choice behaviour as a function of information load: replication and extension. *J Consum Res.* 1974; 3(4): 209-216
- <sup>149</sup> Chinburapa V, Larson LN, Brucks M, Draugalis J, Bootman JL, Puto CP. Physician prescribing decisions: the effects of situational involvement and task complexity on information acquisition and decision making. *Soc Sci Med.* 1993; 36(11):1473-82
- <sup>150</sup> Daly MA. Task Load and Automation Use in an Uncertain Environment. *Masters Thesis.* 2002. [cited Jun 2011] Available from: <https://research.maxwell.af.mil/papers/ay2002/afit/afit-gaq-env-02m-05.pdf> .

- 
- <sup>151</sup> Biros DP, Daly M, Gunsch G. The Influence of Task Load and Automation Trust on Deception Detection. *Group Decis Negot.* 2004; 13(2): 173-189
- <sup>152</sup> Berner ES, Maisiak RS: Influence of case and physician characteristics on perceptions of decision support systems. *J Am Med Inform Assoc.* 1999; 6(5):428-34.
- <sup>153</sup> Langhals BT. The affect of varying arousal methods upon vigilance and error detection in an automated command and control environment. Thesis. 2001. Faculty of the Graduate School of Engineering and Management of the Air Force Institute of Technology Air University. [cited June 2011] Available from: <http://www.dtic.mil/dtic/tr/fulltext/u2/a390921.pdf>
- <sup>154</sup> Riley V. A general model of mixed-initiative human-machine systems. In *Proceedings of the Human Factors Society. 33<sup>rd</sup> Annual Meeting.* Santa Monica CA; Human Factors and Ergonomics Society; Denver, CO; 1989; 124-128
- <sup>155</sup> Dzindolet M, Pierce L, Beck HP. An examination of the social, cognitive, and motivational factors that affect automation reliance. In D. Andrews, R. Herz, & M. Wolf (Eds.), *Human Factors Issues in Combat Identification.* Farnham, Surrey, UK: Ashgate & Gower; 2010
- <sup>156</sup> Arnold V, Sutton SG. The theory of technology dominance: understanding the impact of intelligent decision aids on decision makers' judgments. *Advances in Accounting Behavioral Res.* 1998; 1: 175-194.
- <sup>157</sup> Ho G, Wheatley D, Scialfa CT. Age differences in trust and reliance of a medication management system. *Interact Comput.* 2005; 17(6): 690–710
- <sup>158</sup> Riley V. Human use of automation. Doctoral dissertation. University of Minnesota; 1994a
- <sup>159</sup> Singh IL, Molloy R, Parasuraman R. Automation-induced “complacency”: Development of the complacency-potential rating scale. *Int J Aviat Psychol.* 1993; 3: 111-122
- <sup>160</sup> Probst AC, Shaffer VA, Chan R. Physician individual differences related to willingness to use a computer-based DSS. *Proceedings: 5<sup>th</sup> Annual Symposium: Graduate Research and Scholarly Projects;* 2009.
- <sup>161</sup> Huber GP. Cognitive Style as a Basis for MIS and DSS Designs: Much ADO About Nothing? *Manage Sci.* 1983; 29(5): 567-579
- <sup>162</sup> Chakraborty I, Hu PJ, Cui D. Examining the effects of cognitive style in individuals' technology use decision making. *Decision Support Systems.* 2008; 45(2): 228-241
- <sup>163</sup> Bergman LG, Fors UG. Computer aided DSM IV diagnostics- acceptance, use and perceived usefulness in relation to users' learning styles. *BMC Med Inform Decis Mak.* 2005; 5:1
- <sup>164</sup> Vessey, I. The theory of cognitive fit: One aspect of a general theory of problem solving? in P. Zhang and D. Galletta (eds.), *Human-computer interaction and management information systems: Foundations,* *Advances in Management Information Systems Series,* Armonk, NY: M.E. Sharpe; 2006.
- <sup>165</sup> Smelcer, JB, Carmel E. The Effectiveness of Different Representations for Managerial Problem Solving: Comparing Tables and Maps. *Decis Sci.* 1997; 28(2): 391-421
- <sup>166</sup> Smelcer, J., & Carmel, E. The effectiveness of different representations for managerial problem solving: Comparing maps and tables. *Decision Sciences,* 1997. 28(2), 391-420

