Statement of originality

This is to certify that this thesis is the product my own work. I also certify that all the assistance I received in preparing this thesis has been acknowledged.

Ploutarchos Kourtidis

Copyright declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-NonCommercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Influences of a diagnostic decision aid on clinical reasoning

Ploutarchos Kourtidis

Faculty of Medicine
Department of Surgery and Cancer
Imperial College London

Supervisors: Dr Olga Kostopoulou, Professor Brendan Delaney

Thesis submitted for the degree of Doctor of Philosophy

June 2022

Abstract

Previous research has highlighted the importance of physicians' early hypotheses for their final diagnostic decisions. It has also been shown that diagnostic accuracy improves when physicians are presented with a list of diagnostic suggestions to consider at the start of the patient encounter. The psychological mechanisms underlying this improvement in accuracy have not yet been explored. It is possible that the provision of diagnostic suggestions, early in the clinical consultation, disrupts physicians' intuitive thinking and reduces their initial diagnostic confidence. This possibly encourages physicians to seek more patient information before reaching a conclusion, evaluate this information in a more balanced fashion and be more inclined to change diagnosis when encountering new information.

Three online studies explored the effects of early diagnostic suggestions, provided by a hypothetical decision aid, on different aspects of the diagnostic process. General Practitioners (GPs) assessed up to two patient scenarios with and without the decision aid. We measured associations with certainty in the initial diagnostic hypothesis (initial confidence), information search and evaluation, and frequency of diagnostic changes. We detected mainly trends that were not statistically significant but were in the expected direction. We did measure a statistically significant impact of the decision aid on initial confidence (Study 1) and diagnostic changes (Study 3). We also measured significant associations between participants' Actively Open-Minded Thinking (Baron, 2019) and GPs' initial confidence and information search: GPs with higher scores in Actively Open-minded Thinking (AOT) scale had lower confidence in their initial diagnosis (Studies 1 and 3) and engaged in a more extensive search of information before providing their conclusion (Studies 2 and 3).

In all three studies, initial confidence was found to have significant downstream effects: it was associated with less extensive information search, more biased evaluation of neutral information, fewer diagnostic changes when new information was encountered, and a shorter list of differential diagnoses recorded. A systematic review on initial confidence and resistance to change was subsequently conducted to assess the evidence for this phenomenon and found that initial confidence can determine whether, and to what extent, an initial judgement tends to change over time.

In conclusion, this thesis found that early diagnostic suggestions may impact physicians' confidence, information search and evaluation, but the most prominent effects on behaviour are determined by the physician's confidence in their initial diagnostic hypothesis.

Table of Contents

List of Tables 8

List of Figures 10

CHAPTER 1: General introduction: biases in decision making and debiasing techniques 11

1.1. Overconfidence 14
1.2. Confirmation bias 20
1.2.1. Selective information search 20
1.2.2. Biased information evaluation 22
1.3. Premature closure 28
1.4. Debiasing strategies 32
1.4.1. Decision aids 34
1.5. Research questions and objectives 37

CHAPTER 2 : Effects of diagnostic suggestions on physicians' diagnostic confidence (Study 1) 39

2.1. Introduction 39
2.2. Research questions and objectives 39
2.3. Hypotheses 40
2.4. Methods 40
2.4.1. Participants 40
2.4.2. Sample size 40
2.4.3. Recruitment 41
2.4.4. Materials 41
2.4.4.1. Actively Open-minded Thinking scale (AOT) 42
2.4.5. Design 44
2.4.6. Procedure 44
2.5. Analyses 48
2.5.1. Dependent variables 48
2.5.2. Independent variables 48
2.5.3. Preparing the data for analysis 48
2.5.4. Statistical tests 49

2.5.5. Data exclusion 50
2.6. Results 50
2.6.1. Change in initial certainty (H1) 51
2.6.2. Change in diagnosis (H2) 54
2.6.3. Exploratory analyses 57
2.6.3.1. Actively Open-minded Thinking 57
2.6.3.2. Initial diagnosis 58
2.6.3.3. Initial certainty 58
2.6.3.4. Final diagnosis 59
2.6.3.5. Final certainty 60
2.6.3.6. Number of investigations 61
2.6.3.7. Gender 61
2.6.3.8. Experience 61
2.7. Discussion 61
CHAPTER 3: Effects of diagnostic suggestions on physicians' confidence, information search and evaluation (Study 2) 67
3.1. Introduction 67
3.2. Research questions and objectives 68
3.3. Hypotheses 68
3.4. Methods 69
3.4.1. Participants 69
3.4.2. Sample size 69
3.4.3. Recruitment 69
3.4.4. Materials 70
3.4.5. Design 73
3.4.6. Procedure 73
3.5. Analyses 79
3.5.1. Dependent variables 79
3.5.2. Independent variables 80
3.5.3. Preparing the data for analysis 81
3.5.4. Statistical tests 81
3.5.5. Data exclusion 82
3.6. Results 82
3.6.1. Initial and final diagnoses 83

3.6.2. Number of cues requested (H_1) 85
3.6.3. Perceived cue support (H_2) 88
3.6.4. Change in initial certainty (H_3) 91
3.6.5. Change in diagnosis (H_4) 94
3.6.6. Exploratory analyses 96
3.6.6.1. Initial certainty 96
3.6.6.2. Final certainty 97
3.6.6.3. Number of differential diagnoses 98
3.6.6.4. Appropriateness of management 99
3.6.6.5. Gender 100
3.6.6.6. Experience 100
3.6.6.7. Actively Open-minded Thinking 100
3.7. Discussion 100
CHAPTER 4 : Removing GPs' own diagnoses from the list of suggestions – a follow up (Study 3) 105
4.1. Introduction 105
4.2. Research questions and objectives 105
4.3. Hypotheses 107
4.4. Methods 107
4.4.1. Participants 107
4.4.2. Sample size 107
4.4.3. Recruitment 108
4.4.4. Materials 108
4.4.5. Design 108
4.4.6. Procedure 109
4.5. Analyses 109
4.5.1. Dependent variables 110
4.5.2. Independent variables 111
4.5.3. Preparing the data for analysis 111
4.5.4. Statistical tests 111
4.5.5. Data exclusion 112
4.6. Results 112
4.6.1. Initial and final diagnoses 114
4.6.2. Number of cues requested (H_1) 116

4.6.3. Perceived cue support (H_2) 119
4.6.4. Change in initial certainty (H_3) 124
4.6.5. Change in diagnosis (H_4) 127
4.6.6. Exploratory analyses 130
4.6.6.1. Initial certainty (H5) 130
4.6.6.2. Final certainty (H6) 131
4.6.6.3. Number of differential diagnoses 132
4.6.6.4. Appropriateness of management 133
4.6.6.5. Gender 133
4.6.6.6. Experience 134
4.6.6.7. Actively Open-minded Thinking 134
4.6.7. Attitudes towards the decision aid 135
4.7. Discussion 135
CHAPTER 5: Initial confidence in a judgement and changing of mind: a systematic review 141
5.1. Introduction 141
5.1.1. Theoretical and empirical background 141
5.1.2. Objectives and links to Studies 1, 2 and 3 145
5.2. Methods 146
5.2.1. Search strategy 146
5.2.2. Inclusion and exclusion criteria 148
5.2.3. Reference management 148
5.2.4. Selection and data collection process 149
5.2.5. Quality assessment 149
5.3. Results 150
5.3.1. Methodological quality 152
5.3.2. Studies that found evidence for a relationship between initial confidence and change of judgement 153
5.3.2.1. Medical studies 154
5.3.2.2. Non-medical studies 158
5.3.3. Studies that did not find evidence for a relationship between initial confidence and change of judgement 161
5.3.3.1. Medical studies 162
5.3.3.2. Non-medical studies 164
5.3.4. Studies with mixed results 166

5.3.4.1. Medical studies 166
5.3.4.2. Non-medical studies 168
5.4. Discussion 169
5.4.1. Summary 169
5.4.2. Limitations 171
5.4.3. Implications for future research 173
CHAPTER 6: General discussion 175
6.1. Summary 175
6.2. General conclusions 178
CHAPTER 7 : Appendices 185
7.1. Appendix 1 185
7.2. Appendix 2 187
7.3. Appendix 3 188
7.4. Appendix 4 190
7.5. Appendix 5 196
7.6. Appendix 6 198
7.7. Appendix 7 200
7.8. Appendix 8 216
7.8.1. Piloting for Study 1 216
7.8.2. Piloting for Study 2 217
7.8.3. Piloting for Study 3 220
Patient and Public Involvement (PPI) 221
References 222
Acknowledgements 240

List of Tables

Table 2.1. Results by condition (with and without diagnostic suggestions). 50
Table 2.2. Regression model of change in initial Certainty 52
Table 2.3. Regression model of change in initial certainty (James) 52
Table 2.4. Regression model of change in initial certainty (Derek) 53
Table 2.5. Regression model including variable “Condition order” 53
Table 2.6. Measurements of initial certainty and change in initial certainty across scenarios and conditions. 54
Table 2.7. Regression model of change in diagnosis 55
Table 2.8. Regression model including Condition order 56
Table 2.9. Frequency of diagnostic changes across patient scenarios. 56
Table 2.10. Regression model of change in diagnosis (James). 57
Table 2.11. Regression model of change in diagnosis (Derek). 57
Table 2.12. Most frequent initial diagnoses across the two patient scenarios. 58
Table 2.13. Most frequent final diagnoses across the two patient scenarios. 60
Table 2.14. Certainty measurements across conditions and scenarios. 60
Table 3.1. Results by condition (with and without the decision aid). 83
Table 3.2. Proportion of initial diagnoses across patient scenarios. 84
Table 3.3. Final diagnoses across patient scenarios. 85
Table 3.4. Regression models of number of cues requested. 87
Table 3.5. Number of cues requested across patient scenarios. 88
Table 3.6. Regression model of perceived cue support. 89
Table 3.7. Mean cue ratings in chest pain scenario (David), MSK version. 90
Table 3.8. Mean cue ratings in chest pain scenario (David), cardiac version. 90
Table 3.9. Mean cue ratings in constipation scenario (Barbara), IBS version. 91
Table 3.10. Mean cue ratings in constipation scenario (Barbara), colorectal cancer version 91
Table 3.11. Change in initial certainty across patient scenarios. 93
Table 3.12. Measurements of initial certainty. 93
Table 3.13. Regression model of change in initial certainty. 94
Table 3.14. Frequency of diagnostic changes across patient scenarios. 95
Table 3.15. Regression model of change in diagnosis 96
Table 3.16. Initial certainty across patient scenarios and conditions. 97