- 
- <sup>167</sup> Dreiseitl S, Binder M. Do physicians value decision support? A look at the effect of decision support systems on physician opinion. *Artif Intell Med.* 2005; 33: 25-30
- <sup>168</sup> Sanchez J, Fisk AD, Rogers WA. What Determines Appropriate Trust of and Reliance on an Automated Collaborative System? Effects of Error Type and Domain Knowledge Control, Automation, Robotics and Vision, 2006. ICARCV. 9th International Conference. 2006; 5: 1-6
- <sup>169</sup> Lee Z, Wagner C, Shin HK. The effect of decision support system expertise on system use behaviour and performance. *Inform Manage.* 2008; 45(6): 349–358
- <sup>170</sup> Bliss JP, Acton SA. Alarm mistrust in automobiles: how collision alarm reliability affects driving. *Appl Ergon.* 2003. 34 (6), 499-509
- <sup>171</sup> de Vries, P, Midden C, Bouwhuis D. The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *Intl J Hum-Comput St.* 2003; 58(6): 719-735
- <sup>172</sup> Wiegmann DA. Agreeing with automated diagnostic aids: A study of users' concurrence strategies. *Hum Factors.* 2002; 44(1): 44-50
- <sup>173</sup> Wickens CD, Xu X. Automation trust, reliability and attention. Institute of Aviation Tech. Report AHFD-02-14 / MAAD-02-2. Savoy: University of Illinois, Aviation Research Lab
- <sup>174</sup> Molloy R, & Parasuraman R. (1996). Monitoring an automated system for a single failure: Vigilance and task complexity effects. *Hum Factors.* 1996; 38, 311-322
- <sup>175</sup> Wickens CD, Helleberg J, Xu X. Pilot maneuver choice and workload in free flight. *Hum Factors,* 2002; 44: 171 – 188
- <sup>176</sup> Sanchez J. Factors that affect trust and reliance on an automated aid, Ph.D. dissertation, Georgia Institute of Technology, May 2006
- <sup>177</sup> Mosier KL, Skitka LJ, Burdick MD, Heers ST. Automation Bias, Accountability, and Verification Behaviors. Human Factors and Ergonomics Society Annual Meeting Proceedings, Cognitive Engineering and Decision Making. 1996; 204-208
- <sup>178</sup> Kantowitz BH, Hanowski RJ, Kantowitz SC. Driver acceptance of unreliable traffic information in familiar and unfamiliar settings. *Hum Factors.* 1997; 39: 164-176
- <sup>179</sup> Tenney YJ, Rogers WH, Pew RW. Pilot opinions on cockpit automation issues. *Int J Aviat Psychol.* 1998; 8: 103-120
- <sup>180</sup> Muir BM. Trust in automation: 1. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics,* 1994; 37: 1905–1922
- <sup>181</sup> Itoh M, Abe G, Tanaka K. Trust in and use of automation: their dependence on occurrence patterns of malfunctions. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics. Piscataway, NJ, 1999; 3: 715-720
- <sup>182</sup> Parasuraman R, Sheridan TB, Wickens CD. Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs. *J Cogn Eng and Decis Mak.* 2008; 2(2): 140–160
- <sup>183</sup> Gibbons AM, Sniezek JA, Dalal RS. Antecedents and consequences of unsolicited versus explicitly solicited advice. In D. Budescu (Chair), Symposium in Honor of Janet Sniezek. Symposium presented at the annual meeting of the Society for Judgment and Decision Making, Vancouver, BC; 2003

- 
- <sup>184</sup> Prinzel LJ, De Vries H, Freeman FG, Mikulka P. Examination of Automation-Induced Complacency and Individual Difference Variates. Technical Memorandum No. TM-2001-211413, NASA Langley Research Center, Hampton, VA; 2001
- <sup>185</sup> Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP. The role of trust in automation reliance. *Int J Hum-Comput St.* 2003; 58(6): 697-718
- <sup>186</sup> Isen AM, Rosenzweig AS, Young MJ. The Influence of Positive Affect on Clinical Problem solving. *Med Decis Making.* 1991;11:221-227
- <sup>187</sup> Cooper RS. Information processing in the judge–adviser system of group decision-making. Unpublished master’s thesis, University of Illinois, Urbana-Champaign; 1991
- <sup>188</sup> Budescu DV, Rantilla AK, Yu H, Karelitz TK. The effects of asymmetry among advisors on the aggregation of their opinions. *Organ Behav Hum Dec.* 2003; 90: 178–194
- <sup>189</sup> Budescu DV, Rantilla AK. Confidence in aggregation of expert opinions. *Acta Psychologica.* 2000; 104: 371–398
- <sup>190</sup> Sniezek JA, Buckley T. Cueing and cognitive conflict in judge–advisor decision making. *Organ Behav Hum Dec.* 1995; 62: 159–174
- <sup>191</sup> Soll JB, Klayman J. Overconfidence in interval estimates. *J Exp Psychol: Learn.* 2004; 30: 299–314.
- <sup>192</sup> Westbrook JI, Gosling AS, Coiera EW. The Impact of an Online Evidence System on Confidence in Decision Making in a Controlled Setting. *Med Decis Mak.* 2005; 25(2): 178-185
- <sup>193</sup> Lee JD, Moray N. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics.* 1992; 35: 1243-1270
- <sup>194</sup> Case K, Sinclair MA, Rani MRA. An experimental investigation of human mismatches in machining. *Proceedings of the Institution of Mechanical Engineers Part B – Journal of Engineering Manufacture.* 1999; 213: 197–201
- <sup>195</sup> Bandura, A. *Self-efficacy: The exercise of control.* New York: Worth Publishers; 1997
- <sup>196</sup> Prinzel LJ. *The Relationship of Self-Efficacy and Complacency in Pilot-Automation Interaction (Technical Memorandum No. TM-2002-211925).* NASA Langley Research Center, Hampton, VA; 2002
- <sup>197</sup> Boksem MAS, Meijman TF, Lorist MM. Effects of mental fatigue on attention: An ERP study. *Cognitive Brain Res.* 2005; 25(1): 107-116
- <sup>198</sup> Baumeister RF. The Psychology of Irrationality. In I Brocas, JD Carrillo: *The Psychology of Economic Decisions: Rationality and well-being.* 2003; p. 1–15,
- <sup>199</sup> Danziger S, Levav J, Avnaim-Pesso L. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences, USA.* 2011; 108(17): 6889–6892
- <sup>200</sup> Wu JH, Shen WS, Lin LM, Greenes R, Bates DW. Testing the technology acceptance model for evaluating healthcare professionals’ intention to use an adverse event reporting system. *Int J Qual Health Care.* 2008; 20(2): 123–129