Table 3.17. Final certainty across patient scenarios and conditions. 98
Table 3.18. Certainty measurements across conditions and scenarios. 98
Table 3.19. Number of differential diagnoses that GPs recorded after providing their final diagnosis. 99
Table 4.1. Results across conditions. 113
Table 4.2. Proportion of initial diagnoses across patient scenarios. 115
Table 4.3. GPs' final diagnoses across patient scenarios. 116
Table 4.4. Regression model of number of cues requested (Control versus Aided 2). 117
Table 4.5. Regression model of number of cues requested (Aided 1 versus Aided 2). 118
Table 4.6. Number of cues requested across scenarios and conditions. 119
Table 4.7. Regression model of perceived cue support (Control versus Aided 2). 121
Table 4.8. Regression model of perceived cue support (Aided 1 versus Aided 2). 121
Table 4.9. Mean cue ratings in chest pain scenario, MSK steer. 122
Table 4.10. Mean cue ratings in chest pain scenario, cardiac steer. 123
Table 4.11. Mean cue ratings in constipation scenario, IBS version. 123
Table 4.12. Mean cue ratings in constipation scenario, colorectal cancer version. 124
Table 4.13. Regression model of change in initial certainty (Control versus Aided 2). 126
Table 4.14. Regression model of change in initial certainty (Aided 1 versus Aided 2). 126
Table 4.15. Changes in certainty across conditions and patient scenarios. 127
Table 4.16. Regression model of diagnostic change (Control versus Aided 2). 128
Table 4.17. Regression model of diagnostic change (Aided 1 versus Aided 2). 129
Table 4.18. Frequency of changes in diagnosis across conditions and scenarios. 130
Table 4.19. Initial certainty per scenario and condition. 131
Table 4.20. Final certainty per scenario and condition. 132
Table 4.21. Number of differential diagnoses that GPs recorded by scenario and scenario version. 133
Table 5.1. Main components of the search and search terms used in Web of Science database. 147
Table 5.2. Supportive studies. 154
Table 5.3. Non-supportive studies. 162

Table 5.4. Studies providing partial support. 166

List of Figures

Figure 2.1. Initial diagnosis and certainty measurement scale. 45
Figure 2.2. The decision aid: a list of diagnostic suggestions (James scenario). 46
Figure 2.3. Flow of activity 47
Figure 3.1. Chest pain scenario. 71
Figure 3.2. Constipation scenario. 72
Figure 3.3. Initial diagnosis and certainty scale (chest pain scenario). 74
Figure 3.4. The decision aid (chest pain scenario). 75
Figure 3.5. Additional cues (chest pain scenario). 76
Figure 3.6. “General physical examination” cue information (chest pain scenario). 76
Figure 3.7. Initial diagnosis and certainty scale (chest pain scenario). 77
Figure 3.8. Patient management. 78
Figure 3.9. Flow of activity. 79
Figure 5.1. PRISMA flow diagram. 151

CHAPTER 1: General introduction: biases in decision making and debiasing techniques

Diagnosis is a core task in healthcare. It is particularly important in primary care, where patients presenting with new medical problems need to be managed appropriately, without being indiscriminately subjected to expensive or invasive investigations. Due to their overlapping symptoms, it is often difficult for General Practitioners (GPs) to differentiate between serious and less serious diseases before specialist care, and this could have serious consequences. Failure to provide timely and accurate diagnosis can lead to diagnostic delays, especially in primary care where the volume of patients is high (Singh *et al.*, 2017); for this reason, the World Health Organization (WHO) has set patient safety as its top priority in primary care (World Health Organization, 2016). Given that most people will experience at least one diagnostic error in their lifetime (IoM, 2015), it is particularly important to find ways of supporting medical diagnosis, as it is the task that leads to the greatest potential for error (Bhasale, 1998; Fisseni, Pentzek and Abholz, 2008; Kostopoulou, Delaney and Munro, 2008).

Medical diagnosis could be supported in different ways. At system level, national guidelines, frameworks, checklists as well as performance feedback can be used. At individual level, training and education could make physicians more proficient, metacognitively aware, well-calibrated and less prone to biases. Toward this direction, innovative interventions have been developed in healthcare that focus on both the system and the individual. Such interventions include computerised decision aids that support medical decision making and improve diagnostic accuracy. Very broadly, decision aids are computerised, typically web-based, tools, algorithms or platforms that provide advice to support decision making (Berner *et al.*, 1999; Short, Frischer and Bashford, 2003). In healthcare, this advice could range from risk calculation and procedural guidance to diagnostic recommendations and/or treatment plans (e.g., Isabel[1], DXPlain[2], DynaMed[3], BMJ Best Practice[4], Infera[5],

[1] https://uk.isabelhealthcare.com
[2] http://www.mghlcs.org/projects/dxplain
[3] https://www.dynamed.com/
[4] https://bestpractice.bmj.com/info/
[5] http://www.inferscience.com/about-infera/

Qcancer[6]). The type of decision aid that this thesis is concerned with takes the form of diagnostic suggestions, provided early in the clinical consultation, and was developed to support diagnosis in primary care (Kostopoulou, Rosen, *et al.*, 2015; Kostopoulou, Porat, *et al.*, 2017).

Previous research has highlighted the importance of physicians' initial hypotheses for their subsequent diagnostic decisions. Kostopoulou and colleagues (2017) recruited physicians (GPs) to participate in an online-simulated consultation. GPs were presented with six patient scenarios to diagnose, three of them slightly suggestive of cancer, and were asked to think aloud while assessing the patients. The results showed that physicians who explicitly mentioned cancer as a possible diagnosis at the start (i.e., after a brief patient description) were significantly more likely to finally diagnose cancer and make appropriate referrals in the end. These results suggested that physicians' first hypotheses are strongly related to their subsequent diagnostic and management decisions (Kostopoulou, Sirota, *et al.*, 2017). Based on these findings, the researchers suggested that any intervention to improve medical decisions should be made early, when physicians form their very first hypotheses about the patient, and before they start testing them by gathering further information (Kostopoulou, Sirota, *et al.*, 2017).

Towards this direction, and based on the principle of early intervention, Kostopoulou and colleagues (2015, 2017) developed a prototype computerised decision aid to support diagnoses in primary care. This aid provided diagnostic suggestions to GPs (i.e., a list of alternative diagnoses) at the start of the clinical consultation, as soon as GPs enter the patients' principal symptom. Suggestions were based on the principal symptom and patient demographics and risk factors in the electronic health record. The decision aid was evaluated in a series of studies with computer-simulated patient scenarios, as well as in high fidelity consultations with GPs consulting with actors as patients. The use of the decision aid was found to significantly improve physicians' diagnostic accuracy, without increasing consultation time or number of investigations ordered (Kostopoulou, Lionis, *et al.*, 2015; Kostopoulou, Rosen, *et al.*, 2015; Kostopoulou, Porat, *et al.*, 2017). This series of studies was part of the

[6] https://www.qcancer.org/

TRANSFoRm project, a European research collaboration on the application of Artificial Intelligence (AI) in primary care (Delaney, 2015).

How the improvement in diagnostic accuracy occurs, as well as how the decision aid affects clinical reasoning have not yet been explored. Based on the tenets of dual process theories, there are two distinct types of reasoning: “intuitive” and fast, and “analytical” and slow (Sloman, 1996; Kahneman, 2003, 2011), though other theorists see these on a continuum of thinking (Horstmann, Ahlgrimm and Glöckner, 2009). Since the generation of first hypotheses has been attributed to intuitive thinking (Kahneman and Tversky, 1974; Kahneman, 2011), we can hypothesise that physicians’ first hypotheses about a patient case are formed quickly, as an intuitive response to a presenting problem.

Thomas and colleagues (2008, 2014) linked the generation of initial hypotheses with long-term memory (Thomas *et al.*, 2008; Thomas, Dougherty and Buttaccio, 2014). Specifically, Thomas and colleagues (2008) suggested that, after dealing with a patient case, physicians store related information in long-term memory. When they face a patient with similar symptoms, physicians tend to associate these symptoms with the diagnostic hypotheses they generated in previous, similar patient cases (a form of “availability bias”). Subsequently, physicians retrieve their diagnostic hypotheses for the present patient case from long-term memory and maintain them in working memory (Thomas *et al.*, 2008; Thomas, Dougherty and Buttaccio, 2014). However, during the diagnostic task, physicians can only retrieve a limited number of hypotheses from long-term memory that will be maintained in working memory. As a result, the number of hypotheses that physicians will actually consider are subject to their working memory limitations, and therefore, people with low working memory capacity will consider fewer hypotheses. Thomas and colleagues (2014) suggest that the hypotheses generated first are often the most likely, as these hypotheses occur more frequently in one’s experience and, therefore, are more prevalent in long-term memory (Thomas, Dougherty and Buttaccio, 2014).

Deviating from the working memory hypothesis, there could be a number of psychological mechanisms and cognitive phenomena that could possibly explain how early diagnostic suggestions improve accuracy and, by extension, physicians’

thinking. The underlying hypothesis of this thesis is that the provision of diagnostic suggestions could have a debiasing effect by encouraging physicians to think more analytically and consider more diagnostic alternatives. Below, there is a brief list of cognitive biases that could operate during the diagnostic process and affect physicians' thinking; how these biases could be minimised by the provision of early diagnostic suggestions is also discussed.