- 
- <sup>201</sup> Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. *Manag Sci.* 2000;46:186–204.
- <sup>202</sup> Yeh YY, Wickens CD. The dissociation of subjective measures of mental workload and performance. *Hum Factors.* 1988; 30: 111-120
- <sup>203</sup> Wickens CD, Dixon SR. Is there a Magic Number 7 (to the Minus 1)? The Benefits of Imperfect Diagnostic Automation: A Synthesis of the Literature. *Theor Issues Ergonomics Sci.* 2007; 8(3): 201-212
- <sup>204</sup> Kahneman D. *Attention and effort.* Upper Saddle River, NJ: Prentice Hall; 1973
- <sup>205</sup> Endsley MR. Toward a theory of situation awareness in dynamic systems. *Hum Factors,* 1995; 37(1): 32–64
- <sup>206</sup> Durso F, Rawson K, Girrotto S. Comprehension and situation awareness. In F Durso (Ed.), *Handbook of applied cognition.* West Sussex, UK: Wiley; 2007; p. 163–194
- <sup>207</sup> Bagheri N, Jamieson GA. Considering subjective trust and monitoring behavior in assessing automation-induced “complacency”. In: Vicenzi DA, Mouloua M, Hancock PA (Eds), *Human Performance, Situation Awareness, and Automation: Current Research and Trends.* Erlbaum, Mahwah, NJ; 2004; p. 54–59
- <sup>208</sup> Kirlik A. Modeling Strategic Behavior in Human-Automation Interaction: Why an "Aid" Can (and Should) Go Unused. *Hum Factors.* 1993; 35(2): 221-242
- <sup>209</sup> Madhavan P, Wiegmann DA. Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Hum Factors.* 2007; 49: 773-785
- <sup>210</sup> Madhavan P, Wiegmann DA, Lacson FC. Automation failures on tasks easily performed by operators undermine trust in automated aids. *Hum Factors.* 2006; 48: 241-256
- <sup>211</sup> Madhavan P, Wiegmann DA. Expertise levels of human versus automated decision aids influence response biases in a visual search task. *Proceedings of the Human Factors and Ergonomics Society 50<sup>th</sup> Annual Meeting.* 2006; 50(3): 230-234
- <sup>212</sup> Singh IL, Molloy R, Parasuraman R. Automation-induced monitoring inefficiency: role of display location. *Int J Hum-Comput St.* 1997; 46(1): 17-30
- <sup>213</sup> Meyer J. Effects of warning validity and proximity on responses to warnings. *Hum Factors.* 2001; 43(4): 563-572
- <sup>214</sup> Jamieson GA, Wang L. *Developing Human-Machine Interfaces to Support Appropriate Trust and Reliance on Automated Combat Identification Systems.* 2008. Cognitive Engineering Laboratory. Defence R&D Canada - Toronto, Toronto ONT (CAN); Toronto Univ, Toronto Ont (CAN); 2008 [cited June 2011] Available from: <http://cradpdf.drdc.gc.ca/PDFS/unc75/p529837.pdf>
- <sup>215</sup> Tseng S, Fogg BJ. Credibility and computing technology. *Communications of the Association for Computing Machinery.* 1999; 42(5)
- <sup>216</sup> Kim J. Moon JY. Designing towards emotional usability in customer interfaces – Trustworthiness of cyber-banking system interfaces. *Interact Comput.* 1998; 10: 1–29.

- 
- <sup>217</sup> Karvonen K., Parkkinen J. Signs of trust: A semiotic study of trust formation in the Web. In: MJ Smith, G Salvendy, D Harris, RJ Koubek (Eds), *First International Conference on Universal Access in Human-Computer Interaction*. Mahwah, NJ: Erlbaum. 2001; 1076-1080
- <sup>218</sup> Gong L, Nass C, Simard C, Takhteyev Y. When non-human is better than semi-human: Consistency in speech interfaces. In: MJ Smith, G Salvendy, D Harris, R. Koubek (Eds), *Usability evaluation and interface design: Cognitive engineering, intelligent agents, and virtual reality*. Mahwah, NJ: Erlbaum. 2001; 1558–1562
- <sup>219</sup> Berner ES, Maisiak RS, Heudebert GR, Young KR. Clinician Performance and Prominence of Diagnoses Displayed by a Clinical Diagnostic Decision Support System. *AMIA Annu Symp Proc*. 2003; 76-80
- <sup>220</sup> Teich, JM, Merchia, PR, Schmiz, JL, Kuperman, GJ, Spurr, CD, Bates DW. Effects of computerized physician order entry on prescribing practices. *Arch Intern Med*. 2000; 160(18): 2741-7
- <sup>221</sup> Meyer J, Feinshreiber L, Parmet Y. Levels of automation in a simulated failure detection task. In *IEEE International Conference on Systems, Man and Cybernetics*. 2003; 2101- 2106
- <sup>222</sup> Parasuraman R, Cosenzo KA, De Visser E. Adaptive Automation for Human Supervision of Multiple Uninhabited Vehicles: Effects on Change Detection, Situation Awareness, and Mental Workload. *Mil Psychol*. 2009; 21(2): 270 - 297
- <sup>223</sup> Kaber EB, Wright MC, Prinzel MP. Adaptive Automation of Human-Machine System Information-Processing Functions. *Hum Factors*, 2005; 47(4): 730-741
- <sup>224</sup> Billings RS, Scherer LL. The effects of response mode and importance on decision-making strategies: judgment versus choice, *Organ Behav Hum Dec*. 1988; 41(1): 1–19
- <sup>225</sup> Moyer RS, Landauer TK, Nature. Time required for judgments of numerical inequality. *Nature*. 1967; 215: 1519-1520
- <sup>226</sup> Zimmer K. Examining the validity of numerical ratios in loudness fractionation. *Percept Psychophys*. 2005; 67: 569–579
- <sup>227</sup> Kostopoulou O, Oudhoff J, Nath R, Delaney BC, Munro CW, Harries C, Holder R. Predictors of Diagnostic Accuracy and Safe Management in Difficult Diagnostic Problems in Family Medicine. *Med Decis Making*. 2008; 28(5): 668-680
- <sup>228</sup> Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: Toward a unified view. *MIS QUART*, 2003; 27(3): 425-478
- <sup>229</sup> Workman M. Expert decision support system use, disuse, and misuse: a study using the theory of planned behaviour. *Comput Hum Behav*. 2005; 21(2): 211–231
- <sup>230</sup> Dzindolet M, Pierce L, Beck HP. An examination of the social, cognitive, and motivational factors that affect automation reliance. In D. Andrews, R. Herz, & M. Wolf (Eds), *Human Factors Issues in Combat Identification*. Farnham, Surrey, UK: Ashgate & Gower; 2010
- <sup>231</sup> Alberdi E, Strigini L, Povyakalo AA, Ayton P. Why are people's decisions sometimes worse with computer support? *SAFECOMP '09 Proceedings of the 28th International Conference on Computer Safety, Reliability, and Security* Springer-Verlag Berlin, Heidelberg; 2009
- <sup>232</sup> Arnold V, Collier PA, Leech SA, Sutton SG. Impact of intelligent decision aids on expert and novice decision-makers' judgments. *Acc Finan*. 2004; 44(1): 1-26