1.1. Overconfidence

Daniel Kahneman (2011) described confidence as a feeling that does not necessarily reflect the reality. Specifically, he suggested that *"subjective confidence in a judgement is not a reasoned evaluation of the probability that this judgement is correct. Confidence is a feeling, which reflects the coherence of the information and the cognitive ease of processing it. It is wise to take admissions of uncertainty seriously, but declarations of high confidence mainly tell you that an individual has constructed a coherent story in his mind, not necessarily that the story is true"* (Kahneman, 2011, p. 212). These admissions of uncertainty have been of interest in the medical literature since at least the 1900s, when Sir William Osler characterised medicine as *"a science of uncertainty and an art of probability"* (Bean and Bean, 1950, p. 125). Along the same lines, Jay Katz argued that physicians are generally reluctant to admit uncertainty (Katz, 1984), while Alvan Feinstein, a few years later, highlighted clinicians' unwillingness to consider uncertainty as an option (Feinstein, 1990).

Nevertheless, uncertainty is an inevitable characteristic of medical practice (Katz, 1984; Baumann, Deber and Thompson, 1991). From the early stages of their training, physicians become aware of the uncertainty in their field. However, it seems there is a distance between physicians' theoretical knowledge and clinical practice: physicians are often, consciously or unconsciously, reluctant to acknowledge uncertainty (Katz, 1984; Simpkin and Schwartzstein, 2016). As Simpkin and Schwartzstein (2016) describe, physicians tend to transform a "grey-scale" patient history into a "black and white" diagnosis that can be classified more easily, and therefore, better understood (Simpkin and Schwartzstein, 2016). Physicians are also

reluctant to disclose mistakes, as they think that acknowledging uncertainty is a sign of weakness or ignorance (Brezis *et al.*, 2019). In addition, this intolerance of uncertainty often leads to excessive ordering of tests and investigations (Kassirer, 1989). As a response to uncertainty and to patients' demand for the "one and only" right answer, physicians may try to appear confident and hold authority by expressing unwarranted certainty in their ability to make accurate diagnoses (Katz, 1984; Baumann, Deber and Thompson, 1991). This overconfidence, in turn, can trigger other cognitive biases, such as selective information seeking, information distortion and premature closure, and can therefore, increase the risk of diagnostic error. For these reasons, as Simpkin and Schwartzstein (2016) suggested, it is important to acknowledge uncertainty in medicine and change the existing culture.

Overconfidence, indeed, has been linked to diagnostic error in medicine (Friedman *et al.*, 2005; Berner and Graber, 2008; Croskerry and Norman, 2008; Meyer *et al.*, 2013). As Berner and Graber (2008) describe, it is possible that physicians develop an attitude of knowing all they need to know about a patient case, which makes them reluctant to seek advice or consider other possibilities. Along with the attitudinal aspect of overconfidence, physicians often develop a sense, or an "illusion", of validity which makes them overestimate the accuracy of their judgements. This mismatch between perceived and actual accuracy may prevent physicians from (metacognitively) evaluating their own performance and make them prone to diagnostic error (Berner and Graber, 2008; Kahneman, 2011). Several studies provide evidence for the existence of overconfidence in clinical reasoning (Baumann, Deber and Thompson, 1991; Berner and Maisiak, 1999; Friedman *et al.*, 2005; Meyer *et al.*, 2013; Cavalcanti and Sibbald, 2014; Brezis *et al.*, 2019). However, overconfidence is not limited to medical decision making. It has been described as a general feature of human behaviour (Kahneman and Tversky, 1974; Croskerry and Norman, 2008; Kahneman, 2011; Thompson, Prowse Turner and Pennycook, 2011), and has been observed in legal and forensic decisions (Sieck and Arkes, 2005), forecasting (Arkes, 2001), gambling (Goodie, 2005), as well as in financial decision making (Nosic and Weber, 2010).

Moore and colleagues (2008, 2017) identified three different forms of overconfidence in the literature: over-estimation, over-placement, and over-precision (Moore and

Healy, 2008; Moore and Schatz, 2017). Over-estimation occurs when individuals perceive themselves to perform better than they actually do. Over-placement occurs when individuals (wrongfully) perceive their performance to be better than others'. Over-precision occurs when individuals are overly confident about the accuracy of their judgements. Furthermore, Moore and colleagues (2008, 2017) suggested that overconfidence changes as a function of task difficulty: on difficult tasks, people tend to overestimate their performance and under-place themselves relative to others. On easy tasks, people tend to under-estimate their performance and over-place themselves relative to others.

Apart from the three different aspects of overconfidence identified by Moore and colleagues, overconfidence has also been described as a metacognitive phenomenon (Berner and Graber, 2008; Thompson, Prowse Turner and Pennycook, 2011; Thompson, Prowse Turner, *et al.*, 2013; Roy *et al.*, 2014; Molenberghs *et al.*, 2016). When people are not certain about a decision, they usually spend more time re-evaluating their conclusions, considering alternatives, and seeking advice. On the other hand, people who are very certain about their decisions or actions, they often fail to reflect on their own reasoning and re-evaluate their decisions by considering other possibilities. Thompson and colleagues (2011, 2013) linked this feeling of certainty (i.e., confidence) that people have in their decisions with human intuition. In line with Kahneman (2003, 2011) and according to the assumptions of dual process theories, they suggested that people's first intuitive response to a problem is accompanied by a feeling that they are right. As they suggest, this "feeling of rightness" determines whether, and to what extent, analytical thinking processes will be triggered (Thompson, Prowse Turner and Pennycook, 2011; Thompson, Prowse Turner, *et al.*, 2013).

To test this hypothesis, Thompson and colleagues (2011, 2013) conducted a series of studies with university students and staff as participants. In these studies, they presented participants with a reasoning task (i.e., to evaluate the validity of conditional statements in the form of *if p then q*). Their task was to indicate whether the conclusion from the conditional statements followed logically or not. Participants also indicated their feeling or rightness (i.e., how confident they were) in their responses on a scale from 1 (guessing) to 7 (certain I am right). Participants were

divided into two groups. In one group, participants were asked to give a fast, intuitive response as soon as they read the conditional statements. In the other group, in addition to their first response, participants were also asked to give a second, final response without any time constraints. The results showed that responses that were given “fluently” (i.e., intuitively, quickly and easily, measured by participants’ response time) produced a higher feeling of rightness compared to responses that were given “disfluently” (i.e., slowly and analytically). These fluent responses were less likely to change in the second-response group (Thompson, Prowse Turner and Pennycook, 2011). Likewise, participants with lower feeling of rightness appeared more likely to engage in analytical thinking: they spent more time re-evaluating their responses and changed them more frequently (Thompson, Prowse Turner and Pennycook, 2011; Thompson, Prowse Turner, *et al.*, 2013).

Overall, the results from Thompson and colleagues’ studies (2011, 2013) suggest that feelings of rightness (i.e., confidence in an initial, intuitive response) can predict the time that people take to reach a conclusion, as well as the likelihood of them changing their mind. However, it is important to note that these results come from a series of studies where the decision task was completed by students and lay participants. Furthermore, the task itself was relatively simple, and inaccurate judgements had no consequences. In healthcare though, decisions are made by professionals, problems are often complex, and inaccurate judgements can have serious consequences.

With regards to clinical reasoning, Baumann and colleagues (1991) investigated physicians’ confidence in their treatment choices for patients diagnosed with breast cancer (Baumann, Deber and Thompson, 1991). In that study, the authors made a distinction between micro-(un)certainty and macro-(un)certainty: the former refers to physicians’ levels of confidence in their treatment decisions, while the latter refers to the variability of treatment decisions across physicians. Study participants were physicians from various specialties (e.g., haematologists, oncologists, surgeons, GPs). Physicians were presented with 10 breast cancer patient scenarios and a list of treatment options to choose from (e.g., removal of some of the breast, radiotherapy as primary treatment, chemotherapy). They were also asked to indicate how certain (i.e., confident) they were in their treatment choices on a scale from 1

("no uncertainty") to 5 ("a great deal of uncertainty"). The results showed that there was large variation in treatment choices across physicians, yet physicians were very certain of their treatment choice (Baumann, Deber and Thompson, 1991). These findings suggested that in the absence of a consensus about the optimal treatment in each case, physicians' high confidence in their decisions would appear unjustifiable.

Friedman and colleagues (2005) provided further evidence for the effects of overconfidence on clinical judgement. In their study, Friedman and colleagues explored whether physicians were well-calibrated by measuring the relationship between physicians' confidence in their diagnoses (i.e., perceived diagnostic accuracy) and the correctness of their diagnoses (i.e., actual diagnostic accuracy). In a within-participant design, they recruited 216 participants (72 medical students, 72 medical residents and 72 faculty internists) to assess nine patient scenarios. In the control condition, after being presented with the patient scenarios, participants were asked to provide up to six diagnostic hypotheses and indicate how confident they were in their diagnoses. Confidence was measured in a behavioural, rather than an explicit way: participants were asked to indicate the likelihood of seeking assistance in reaching a diagnosis, on a scale from 1 ("unlikely", indicating high confidence) to 4 ("likely", indicating low confidence). In the experimental condition, participants were provided with decision support in the form of diagnostic advice. The results showed that participants were well-calibrated (i.e., appropriately confident) in most of the cases (68%). However, when confidence and accuracy were not aligned, residents, internists and students were found to be overconfident in 41%, 36% and 25% of the cases respectively (Friedman *et al.*, 2005). The results also showed that decision support improved participants' diagnostic accuracy.

In a more recent study, Meyer and colleagues (2013) explored the relationship between physicians' diagnostic accuracy and confidence in judgement (Meyer *et al.*, 2013). They recruited 118 internal medicine physicians to assess four hypothetical patient scenarios in a within-participant design. The scenarios were varied in difficulty (i.e., two easier and two more difficult). Scenario difficulty was determined on the basis of average ratings, ranging from 1 to 7, given by experienced clinicians. All physicians assessed the four patient scenarios in a random order. Each scenario was divided into four sequential stages where additional patient information was

provided as the diagnostic process progressed. Specifically, at stage 1, physicians were presented with a short patient description including medical history and presenting complaint. They were then asked to provide up to three diagnoses (in order of likelihood) and indicate their confidence in each diagnosis on a scale from 0 (low confidence) to 10 (high confidence). At stage 2, physicians were presented with physical examination results and were asked to provide again up to three diagnoses, as well as their confidence ratings. At stage 3, physicians were presented with general laboratory and imaging results. At this stage, along with their diagnoses and confidence ratings, physicians were asked to indicate what additional information they would request in order to reach a diagnostic conclusion. The available options included advice from a colleague, additional laboratory results and use of electronic references or books. At the last stage (stage 4), physicians were presented with definitive or specialised laboratory and imaging results. At this stage, they were asked to provide only one final diagnosis and a confidence rating.