- 
- <sup>233</sup> Hampton C. Determinants of reliance: An empirical test of the theory of technology dominance. *Int J Accounting Inf*, 2006. 6(1): 217-240
- <sup>234</sup> Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O, Peacock R. Storylines of research in diffusion of innovation: a meta-narrative approach to systematic review. *Soc Sci Med*. 2005; 61: 417–430
- <sup>235</sup> Noy N, McGuinness, D. Ontology development 101: A guide to creating your first ontology. Stanford Knowledge Aytms Laboratory Technical Report. KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880; 2001
- <sup>236</sup> Yu AC. Methods in biomedical ontology. *J Biomed Inform*. 2006; 39: 252-66
- <sup>237</sup> Bright T, Furaya EY, Kuperman GJ, Cimino JJ, Bakken S. Development and Evaluation of an Ontology for Guiding Appropriate Antibiotic Prescribing. *J Biomed Inform*. 2012; 45(1): 120-128
- <sup>238</sup> Lopez JM, Gil R, Garcia R, Cearreta I, Garay N. Towards an ontology for describing emotions. Proceedings of the 1st world summit on The Knowledge Society: Emerging Technologies and Information Systems for the Knowledge Society. 2008: 96-104
- <sup>239</sup> Nykanen P. On the Ontology of a Decision Support System in Health Informatics. In GA Forgionne, J Gupta, M Mora. Decision making support systems. IGI Publishing Hershey, PA, USA; 2003
- <sup>240</sup> Masolo C, Borgo S, Gangemi A, Guarino A, Oltramari A. Wonderweb Deliverable D18: Ontology library (final). Technical report. Laboratory for Applied Ontology – ISTC-CNR, Trento; 2003.
- <sup>241</sup> Ajzen I, Fishbein M. Understanding attitudes and predicting social behavior. Englewood Cliffs, NJ: Prentice-Hall; 1980
- <sup>242</sup> Eapen BR. ONTOderm – a domain ontology for dermatology. *Dermatology Online Journal*. 2008; 14(6): 6
- <sup>243</sup> Corbridge, C, Rugg G, Major NP, Shadbolt NR, Burton AM. Laddering: technique and tool use in knowledge acquisition. *Knowl Acquis*. 1994; 6: 315-341.
- <sup>244</sup> Bannister D, Fransella F. Inquiring Man: The Psychology of Personal Constructs (3<sup>rd</sup> Ed). London: Routledge; 1989.
- <sup>245</sup> Shaw MLG, Gaines. KITTEN: Knowledge initiation and transport tools for experts and novices. In J Boose, BR Gaines Eds. Knowledge-based systems Volume 2: Knowledge Acquisition Tools for Expert Systems. New York: Academic Press; 1988: 309-338
- <sup>246</sup> Stewart V, Stewart A. Business applications of repertory grid. London: McGraw-Hill; 1981
- <sup>247</sup> Han YY, Carcillo JA, Venkataraman ST, et al. Unexpected Increased Mortality After Implementation of a Commercially Sold Computerized Physician Order Entry System. *Pediatrics*. 2005;116(6):1506–1512.
- <sup>248</sup> Ornstein C. Hospital heeds doctors, suspends use of software. *Los Angeles Times*; 2003
- <sup>249</sup> Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional integration. *Hum Factors*. 2010; 52(3):381-410
- <sup>250</sup> Juni P, Witschi A, Bloch R, Egger M. The Hazards of Scoring the Quality of Clinical Trials for Meta-analysis. *J Am Med Assoc*. 1999; 282:1054-1060.

- 
- <sup>251</sup> Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977; 33: 159-174
- <sup>252</sup> Alberdi E, Povyakalo A, Strigini L, Ayton P. Effects of incorrect CAD output on human decision making in mammography. *Acad Radiol*. 2004; 11( 8): 909-918
- <sup>253</sup> Tsai TL, Fridsma DB, Gatti G. Computer decision support as a source of interpretation error: the case of electrocardiograms. *J Am Med Inform Assoc*. 2003; 10(5): 478-483
- <sup>254</sup> Hillson SD, Connelly DP, Liu Y. The Effects of Computer-Assisted Electrocardiographic Interpretation on Physicians' Diagnostic Decisions. *Med Decis Making*, 1995; 15(2): 107-112
- <sup>255</sup> Bogun F. Misdiagnosis of atrial fibrillation and its clinical consequences. *Am J Med*. 2004; 117(9):636-42
- <sup>256</sup> Guerlain S. The antibody identification assistant (AIDA), an example of a cooperative computer support system. *Systems, Man and Cybernetics, IEEE Conference paper*. 1995; 2:1909-1914
- <sup>257</sup> Wyatt J. Lessons learned from the field trial of ACORN, an expert system to advise on chest pain. In: Barber B, Cao D, Qin D, eds. *Proc. Sixth World Conference on Medical Informatics*, Singapore. Amsterdam: North Holland 1989: 111-115
- <sup>258</sup> Human Error in Aviation. Available at: <http://www.carrielee.net/pdfs/HumanError.pdf> [cited on 01/01/2011]
- <sup>259</sup> Human Error Rates. Available at: <http://panko.shidler.hawaii.edu/HumanErr/> [cited on 01/01/2011]
- <sup>260</sup> Ruben G, George A, Chinn DJ, Richardson C. Errors in general practice: development of an error classification and pilot study of a method for detecting errors. *Qual Saf Health Care*. 2003;12:443-447
- <sup>261</sup> Westbrook JI, Coiera E, Gosling AS. Do online retrieval systems help experienced clinicians answer clinical questions? *J Am Med Assoc*. 2005; 12: 315-321
- <sup>262</sup> McKibbin A, Fridsma DB. Effectiveness of clinician-selected electronic information resources for answering primary care physicians' information needs. *J Am Inform Med Assoc*. 2006; 13(6): 653-659
- <sup>263</sup> Southern WN. The Effect of Erroneous Computer Interpretation of ECGs on Resident Decision Making. *Med Decis Making*. 2009; 29(3): 372-376
- <sup>264</sup> Sowan AK, Gafoor m, Soeken K, Mills ME, Johantgen M, Vaidya V. A Comparison of Medication Administrations Errors Using CPOE Orders vs. Handwritten Orders for Pediatric Continuous Drug Infusions. *AMIA Annu Symp Proc*. 2006; 1105
- <sup>265</sup> Durieux P. A clinical decision support system for prevention of venous thromboembolism - Effect on physician behaviour. *J Am Med Assoc*. 2000; 283(21): 2816-2821
- <sup>266</sup> Guerlain S. Factors influencing the cooperative problem-solving of people and computers. *Proc Hum Factors Ergonom Soc 37th Annual Meeting, Human Factors and Ergonomics Society, Santa Monica, CA*. 1993; 387-391
- <sup>267</sup> Ikeda M, Ishigaki T, Yamauchi K. A Signal-detection Experiment Measuring the Effect of Computer-aided Detection on Radiologists' Performance. *Med Dcis Making*. 2000; 20(3); 343-351
- <sup>268</sup> Li F, Li Q, Engelmann R, Aoyama M, Sone S, MacMahon D, Doi K. Improving radiologists' recommendations with computer-aided diagnosis for management of small nodules detected by CT. *Acad Radiol*. 2006; 13(8): 943-950

- 
- <sup>269</sup> Fenton JJ, Taplin SH, Carney PA, Abraham L, Sickles EA, D'Orsi C, Berns EA, Cutter G, Hendrick RE, Barlow WE, Elmore JG. Influence of Computer-Aided Detection on Performance of Screening Mammography. *New Eng J Med*. 2007; 356:1399-1409
- <sup>270</sup> Morimoto T, Iinuma G, Shiraishi J, Arai Y, Moriyama N, Beddoe G, Nakijima Y. Computer-aided detection in computed tomography colonography: current status and problems with detection of early colorectal cancer. *Jpn J Radiol*. 2008; 26(5): 261-269
- <sup>271</sup> Zheng B, Ganott MA, Britton CA, Hakim CA, Hardesty LA. Soft-Copy Mammographic Readings with Different Computer-assisted Detection Cuing Environments: Preliminary Findings. *Radiology*. 2001; 221(3): 633-640
- <sup>272</sup> Hadjiski L, Chan HP, Sahiner B, Helvie MA, Roubidoux MA. Improvement in Radiologists' Characterization of Malignant and Benign Breast Masses on Serial Mammograms with Computer-aided Diagnosis: An ROC Study. *Radiology*. 2004; 255-265
- <sup>273</sup> Helvie MA, Hadjiski L, Makariou E, Chan HP, Petrick N. Sensitivity of Noncommercial Computer-aided Detection System for Mammographic Breast Cancer Detection: Pilot Clinical Trial. *Radiology*. 2004; 231(1):208-214
- <sup>274</sup> Petrick N, Haider M, Summers RM, Yeshwant SC, Brown L, Iuliano EM, Louie A, Choi JR, Pickhardt PJ. CT Colonography with Computer-aided Detection as a Second Reader: Observer Performance Study. *Radiol*. 2008; 246: 148-156
- <sup>275</sup> Quek ST, Thng CH, Khoo JBK, Koh WL. Radiologists' detection of mammographic abnormalities with and without a computer-aided detection system. *Australas Radiol*. 2003; 47(3): 257-260
- <sup>276</sup> Gakenheimer DC. The efficacy of a computerized caries detector in intraoral digital radiography. *J Am Dent Asso*. 2002; 133(7): 883-89
- <sup>277</sup> Moberg K. Computed assisted detection of interval breast cancers. *Eur J Radiol*, 2001; 39(2): 104-110
- <sup>278</sup> Kobayashi T, Xu XW, MacMahon H, Metz CE, Doi K. Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. *Radiol*. 1996; 199: 843-848
- <sup>279</sup> Xu X, Wickens CD, Rantanen EM. Effects of conflict alerting system reliability and task difficulty on pilots' conflict detection with cockpit display of traffic information. *Ergonomics*. 2007; 50(1):112-30.
- <sup>280</sup> Dixon SR, Wickens CD. Automation Reliability in Unmanned Aerial Vehicle Control: A Reliance-Compliance Model of Automation Dependence in High Workload. *Hum Factors*. 2006; 48(3): 474-486
- <sup>281</sup> Mosier KL, Skitka LJ, Heer S, Burdick M. Automation bias - Decision making and performance in high-tech cockpits. *Int J Aviat Psychol*. 1997; 8(1): 47-63
- <sup>282</sup> Bailey NR, Scerbo MW. The effects of operator trust, complacency potential, and task complexity on monitoring a highly reliable automated system.. *Dissertation Abstracts International: Section B: The Sciences and Engineering*. US, ProQuest Information & Learning; 2005.
- <sup>283</sup> Ma R, Kaber DB. Situation awareness and driving performance in a simulated navigation task. *Ergonomics*. 2007; 50(8); 1352-1364
- <sup>284</sup> Moray N, Inagaki T, Itoh M. Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *J Exp Psychol Appl*. 2000; 6(1): 44-58