The results from Meyer and colleagues (2013) study showed that diagnostic accuracy was significantly lower in more difficult scenarios (5.80% versus 55.30%). The same was true for physicians' confidence. However, the difference in confidence between easier and more difficult scenarios was minor (7.2 versus 6.4 respectively), suggesting that physicians were not well-calibrated in the more difficult scenarios (they appeared to be overconfident). In addition, more confident and more experienced physicians were less likely to request further patient information or advice (at stage 3), suggesting that physicians were not aware of the (in)accuracy of their own diagnostic judgements and the need to seek assistance in order to reach a conclusion (Meyer *et al.*, 2013).

The empirical evidence of overconfidence in clinical reasoning may be limited. However, the studies described above provide some indication of the robustness of overconfidence in medical decision making and its consequences in terms of patient safety. It would be therefore important to find ways of reducing this effect.

1.2. Confirmation bias

Confirmation bias is one of the most robust phenomena in the literature of judgement and decision making. Broadly, it refers to the tendency to seek or interpret information in a way that fits prior beliefs (Nickerson, 1998; Kahneman, 2011). Similar to overconfidence, confirmation bias is not limited to medical decision making, but it also operates in financial decisions (Gail, 2000; Huang, Hsu and Ku, 2012), legal and forensic contexts (Kassin, Dror and Kukucka, 2013; Lidén, Gräns and Juslin, 2019), as well as in forecasting (Cipriano and Gruca, 2014) and gambling (Gilovich, 1983).

1.2.1. Selective information search

The first instantiation of confirmation bias according to Nickerson's (1998) definition refers to the selective search of information. This phenomenon has been explored by Peter Wason in a seminal programme of work during 1960s and early 1970s. In those studies, Wason tested participants' information search strategy about validating or falsifying a given rule, by seeking for either confirming or disconfirming information (Wason, 1960, 1968; Wason and Shapiro, 1971). In his first study, Wason recruited 20 undergraduate students and presented them with a sequence of numbers (i.e., 2, 4, 6) that followed a certain rule (i.e., any three numbers in increasing order). Participants' task was to find the rule, after having generated different sequences of numbers and justified their response (i.e., explain why they had chosen those three numbers). For each sequence that participants generated, they were provided with feedback about the correctness of their response (i.e., whether the sequence generated conformed to the rule or not; if not, they continued generating sequences – with feedback – until the rule was discovered). Over many rounds of hypothesis-generation and -testing, Wason observed that participants generally formed a hypothesis about a given rule and tended to generate sequences that followed the rule they had in mind (Wason, 1960). This finding suggests that participants sought information that would confirm (rather than test) their hypothesised rule (i.e., to find the certain rule that the given sequence of numbers conformed to).

Further evidence for this hypothesis-confirming strategy was provided in subsequent studies by Wason. In a series of experiments on deductive reasoning, undergraduate students were given the (well-known) Wason selection task (Wason, 1968; Wason and Shapiro, 1971). The task included a) a set of four different cards, each of them having a letter on one side and a number on the other side (e.g., D, 3, B, 7), and b) a conditional rule associated with the cards (e.g., if a card has a "3" on its front side, then there is a "D" on the back side)[7]. Participants were asked to select which cards they needed to turn over in order to conclude whether the rule was true or false. In line with the 1960 study, these studies showed that on average, participants tended to select cards that would validate the rule (e.g., "3" and "D"), rather than cards that could falsify the rule (e.g., "7"). As before, these findings suggest that people tend to focus selectively on information that confirms their hypotheses, disregarding evidence that contradicts their initial choice.

Analogies of selective information search have been observed in clinical reasoning as well. There is evidence that physicians tend to anchor on their initial hypotheses, as they believe they are correct, and they are often reluctant to seek advice (Dreiseitl and Binder, 2005; Arkes, 2013). When they do, they often seek information that only supports their prior hypotheses – i.e., selective search (Mendel *et al.*, 2011; Dani, Bowen-Carpenter and McGown, 2019), and/or misinterpret this information to support their initial diagnostic hypotheses – i.e., information distortion (Leblanc, Norman and Brooks, 2001; Leblanc, Brooks and Norman, 2002; Kostopoulou, Mousoulis and Delaney, 2009; Kostopoulou *et al.*, 2012, also see 1.2.2. below).

Mendel and colleagues (2011) conducted a study on confirmation bias in medical decision making, using psychiatrists and medical students as participants (Mendel *et al.*, 2011). In this study, participants were presented with a patient case and asked to choose between two possible diagnoses (dementia or depression). They were then presented with twelve items of additional information, six of them supporting dementia and another six that supported depression. The items were introduced in short titles, summarising their main idea. Participants were asked to indicate which

[7] Throughout the experiments, there were variations of the selection task including cards with shapes and colours, cities and transportation means or vowels and even (as opposed to odd) numbers (see Wason, 1968 and Wason and Shapiro, 1971).

items of information they would like to receive more information on, and they were free to obtain as much information as they wished. After seeing all the information that they requested, they provided their final diagnosis. The results showed that 13% of psychiatrists and 25% of medical students conducted a confirmatory information search, suggesting that they selectively searched for information that fitted their prior diagnostic hypotheses (e.g., if their chosen diagnosis was depression, then they searched more items of information that suggested depression as the most likely diagnosis for the patient). Those participants were also less likely to choose the correct diagnosis: psychiatrists who requested confirmatory information made an incorrect diagnosis in 70% of the cases, compared to 27% and 47% of those who requested disconfirming or balanced information respectively. This study not only identifies confirmation bias in clinical reasoning, but also suggests that selective information search can lead to diagnostic error (Mendel *et al.*, 2011).

1.2.2. Biased information evaluation

The second manifestation of confirmation bias according to Nickerson's (1998) definition refers to the biased interpretation of information in favour of an existing or prior belief (i.e., information distortion). This biased evaluation of information has traditionally been thought to occur post-decisionally; that is after a decision is made (Festinger, 1957). In Cognitive Dissonance Theory, Leon Festinger (1957) proposed that people generally feel a desire to maintain internal cognitive consistency between their opinions/beliefs and their actions/attitudes (Festinger, 1957). To maintain consistency and avoid conflict, they tend to seek information that confirms their beliefs. However, when people are exposed to information that contradicts their beliefs, this information produces cognitive dissonance, a form of "conflict" between the negative elements of the chosen action and the positive elements of the alternative actions (Festinger, 1957). As people wonder whether they have made the best decision/choice, this internal conflict causes psychological discomfort. To reduce psychological discomfort, as Festinger (1957) suggested, people tend to either invalidate contradictory information or overweight confirmatory information. To this extent, information distortion instantiates a form of confirmation bias.

Some years later, researchers began to investigate whether information distortion also occurs pre-decisionally; that is, during the course of decision making, and before a final decision is made. Russo and colleagues (1996) suggested that distortion of information can be attributed to a pre-established preference, and not necessarily to post-decisional cognitive dissonance, as Festinger (1957) suggested.

In a series of experiments, Russo and colleagues (1996) tested whether pre-established preferences or prior beliefs might bias decision makers' evaluation of incoming information in a way that favours their preferred option (Russo, Medvec and Meloy, 1996). In these experiments, Russo and colleagues (1996) presented participants (students) with a decision task: to choose between two restaurants/hotels. To induce an initial leaning (i.e., a pre-established preference), the researchers linked one of the options to a positive event in the participants' history (e.g., "the owner of hotel A previously accommodated your parents when all the local hotels were fully booked"). Participants were then given five, relatively neutral, items of information, which favoured neither choice (e.g., "hotel A offers a room for $75 and breakfast for an additional $5, whereas hotel B offers a room for $80 including breakfast"). For each item of information, participants were asked to rate how much it favoured one choice or the other, on a scale from 1 ("Strongly favours A") to 7 ("Strongly favours B"). After each item of information, participants were asked whether they would like to make their final choice or wanted to obtain more information. When participants indicated their final choice, they indicated how confident they were, on a scale from 50 ("very uncertain") to 100 ("very certain"). If participants made their choice without requesting all the items, they were shown the remaining items and asked to evaluate them (the researchers therefore collected both pre-decisional and post-decisional evaluations). A separate control group completed the same task, but no pre-established preference was installed. Distortion was calculated as the difference in item ratings between the control group (no pre-established preference) and the experimental group (pre-established preference).

The results showed that participants with a pre-established preference (the experimental group) distorted information in favour of their preferred choice, as distortion ratings were higher compared to the control group (Russo, Medvec and Meloy, 1996). They also found that participants without a pre-established preference

(the control group) developed a preference “spontaneously”, (while evaluating the items); subsequently, they too began to distort information in its favour. In the experimental group, pre-decisional information distortion was significantly greater than post-decisional information distortion, which suggests that prior beliefs had a stronger effect than post-decisional dissonance.

In later work, Russo and colleagues (1996, 2008) endeavoured to explain why pre-decisional information distortion occurs. In a series of experiments, they investigated three potential causes (or driving forces): a) minimisation of effort, b) separation between alternative choices and c) information consistency (Russo, Medvec and Meloy, 1996; Russo *et al.*, 2008). Minimisation of effort refers to the tendency to reduce cognitive load. It has been hypothesised to underpin many cognitive phenomena (heuristics and biases), including confirmation bias (Russo, Medvec and Meloy, 1996; Mendel *et al.*, 2011; Huang, Hsu and Ku, 2012) and premature closure (Webster and Kruglanski, 1994; Choi *et al.*, 2008; Roets *et al.*, 2014). Separation refers to the distinction between alternative choices in order to make a clearer choice. This separation between alternatives increases when information is distorted to support a preferred choice (Russo, Medvec and Meloy, 1996). Finally, maintaining information consistency refers to the desire to make coherent judgements and avoid information that conflicts with your beliefs or choices (Holyoak and Simon, 1999; Russo *et al.*, 2008). It is similar to Festinger’s cognitive consistency notion described above, with the difference that Russo and colleagues’ consistency is thought to occur pre-decisionally. In binary-choice tasks, similar to the ones used in 1996 study, Russo and colleagues (2008) used different methodologies including (subliminal) priming and (conscious) retrospective assessment in order to activate the driving forces described above. Both methods showed that activation of consistency led to significantly greater distortion of information compared to other potential causes (e.g., separation or minimisation of effort). The findings from their series of experiments suggest that the main driver of distortion was people’s desire for consistency (Russo *et al.*, 2008).

Analogies of biased information evaluation to support an initial choice have been found in clinical reasoning as well. LeBlanc and colleagues (2001, 2002) investigated the effects of initial hypotheses and diagnostic suggestions on the evaluation of

clinical information. In their studies, they recruited medical students and residents to assess eight photographs of patients (portraits). The photographs were taken from medical textbooks as representatives of specific diseases but did not rule out other possible diseases. For each photograph, participants were presented with a brief patient history that was suggestive of a specific disease (diagnosis). Participants were also provided with a plausible differential diagnosis (only in the 2001 study). Their task was to assess the photographs for any important features (i.e., physical characteristics) that could help to confirm or rule out any diagnoses. They were also asked to provide ratings of likelihood for the suggested diagnosis, the differential diagnosis (2001 study only), as well as any other diagnoses that they themselves deemed plausible. When participants were steered towards a differential diagnosis ("diagnosis B"), they ended up interpreting textbook features of diagnosis A as symptoms of diagnosis B. They also gave higher likelihood ratings to diagnosis B than diagnosis A (2001 study), even though the photograph was a textbook example of diagnosis A. In short, both studies showed that the interpretation of features presented in the photographs was consistent with the suggested diagnosis (Leblanc, Norman and Brooks, 2001; Leblanc, Brooks and Norman, 2002). That is, when participants were initially primed to think of a specific diagnosis (e.g., liver disease), their subsequent interpretations of patient features were towards the prime and not towards another diagnosis (e.g., tanned skin interpreted as a sign of liver disease rather than a stomach disease, or no disease at all).

In a similar vein, Kostopoulou and colleagues (2009) conducted a study on information search and information evaluation for diagnosis in primary care. In a between-participants design, 84 General Practitioners (GPs) assessed a patient scenario with ambiguous features. The scenario started with a brief patient description with chest pain as the presenting problem. The patient description contained a number of clinical cues, some of them suggesting a cardiac diagnosis (i.e., 60 years old male, ex-smoker with family history of ischemic heart disease), and others suggesting that a musculoskeletal cause was more likely (i.e., onset of pain while lifting a washing machine). Subsequently, physicians were able to request further information about the patient from a list of additional clinical cues. As in the initial patient description, some of the additional clinical cues favoured a cardiac diagnosis (e.g., pain relieved at rest) while some other cues favoured a

musculoskeletal cause (e.g., tenderness over the intercostal muscles). There was also an ambiguous cue that could fit with both cardiac and musculoskeletal pain (i.e., pain comes on with gardening).

Therefore, patient information was relatively balanced for the two competing diagnoses (i.e., cardiac versus musculoskeletal), and both were plausible. Physicians were able to request as many clinical cues as they wished before providing their final diagnosis (in free text), their level of confidence (0-100 scale), as well as their management decision. After completing the task, 52 of the physicians agreed to be interviewed about their reasoning during the diagnostic task (the authors used a 'stimulated recall' method of interviewing, in which participants viewed a recording of their task performance and answered questions regarding their thought processes at specific times). The stimulated recall interviews revealed major differences between physicians in the interpretation of the clinical cues: physicians who diagnosed a cardiac cause perceived the cues to indicate cardiac disease whereas physicians who diagnosed a musculoskeletal cause interpreted the cues as more supportive of a musculoskeletal diagnosis (Kostopoulou, Mousoulis and Delaney, 2009).

Furthermore, physicians who diagnosed a cardiac cause (a more severe diagnosis than musculoskeletal) requested more information about the patient (cues) and spent more time on the scenario. In addition, they requested more cues related to a cardiac diagnosis (e.g., a resting electrocardiogram or family history of heart disease). Likewise, physicians who diagnosed a musculoskeletal cause requested more cues related to their (musculoskeletal) diagnosis (e.g., palpation of chest). They were also more certain in their diagnosis compared to physicians who diagnosed a more severe (cardiac) cause (71.09 versus 61.37, $p = 0.89$). The findings of the study also suggest that physicians can be biased in the way they search clinical information, as well as in the way they interpret information in order to make this information fit with their diagnostic decisions (Kostopoulou, Mousoulis and Delaney, 2009). As the authors suggest, this could be due to the need for coherent judgements, as described in Russo and colleagues' studies.

In an attempt to seek more direct evidence for pre-decisional distortion of information in physicians' diagnostic judgements, Kostopoulou and colleagues (2012) conducted a subsequent study using the stepwise evolution of preference method (SEP; Russo, Medvec and Meloy, 1996; Meloy and Russo, 2004; Russo *et al.*, 2008). The SEP method has been widely used in the information distortion literature (Meloy and Russo, 2004; Kostopoulou *et al.*, 2012; Nurek, Kostopoulou and Hagmayer, 2014; Nurek and Kostopoulou, 2016). It was inspired by Russo and colleagues' work on dual-choice decision tasks, as described above (Russo, Medvec and Meloy, 1996), and took its final form in their subsequent studies (2004, 2008). Briefly, in a given dual-choice task, decision makers are presented with some initial information about the two alternatives. They are then presented with subsequent information about the alternatives ("cues") which are presented sequentially. For each cue, decision makers are asked to a) evaluate the cue information (i.e., rate the extent to which it favours one alternative or the other), b) indicate their current choice (i.e., their "leading" alternative) and c) indicate their confidence in that choice (i.e., the strength of their current preference).

Kostopoulou and colleagues (2012) used a between-participants design and recruited 102 GPs to assess three patient scenarios. As in Kostopoulou and colleagues (2009), each scenario started with a brief patient description. However, the patient description did not contain ambiguous or conflicting information (as was the case in the 2009 study). Instead, it contained information that favoured one of two diagnostic alternatives, e.g., Diabetes versus Depression. Participants were asked to indicate which diagnosis is more likely, on a 21-point likelihood scale anchored at 10 on either end (diagnosis A is more likely - diagnosis B is more likely), with a 0 midpoint (i.e., equally likely).

Thereafter, a sequence of neutral (non-diagnostic) items of clinical information was presented. Physicians were asked to rate how much each item favoured either diagnosis A or B on a similar 21-point scale. They were also asked to update the strength of their diagnostic belief (i.e., the likelihood rating). In the third scenario only, and only after the neutral cues, physicians were presented with a set of diagnostic clinical cues that favoured the alternative diagnosis (i.e., the diagnosis that initially appeared as less likely, based on the patient description). This last piece of

information made the overall scenario information balanced, suggesting that both diagnoses were equally likely. By providing physicians with conflicting information (i.e., cues that favoured diagnosis A, neutral cues, and cues that favoured diagnosis B), the researchers tested whether biased evaluation of information contributes to physicians' resistance to change mind in favour of their initial belief.

There was also a separate control group of 36 physicians who provided "baseline" ratings of all the cues, in a random order, without having the opportunity to provide likelihood estimations about possible diagnoses. The cues were scrambled to prevent participants from developing a leading diagnosis (and distorting subsequent information in its favour). In the absence of a leading diagnosis, these ratings were taken to indicate each cue's "true" (or "undistorted") diagnostic value. Distortion was measured as the difference between the ratings provided by the control and experimental groups. The results showed that physicians with higher confidence in their leading diagnosis were more likely to distort subsequent information to support it, and less likely to change diagnosis at the end, after rating the conflicting cues (Kostopoulou *et al.*, 2012).

1.3. Premature closure

Along with overconfidence, premature closure has been identified as one of the most important contributing factors to diagnostic error in medicine (Eva *et al.*, 2010; Jaimes *et al.*, 2013; Phua and Tan, 2013; Van Den Berge and Mamede, 2013; Spaanjaars *et al.*, 2015; Voytovich, Rippey and Suffredini, 1985; Kumar, Kanna and Kumar, 2011; Krupat *et al.*, 2017). Premature closure occurs when decision makers reach a conclusion too early, based on inconclusive information (Eva *et al.*, 2010). It is related, but not limited, to what Kruglanski and colleagues termed "need for cognitive closure", the desire for an effortless but confident final answer or definite knowledge on a given topic, as opposed to uncertainty (Kruglanski, Webster and Klem, 1993; Webster and Kruglanski, 1994; Kruglanski and Webster, 1996). Since premature closure is based on limited information, the authors suggested that people with high need for cognitive closure would generate fewer hypotheses that require less information processing. As a result, confidence in an initial hypothesis would

increase, as this hypothesis appears to be the most likely, given the “lack” of alternatives (Kruglanski and Webster, 1996).

Kruglanski and colleagues (1993, 1996) suggested that people tend not only to reach conclusions prematurely based on limited information, but also to maintain these conclusions for as long as possible (also see Chapter 5: Summary). In their 1993 study, Kruglanski and colleagues investigated this phenomenon in a series of experiments with undergraduate students as participants. In brief, participants were presented with a legal case and a legal analysis, which argued either for the plaintiff or the defendant. Their task was to provide their individual verdict, along with their level of confidence in their decision. Participants subsequently formed separate groups of two people and were asked to provide a verdict by consensus. Each group consisted of one participant with high and one participant with low need for cognitive closure, as measured by the relevant scale developed by the researchers (Kruglanski, Webster and Klem, 1993). Without being aware, each participant within each group had read a different legal analysis: one participant read an analysis arguing for the plaintiff while the other participant read an analysis arguing for the defendant, allocated at random. This manipulation steered participants toward different verdicts and, thus, the researchers were able to test whose verdict would finally predominate. The results showed that participants with higher need for closure, were more confident in their verdicts and more likely to persuade their partners. This finding suggests that participants who reached a conclusion prematurely and were confident therein were less willing to change their mind, or in the authors’ words, less open to persuasion (Kruglanski, Webster and Klem, 1993).