- 
- <sup>285</sup> Marten K, Seyfarth T, Auer F, Wiener E, Grillhösl A, Obenauer S, Rummeny EJ, Engelke C. Computer-assisted detection of pulmonary nodules: performance evaluation of an expert knowledge-based detection system in consensus reading with experienced and inexperienced chest radiologists. *Eur Radiol.* 2004; 14(10): 1930-1938
- <sup>286</sup> Walsham AC, Roberts HC, Kashani HM, Mongiardi CN, NG YL. The use of computer-aided detection for the assessment of pulmonary arterial filling defects at computed tomographic angiography. *J Comput Assist Tomogr.* 2008; 32(6): 913-918
- <sup>287</sup> Sarter NB, Schroeder B. Supporting Decision Making and Action Selection under Time Pressure and Uncertainty: The Case of In-Flight Icing. *Hum Factors.* 2001; 43(4): 573-583
- <sup>288</sup> McGuirol JM, Sarter NB. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Hum Factors* 2006; 48(4): 656-665
- <sup>289</sup> Prinzel LJ, Freeman FG, Prinzel HD. Individual Differences in Complacency and Monitoring for Automation Failures. *Indiv Diff Res.* 2005; 3(1): 27-49
- <sup>290</sup> Yeh M. Display Signaling in Augmented Reality: Effects of Cue Reliability and Image Realism on Attention Allocation and Trust Calibration. *Hum Factors.* 2001; 43(3): 355-365
- <sup>291</sup> Singh IL, Molloy R, Parasuraman R. Individual differences in monitoring failures of automation. *J Gen Psychol.* 1993b; 120: 357-373
- <sup>292</sup> Dieckmann A, Dippold K, Dietrich H. Compensatory versus non-compensatory models for predicting consumer preferences. *Judgm Decis Mak* 2009;4:200-213
- <sup>293</sup> Burdick MD, Skitka LJ, Mosier KL, Heers S. The ameliorating effects of accountability on automation bias. *Proceedings of the 3rd Symposium on Human Interaction with Complex Systems.* 1996; p. 142
- <sup>294</sup> Grubb PL, Warm JS, Dember WN, Berch DB. Effects of Multiple-Signal Discrimination on Vigilance Performance and Perceived Workload. *Proceedings of the Human Factors and Ergonomics Society.* 1995; 39(21): 1360-1364
- <sup>295</sup> McFadden SM, Vimalachandaran A, Blackmore E. Factors affecting performance on a target monitoring task employing an automatic tracker. *Ergonomics.* 2004; 46(3): 257-280
- <sup>296</sup> Mosier KL, Skitka LJ, Dunbar M, et al. Aircrews and automation bias: the advantages of teamwork? *Int J Aviat Psychol* 2001;11:1-14
- <sup>297</sup> Masalonis AJ. Effects of training operators on situation-specific automation reliability. *Conference on Systems, Man and Cybernetics, IEEE International;* 2003
- <sup>298</sup> Parasuraman R, Mouloua M, Molloy R. Effects of adaptive task allocation on monitoring of automated systems. *Hum Factors.* 1996; 38(4): 665-679
- <sup>299</sup> Bonaccio S, Dalal RS. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes.* Volume 101, Issue 2, November 2006, Pages 127-151
- <sup>300</sup> Langer, EJ. *The psychology of control.* Beverly Hills, CA: Sage; 1983.
- <sup>301</sup> Goddard K, Roudsari A, Wyatt JC. Automation Bias - a hidden issue or clinical decision support system use. *Stud Health Technol Inform.* 2011