Apart from resistance to change, research on consumer behaviour has shown that need for cognitive closure can also affect people’s information search strategy, as well as the amount of information that they are willing to process. Choi and colleagues (2008) conducted two computer-based experiments with undergraduate students as participants (77 and 54 respectively). Participants were presented with a 12 x 7 table, containing 12 alternative choices for a technology product category (i.e., laptops), and seven different attributes/features of each alternative, including price, memory capacity and weight. The attribute information for each alternative was initially hidden; to view it, participants had to select it in order to reveal it. Participants

were able to reveal as many attributes as they wanted, until they finally indicated their preferred alternative.

The results showed that participants' search strategy, as well as the amount of information they assessed before they indicated their preferred choice, was associated with their need for cognitive closure, as measured by the relevant scale developed by Webster and Kruglanski (1994). Specifically, participants with higher need for cognitive closure tended to conduct an attribute-based search (examining a particular attribute across alternatives) rather than an alternative-based search (examining all of the belonging to a particular alternative), suggesting a preference for a search strategy that requires less information processing. The results also showed that participants with higher need for cognitive closure revealed fewer attributes across the 12 alternatives (i.e., assessed a smaller amount of information) than those with lower need for cognitive closure (Choi *et al.*, 2008). These findings were replicated in a second experiment by Choi and colleagues (2008), with the difference that the task included time constraints.

The effects of cognitive closure have been shown in clinical reasoning as well. Voytovich and colleagues (1985) conducted an early study on premature closure in medical diagnosis. They recruited 58 participants (physicians, residents, and medical students) to assess three brief patient scenarios. Participants were asked to generate a list of possible diagnoses for each patient scenario. They were also advised to avoid any diagnostic conclusions that cannot be justified by the patient information. None of the patient scenarios contained sufficient information for a definite diagnosis. The results showed that the large majority of participants (i.e., 91.40%) reached a diagnostic conclusion prematurely at least in one scenario, as the authors reported. Across the three patient scenarios, the frequency of premature conclusions was approximately 63% (i.e., 110 out of 174 cases), irrespective of participants' level of experience (Voytovich, Rippey and Suffredini, 1985).

Further evidence of premature closure and its association with diagnostic error was provided by Krupat and colleagues (2017). In an online study, the researchers recruited 75 participants (medical students, residents and faculty physicians) to assess four hypothetical clinical scenarios. Patient information was presented

sequentially: initially there was a brief description of the patient, based on which participants were asked to provide a list of possible diagnoses. They were also asked to provide confidence ratings in each diagnosis on a scale from 1 to 100. Subsequently, participants were asked to either "announce" their diagnostic conclusion or request further information about the patient. There were 10 different pieces of clinical information that participants were able to assess. The patient scenarios were designed in a way that the first four pieces of information created an initial leaning toward a certain diagnosis, while the rest of the information was somewhat inconsistent with the initial leaning. After each piece of information, participants were asked to update their diagnoses and confidence in them. Participants could stop searching for additional patient information at any point, by stating their final diagnosis. However, if they decided to stop before the fifth piece of information, they were then asked to assess a last piece of information (i.e., the disconfirming one compared to the previous four), before providing their final diagnosis. Finally, they were able to either obtain further information or provide their diagnostic conclusion.

In line with Voytovich and colleagues (1985), Krupat and colleagues' (2017) study found that level of experience did not predict participants' accuracy in their diagnoses. Instead, it was the average number of diagnoses generated, as well as the average number of pieces of information requested, that predicted accuracy. Specifically, participants who on average listed a greater number of differential diagnoses were more likely to be accurate in their final diagnosis. Likewise, participants who obtained a larger amount of information, as well as those who sought further patient information after having been presented with the fifth piece (i.e., disconfirming information) were more likely to provide a correct final diagnosis. Finally, participants with lower initial confidence were more likely to change their diagnosis in the presence of additional disconfirming information (Krupat *et al.*, 2017). Overall, these findings suggest that considering fewer alternatives and assessing less information can lead to premature and erroneous decisions. They further suggest that confidence in an initial decision can predict commitment to that decision (i.e., resistance to change in the face of conflicting information).

1.4. Debiasing strategies

The psychological mechanisms described above are closely related to what Kahneman and Tversky (1974) called “anchoring”, the overreliance on an initial piece of information, an initial judgement, or a first impression (Kahneman and Tversky, 1974; Rabin and Schrag, 1999). Initial hypotheses are thought to be generated intuitively and accompanied by a feeling of rightness (Kahneman, 2011; Thompson, Prowse Turner and Pennycook, 2011). These hypotheses often serve as anchors that drive people’s search of information and information evaluation, and they can determine whether, and to what extent an initial judgement will be updated in the light of new information (Hogarth and Einhorn, 1992; Ames *et al.*, 2010; Spaanjaars *et al.*, 2015; Kostopoulou, Sirota, *et al.*, 2017) (also see Chapter 5: Introduction). This sequence of events could be broken by several de-biasing interventions that mainly focus on reducing this anchoring effect and the confidence that accompanies it. Interventions that have the potential to minimise the effects of cognitive biases and induce analytical reasoning are, therefore, of great interest. This is particularly important in medical decision making, where such interventions can make physicians more cautious and better calibrated, in order to reduce diagnostic error.

Larrick (2004) distinguished between motivational, cognitive, and technological debiasing strategies. Motivational strategies include incentives and accountability to improve decision making. Cognitive strategies aim to improve decisions through training in logical rules, statistical reasoning and awareness of one’s own biases (also see Gigerenzer, 2015). A well-known example is ‘consider-the-opposite’: a cognitive technique that directs attention towards disconfirming information and facilitates consideration of alternative hypotheses. Finally, technological strategies aim to improve decision making through the use of tools such as decision analysis and computerised decision aids (Larrick, 2004). In a recent systematic review, Ludolph and Schulz (2018) examined the effectiveness of these debiasing strategies in health-related judgements. They found that the majority of interventions did have a successful debiasing effect (Ludolph and Schulz, 2018). Technological interventions (including decision aids) as well as cognitive interventions (including educational

training and consider-the-opposite rules) were found to be amongst the most effective debiasing strategies.

Consider-the-opposite has been found to be particularly effective on reducing overconfidence (Koriat, Lichtenstein and Fischhoff, 1980; Hirt and Markman, 1995; McKenzie, 1997; Haran, Moore and Morewedge, 2010), anchoring (Mussweiler, Strack and Pfeif, 2000), confirmation bias (Brussel van *et al.*, 2020), and hindsight bias (Arkes *et al.*, 1988). For instance, Mussweiler and colleagues (2000) examined the effects of consider-the-opposite on anchoring. In one of their experiments, they recruited car experts and asked them to estimate the price of two cars, based on certain features such as milage and production year. In brief, participants were provided with either a low or a high anchor (i.e., $1,500 or $2,800), and were asked whether this price was too high or too low. Subsequently, half of the participants were provided with feedback (e.g., a friend's opinion about the anchor as too high/low) and asked to argue against the feedback (e.g., to list reasons that contradict that opinion). The results showed that participants who provided arguments that contradicted the anchor value (i.e., consider-the-opposite) were less influenced by the anchors compared to those who did not provide counterarguments (Mussweiler, Strack and Pfeif, 2000).

Similar effects to the consider-the-opposite technique on hindsight bias were demonstrated by Arkes and colleagues (1988). In their study, Arkes and colleagues recruited psychologists to assess a hypothetical clinical scenario. Participants were divided into two groups: "foresight" and "hindsight". They were presented with the patient case and were asked to estimate the probabilities of three possible diagnoses (i.e., alcohol withdrawal, Alzheimer disease or brain damage due to alcohol abuse). Patient information was different across groups, so that participants in the hindsight group read an additional piece of information about the patient that primed them to think of, or anchored them on, a certain diagnosis (e.g., "The primary diagnosis for the patient is alcohol withdrawal/Alzheimer disease/brain damage"). This information was not revealed to the foresight group. In both groups, half of the participants were asked to indicate one piece of information from the patient scenario that would support each of the three diagnoses as the most likely (i.e., consider reasons for the opposite). The results showed that the provision of reasons for the

alternative diagnoses (i.e., the diagnoses that were not included in the prime) did not completely eliminate hindsight bias (Arkes *et al.*, 1988). However, the proportion of participants in the hindsight group who provided higher estimates than the foresight groups was smaller compared to the proportion of participants in the hindsight group who considered information that supported the alternative diagnoses (58% versus 41%). This suggests that considering evidence that does not support the prime, or the anchor (the suggested diagnosis in this case), significantly reduced the magnitude of the bias.

As mentioned above, the debiasing effect of the consider-the-opposite technique has been demonstrated in studies on overconfidence. An example is Koriat and colleagues' (1980) study where participants (students) were presented with two sets of 10 general knowledge questions. The first set of questions (control) required participants to choose between two alternatives for the correct answer and provide their confidence in each answer. In the second set of questions (experimental), apart from their answers and confidence ratings, participants were asked to provide reasons for and against each of the two alternatives (i.e., to consider-the-opposite). The results showed that participants in the experimental condition were better calibrated, compared to the control condition, where participants were found to be overconfident. This suggests that considering reasons that contradict the chosen option and support the alternative one can reduce overconfidence, thus improving accuracy (Koriat, Lichtenstein and Fischhoff, 1980; Hirt and Markman, 1995; McKenzie, 1997; Haran, Moore and Morewedge, 2010).

1.4.1. Decision aids

Since overconfidence has been linked to diagnostic error in medicine, reducing overconfidence has been proposed as a debiasing strategy to improve accuracy in clinical judgements, especially in mental health (Arkes, 1981; Borum, Otto and Golding, 1993; Smith and Dumont, 1997; Miller, Spengler and Spengler, 2015). As discussed above, consider-the-opposite can be an effective strategy on reducing overconfidence. Despite the fact that Larrick (2004) classified this debiasing technique as a cognitive intervention, research has shown that it can be

implemented through technological strategies as well, such as decision aids that provide alternative hypotheses (Dreiseitl and Binder, 2005; Bhandari, Hassanein and Deaves, 2008; Huang, Hsu and Ku, 2012; Harada *et al.*, 2021; Sibbald *et al.*, 2021), advice about communicating prognosis to patients (Epstein *et al.*, 2021), and heuristic-based strategies to diagnose diseases (Marewski and Gigerenzer, 2012).

Decision aids have, indeed, been proposed as debiasing interventions to improve decision making (Leblanc, Norman and Brooks, 2001; Larrick, 2004; Sieck and Arkes, 2005; Berner and Graber, 2008; Kostopoulou, Porat, *et al.*, 2017; Singh *et al.*, 2017). However, peoples' confidence in their own judgements often makes them reluctant to use decision aids. As Sieck and Arkes (2005) highlighted, this could also be due to peoples' scepticism about the credibility of decision aids, the complexity of their underlying rules, and the perceived effort required to process additional information provided by the aids. Given the robustness of overconfidence, the researchers further suggested that it could be more helpful to try to increase people's confidence in decision aids, rather than trying to reduce their confidence in their judgements (Sieck and Arkes, 2005). In their review of decision technology, Edwards and Fasolo (2001) also raised the issues of human agency and accountability, concluding though that computerised decision aids will be widely adopted in the future, in an attempt to improve decision making (Edwards and Fasolo, 2001)[8].

To investigate the effects of decision aids on reducing cognitive biases, Huang and colleagues (2012) conducted a study on financial decisions. They recruited 241 students to complete a computer-based investment task. Participants were initially provided with brief information about two companies and were asked to indicate their preliminary investment decision, as well as their level of confidence in that decision. Subsequently, participants were able to obtain further information (either positive or negative) about the two alternatives, presented in brief titles. Participants could click on a title to reveal its contents. In half of the participants and depending on their choice, a message window appeared on their screen after they read their chosen piece of information, providing (counter)arguments against this information. That is, when participants chose a piece of information that favoured one of the companies,

[8] Despite the fact that agency and accountability are interesting concepts, further discussion about these would be beyond the scope of this thesis.

the pop-up message provided supportive information about the other company and vice versa.

When ready, participants provided their final decision and confidence in it. The study found that participants who had a preference towards information that favoured their choice (i.e., those who demonstrated confirmation bias) grew more confident in their decisions and were less likely to change them. However, for those who were presented with counterarguments, confirmation bias had a weaker effect, suggesting that reading contradictory information can mitigate the effects of confirmation bias on decision confidence and resistance to change mind (Huang, Hsu and Ku, 2012). Similar results have been found in asset allocation, where decision aids that provide suggestive and informative feedback (either textual or graphical) were found to reduce the effects of representativeness, framing and ambiguity aversion in financial decisions (Bhandari, Hassanein and Deaves, 2008).

In medical decision making, Dreiseitl and Binder (2005) conducted a study to explore whether physicians change their diagnosis when presented with decision support that contradicts their own diagnoses. In a within-participant design, 52 dermatologists were presented with 25 real lesion images on a computer screen, and asked to rate the malignacy of each lesion, on a scale from 0 (benign) to 10 (malignant). Malignancy ratings were used as a proxy for physicians' confidence in recommending a lesion excision or not. Physicians were also asked to provide a dichotomous recommendation (either "excise" or "no excise"). After providing their responses, physicians were presented with diagnostic and treatment suggestions (some of them intentionally incorrect), made by a computerised decision aid. They were then given the opportunity to revise their responses. The results showed that, when the aid provided suggestions that contradicted physicians' decisions, physicians followed the suggestions in 24% of the cases. Also, in line with Harada *et al.*, (2021) more experienced physicians were less likely to accept the decision aid's suggestions. Finally, the study did not report physicians' diagnostic performance and whether the decision aid had any impact of physicians' confidence; however, it reported that more confident physicians were less likely to follow the recommendations of the aid, suggesting that initial confidence possibly affected their willingness to follow advice that contradicted their own diagnoses.

Apart from the cognitive factors described above (e.g., overconfidence), as well as the possible explanations provided by Sieck and Arkes (2005) – e.g., credibility of decision aids – and Edwards and Fasolo (2001) – i.e., accountability and agency, there are other factors that would potentially affect the adoption and effectiveness of decision aids. Physicians' experience, as well as individual differences in cognitive styles and need for cognitive closure are all factors that could affect physician's diagnostic flexibility and openness to suggestions. For instance, experienced physicians are thought to be more confident in their judgements, to respond to a problem more intuitively and are also less likely to seek advice and change their mind (Eva *et al.*, 2010; Phua and Tan, 2013; Lambe, Hevey and Kelly, 2018). Likewise, people with high need for cognitive closure are generally more reluctant to seek further information or consider alternative opinions (Webster and Kruglanski, 1994; Roets and Van Hiel, 2011; Roets *et al.*, 2014). They have a desire to reach a conclusion quickly (i.e., closure), in an attempt to avoid uncertainty that requires cognitive effort (Roets and Van Hiel, 2011). On the other hand, people who engage in Actively Open-minded Thinking (AOT) – a thinking style that involves adopting various perspectives and considering arguments that oppose one's beliefs – are more likely to seek further information and consider alternatives (Haran, Ritov and Mellers, 2013).

The existence of cognitive biases in clinical reasoning highlights the importance of de-biasing approaches for the prevention of diagnostic error. Given that decision aids have been shown to improve diagnostic accuracy (Kostopoulou, Rosen, *et al.*, 2015; Kostopoulou, Porat, *et al.*, 2017; Sibbald *et al.*, 2021), it is important to a) explore how they influence physicians' cognition during the diagnostic process and b) understand why this intervention is effective in order to develop it further.

1.5. Research questions and objectives

This programme of work aimed to understand how decision aids affect and improve clinical reasoning. In a series of experimental studies, in which GPs assessed hypothetical patient scenarios online, we tested the hypothesis that providing diagnostic suggestions (via a decision aid) early in the clinical consultation mitigates

cognitive biases in diagnostic reasoning. Specifically, in Study 1, we tested the effects of the decision aid on physicians' diagnostic confidence and diagnostic change. In Study 2, alongside diagnostic confidence and diagnostic change, we tested the effect of the decision aid on information search and information evaluation. In Study 3, we tested the same mechanisms, with an additional manipulation in the decision aid (i.e., removing GPs' own diagnosis from the list of suggestions). The overarching hypothesis was that the provision of diagnostic suggestions (in the form of a decision aid) would interrupt physicians' intuitive thinking by reducing 1) initial confidence, 2) information distortion, and 3) premature closure, yielding greater openness to alternative hypotheses.

Summary

- Patient safety is a top priority for WHO
- Most people will experience at least one diagnostic error in their lifetime
- Decision aid is important in healthcare to minimise diagnostic error
- Decision making biases (e.g., overconfidence, confirmation bias, premature closure) can operate during the diagnostic process and lead to erroneous clinical decisions.
- Decision aids, such as early diagnostic suggestions, can de-bias physicians' thinking and improve accuracy of medical diagnosis.

CHAPTER 2: Effects of diagnostic suggestions on physicians' diagnostic confidence (Study 1)

2.1. Introduction

The first study of this research was an attempt to explore a potential mechanism via which the decision aid described above impacts cognition and improves diagnostic accuracy. Specifically, we tested whether the provision of early diagnostic suggestions can reduce physicians' certainty about their initial diagnostic hypotheses (i.e., initial confidence)[9], resulting in more frequent diagnostic changes when physicians encounter new information that is not entirely accounted for by the initial hypotheses. This was to extend the initial work by Kostopoulou and colleagues that found that early diagnostic suggestions can improve diagnostic accuracy, without revealing the underlying mechanisms of this improvement (Kostopoulou, Rosen, *et al.*, 2015; Kostopoulou, Porat, *et al.*, 2017). In this research, we also measure diagnostic certainty and diagnostic change over the course of the decision-making process (by taking multiple measures), something that was not explored in the original studies by Kostopoulou and colleagues.

2.2. Research questions and objectives

This study aimed to investigate whether the presence (versus absence) of a decision aid in the form of diagnostic alternatives, provided after a brief patient description, reduces physicians' certainty about their initial diagnosis, which then leads them to change diagnosis more frequently in the presence of inconsistent patient information. It also aimed to investigate whether the effect of the decision aid depends on physicians' thinking style. To test this, a version of the Actively Open-minded Thinking (AOT) scale, developed by Jonathan Baron (2018), was used (see Materials below)[10].

[9] The words "confidence" and "certainty" are used interchangeably. When we refer to the empirical studies conducted for this thesis, we use the word "certainty" to reflect the wording in the online task.

[10] The AOT scale was initially developed by Stanovich and West (1997). The version used in the present work developed by Jonathan Baron (2018).

2.3. Hypotheses

We hypothesised that when GPs received diagnostic suggestions they would:

H_1: indicate lower certainty about their initial diagnosis after receiving the suggestions, and

H_2: change diagnosis more frequently than when not receiving the suggestions.

2.4. Methods

2.4.1. Participants

Participants were fully qualified GPs and GP trainees currently practising in England.

2.4.2. Sample size

We powered the study for the main hypothesis that when GPs receive diagnostic suggestions, they would be less certain about their initial diagnosis than when they do not receive suggestions. To calculate sample size, the G*Power software was used (version 3.0.10). To detect a small effect size (Cohen's $f^2 = 0.05$) in a multiple linear regression (Omnibus: R^2 deviation from zero) with one predictor (suggestions versus no suggestions) and one covariate (AOT score), with 80% power and $\alpha = 0.05$, 196 participants would be required. We increased this to 200 participants, assuming that not all responses would be analysable.

We opted for a within participant design to have greater power to detect an effect of the decision aid with fewer GPs. A between participants design would require three times the sample size (i.e., approximately 600 participants) and it would not be feasible to recruit such a large number of GPs.