- 
- <sup>302</sup> Issenberg SB, McGaghie WC, Hart IR, Mayer JW, Felner JM, Petrusa ER, Waugh RA, Brown DD, Safford RR, Gessner IH, Gordon DL, Ewy GA. Simulation technology for health care professional skills training and assessment. *J Am Med Assoc.* 1999; 282(9): 861-6
- <sup>303</sup> Walton RT, Gieri C, Yudkin P, Mistry H, Vessey MP, Fox J. Evaluation of computer support for prescribing (CAPSULE) using simulated cases. *Brit Med J.* 1997;315:791
- <sup>304</sup> Dillon GF, Clyman SG, Clauser BE, Margolis MJ. The introduction of computer-based simulations into the United States medical licensing examination. *Acad Med.* 2002; 77(10 Suppl): S94-6
- <sup>305</sup> Cohen J. *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum; 1988
- <sup>306</sup> Rieskamp J, Hoffrage U. Inferences under time pressure: How opportunity costs affect strategy selection. *Acta Psychol.* 2008; 127(2); 258–276
- <sup>307</sup> Kittler JE, Menard W, Phillips KA. Weight concerns in individuals with body dysmorphic disorder. *Eating Behav.* 2007; 8(1), 115-120
- <sup>308</sup> Gino F, Moore DA. Effects of task difficulty on use of advice. *J Behav Decis Making.* 2007; 20(1): 21-35
- <sup>309</sup> King WR, Epstein BJ. Assessing information system value: an experimental study. *Decis Sci.* 1982; 13: 34-45
- <sup>310</sup> Fowler FJ. *Survey Research Methods* (2nd ed) Volume 1: Applied Social Research Methods Series, Sage Publication; 1993
- <sup>311</sup> NHS Common User Interface (CUI) Programme. [cited Jun 2011] Available from: <http://www.cui.nhs.uk/Pages/NHSCommonUserInterface.aspx>
- <sup>312</sup> Avery AJ, Savelyich BSV, Sheikh A, Morris CJ, Bowler I, Teasdale. Improving general practice computer systems for patient safety: qualitative study of key stakeholders. *Qual Saf Health Care.* 2007;16: 28–33
- <sup>313</sup> Gordon JS, Akers L, Severson HH, Danaher BG, Boles SM. Successful participant recruitment strategies for an online smokeless tobacco cessation program. *Nicotine Tob Res.* 2006;8 Suppl 1: S35-41
- <sup>314</sup> Evangelista F, Albaum G, Poon P. An empirical test of alternative theories of survey response behavior. *J Market Res Soc.* 1999; 41(2): 227-244
- <sup>315</sup> Thomas Dr, Fenner DP. The central tendency effect in stimulus generalization: the effect of sex of subject. *Am J Psychol.* 1978; 91(3): 509-22
- <sup>316</sup> Bates DW et al. Ten Commandments for effective decision support: making the practice of evidence-based medicine a reality. *J Am Med Inform Assoc.* 2003;10(6):523-530
- <sup>317</sup> Liu J, Wyatt J, Altman D. Decision tools in healthcare: focus on the problem, not the solution. *BMC Med Inform Decis Mak.* 2006;6:4
- <sup>318</sup> Shiffman RN, Liaw Y, Brandt CA, et al. Computer-based guideline implementation systems: a systematic review of functionality and effectiveness. *J Am Med Inform Assoc.* 1999;6:104-114

---

<sup>319</sup> Friedman CP, Wildemuth BM, Muriuki M, Gant SP, Downs SM, Twarog RG, de Bliet R. A Comparison of Hypertext and Boolean Access to Biomedical Information. Proc AMIA Annu Fall Symp. 1996; 2–6.

<sup>320</sup> Stocks N, Gunnell D. What are the characteristics of general practitioners who routinely do not return postal questionnaires: a cross sectional study. J Epidemiol Community Health. 2000; 54:940-941

<sup>321</sup> Holbrook A, Krosnick J, Pfent A. The Causes and Consequences of Response Rates in Surveys by the News Media and Government Contractor Survey Research Firms. In Advances in telephone survey methodology, ed. JM Lepkowski, NC Tucker, JM Brick, ED De Leeuw, L Japiec, PJ Lavrakas, MW Link, RL Sangster. New York: Wiley; 2007

<sup>322</sup> Schedlbauer A, Prasad V, Mulvaney C, et al. What evidence supports the use of computerized alerts and prompts to improve clinicians' prescribing behavior? J Am Med Inform Assoc. 2009; 16:531-8.

<sup>323</sup> Parasuraman R, Sheridan TB, Wickens CD. A Model for Types and Levels of Human Interaction with Automation, IEEE Transactions on Systems, Man, and Cybernetics. 2000; 30(3): 286-29

<sup>324</sup> Reason J. Human Error. Cambridge University Press; 1990

<sup>325</sup> Phansalkar S, Edworthy J, Hellier E, Seger DL, Schedlbauer A, Avery AJ, Bates DW. A review of human factors principles for the design and implementation of medication safety alerts in clinical information systems. J Am Med Inform Assoc. 2010;17: 493-501

<sup>326</sup> Teich, JM, Merchia, PR, Schmiz, JL, Kuperman, GJ, Spurr, CD, Bates DW. Effects of computerized physician order entry on prescribing practices. Arch Intern Med. 2000; 160(18): 2741-7

<sup>327</sup> Leveson NG, Turner CS. An investigation of the Therac-25 accidents. Computer 1993;18-41