2.4.3. Recruitment

Participants were recruited through a database that the primary supervisor (Dr Olga Kostopoulou) has compiled. The database contains e-mail addresses of GPs who took part in previous studies on medical decision making and have consented to be contacted regarding participation in future studies. Study participants were fully qualified GPs and GP trainees, currently practising in England. Each GP was sent an invitation e-mail with information about the study and a hyperlink to access the study website (Appendix 3). GPs who took part were offered a £10 Amazon voucher as a token of appreciation. The value of the voucher was determined based on the estimated time to complete the study (i.e., approximately 10 minutes). Recruitment was closed as soon as the required sample size was reached.

The study was approved by the Head of Department of Surgery & Cancer and by the Imperial Joint Research Compliance Office (ICREC: 18IC4590).

2.4.4. Materials

Two patient scenarios, initially developed by Kostopoulou and colleagues (2017), were used. The scenarios were adapted for the purposes of the present study. In both scenarios, the amount of information was not sufficient for a definite diagnosis. One patient scenario was about James, a 32-year-old man, who smokes 10 cigarettes a day, with a BMI of 24 and a past medical history of knee injury two years ago. He is currently not taking any medication and his last consultation was about tonsilitis six months ago. James presents with a few weeks' chest pain when he takes a breath. The pain started after a cold. Initially, most GPs were expected to consider musculoskeletal or pleuritic chest pain, as the patient is young and has no risk factors except for smoking. Subsequently, James says that the pain is sharp, on the right side of his arm and that he feels it only when he takes a breath. James also says that he has a dry cough since his cold four weeks ago. James finally says that he has been feeling more tired than usual, although he describes himself as a fit and healthy person. His physical examination is normal (respiratory rate 20 breaths per minute, pulse 70, temperature 36.5 °C, oxygen saturation 98% on air) and when the

GP presses on his chest, James does not feel the pain. At this stage, given the ongoing tiredness, most GPs were expected to think of other causes of the pain, such as acute or more chronic infection, pulmonary embolism, or pneumothorax. Also, the fact that James feels no pain when the GP presses his chest makes musculoskeletal pain less likely.

The other patient scenario was about Derek, a 68-year-old man, who is obese (BMI of 31), a heavy smoker for 50 years (he used to smoke 3 packets a day but is currently smoking one packet a day) and a regular drinker (12 units per week). Derek has no medical history and is not taking any medication. His last consultation was about an infected laceration of his finger four years ago. Derek appears comfortable at rest, but he has been experiencing worsening breathlessness for the past six months. He says that the last time he was rushing to catch the bus he became very breathless. Initially, given his age and risk factors, most GPs were expected to consider chronic obstructive pulmonary disease (COPD) and/or ischaemic heart disease (IHD). Subsequently, Derek said that he feels tired lately. His physical examination results revealed raised blood pressure (132/104). Given this information, COPD and IHD are still possible. However, the reported fatigue, along with his risk factors (i.e., smoking and obesity), introduce more diagnostic possibilities, such as heart failure and lung cancer.

For each patient scenario, there was a list of diagnostic suggestions provided to participants (18 diagnoses for patient James and 23 diagnoses for patient Derek). The diagnostic suggestions were used in the Kostopoulou and colleagues (2017) study and generated by two academic GPs (also see Kostopoulou *et al.*, 2017). The suggestions on the list were presented in alphabetical order to avoid unintended effects (e.g., perceived likelihood of a disease based on the order presented).

2.4.4.1. Actively Open-minded Thinking scale (AOT)

Jonathan Baron (2006) described AOT as “good thinking”. AOT is a model of rational thinking and has been termed as such because it is a) “open” to alternative explanations that oppose an initial judgement (open-mindedness) and b) “active” in

searching for evidence to disconfirm pre-established beliefs (Baron, 2006, 2019). In order to measure AOT, a scale was initially developed by Stanovich and West (1997). However, throughout this thesis, a shorter and more recent version of the original AOT scale was used, after being advised by Professor Jonathan Baron (personal communication, 2nd of May 2018). The scale consists of eleven statements that measure individual differences in the way people evaluate information and form their beliefs. The statements are presented below:

1. Allowing oneself to be convinced by a solid opposing argument is a sign of good character.
2. People should take into consideration evidence that goes against conclusions they favour.
3. Being undecided or unsure is the result of muddled thinking.
4. People should revise their conclusions in response to relevant new information.
5. Changing your mind is a sign of weakness.
6. People should search actively for reasons why they might be wrong.
7. It is OK to ignore evidence against your established beliefs.
8. It is important to be loyal to your beliefs even when evidence is brought to bear against them.
9. When we are faced with a new question, the first answer that occurs to us is usually best.
10. Good thinking leads to uncertainty when there are good arguments on both sides.
11. When faced with a new question, we should consider more than one possible answer before reaching a conclusion.

For each statement, respondents are asked to indicate their agreement on a 5-point scale, ranging from “Completely disagree” to “Completely agree” with midpoint “Neutral”. Statements 3, 5, 7, 8 and 9 are reverse-scored.

Baron further suggested that AOT is a way to prevent various biases to occur, including overconfidence, confirmation bias, premature closure, and resistance to change. These cognitive biases potentially occur during the diagnostic process and

can affect clinical reasoning. These are also the precise psychological mechanisms that the present research explored. For this reason, we chose the AOT scale and tested whether it moderates the effects of the diagnostic suggestions. Since people with high scores in AOT consider more hypotheses and seek evidence against their prior beliefs, we expected that any debiasing effect of the diagnostic suggestions would be diminished, compared to people with low AOT.

2.4.5. Design

The design of the present study was finalised after piloting different patient cases and procedures; The process is described in Appendix 8.

A within-participant design with one independent variable (Condition: decision aid versus no aid) was employed. Each GP was presented with the two patient scenarios described above. In one patient scenario, GPs received the decision aid in the form of diagnostic suggestions (Aided condition) while in the other scenario they did not (Control condition). The scenarios were presented in a random order. The presentation of the decision aid was counterbalanced: half of the GPs received diagnostic suggestions in the first patient scenario while the other half received the suggestions in the second scenario. Each scenario was presented with and without diagnostic suggestions an equal number of times.

2.4.6. Procedure

The study was conducted entirely online. The two patient scenarios were administered using the Qualtrics software platform (Qualtrics.com). Participating GPs were sent an invitation e-mail that contained a brief description of the study, as well as a hyperlink via which they could access the study website. GPs were able to complete the study at a time and place that was convenient to them. Upon accessing the study site, GPs read the information sheet and indicated consent (Appendices 1 and 2 respectively). They were then asked to indicate their gender (male or female) and GP status (fully qualified GP or GP trainee). Fully qualified GPs were asked to

provide their year of GP certification. GPs were then presented with the two patient scenarios in a random order. In one scenario (Aided condition), GPs received diagnostic suggestions. In the other scenario (Control condition) GPs did not receive suggestions.

Patient information in both scenarios was presented in two steps. In the first step, GPs saw a patient description, containing a brief patient history and the presenting problem. They were then asked to provide their initial diagnosis in a text box and indicate how certain they were on a visual analogue scale (VAS), ranging from 0 to 10 (Figure 2.1.)[11].

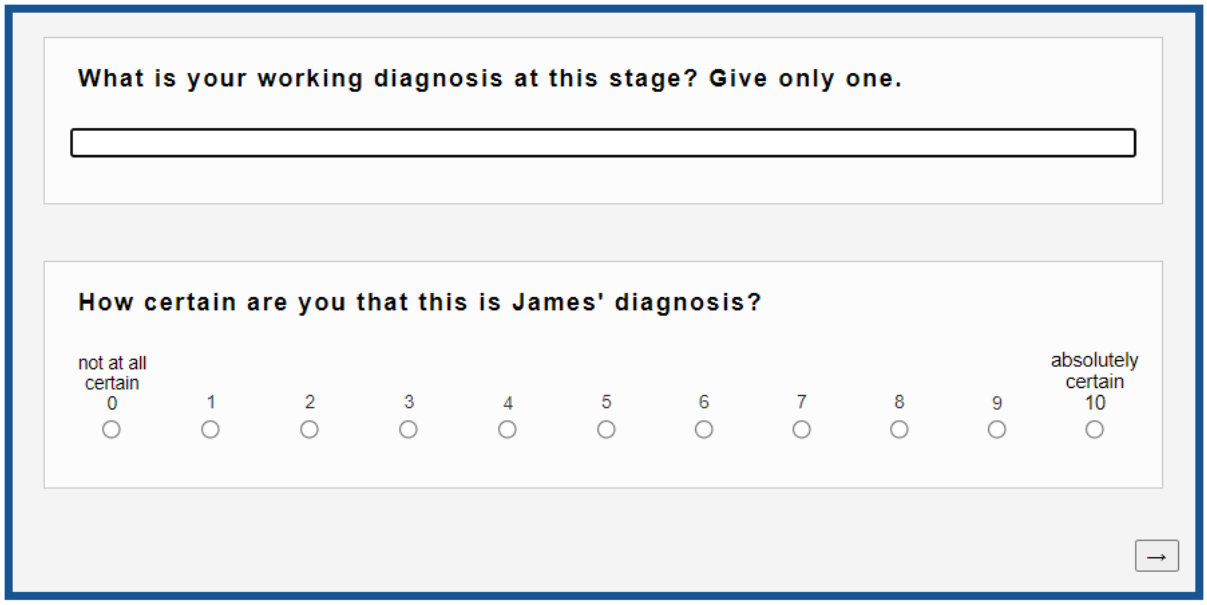


Figure 2.1. Initial diagnosis and certainty measurement scale.

After GPs provided their initial diagnosis and certainty, in the Aided condition only, they were provided with a list of diagnostic suggestions (Figure 2.2.). In the Control condition, GPs did not receive diagnostic suggestions. To ensure that GPs would not proceed without reading the diagnostic suggestions, the suggestions remained on the screen for a minimum of ten seconds and GPs were asked to confirm that they had read them. After this, they could proceed to the next screen.

[11] This scale was previously used by Kostopoulou, Porat and colleagues (2017) in their study on the evaluation of the decision described in the Introduction. They chose the term "certain" over the term "confident", as less emotionally loaded. This term has also been used in the experiments for this thesis in order to make comparisons with the findings of Kostopoulou et al. (2017).

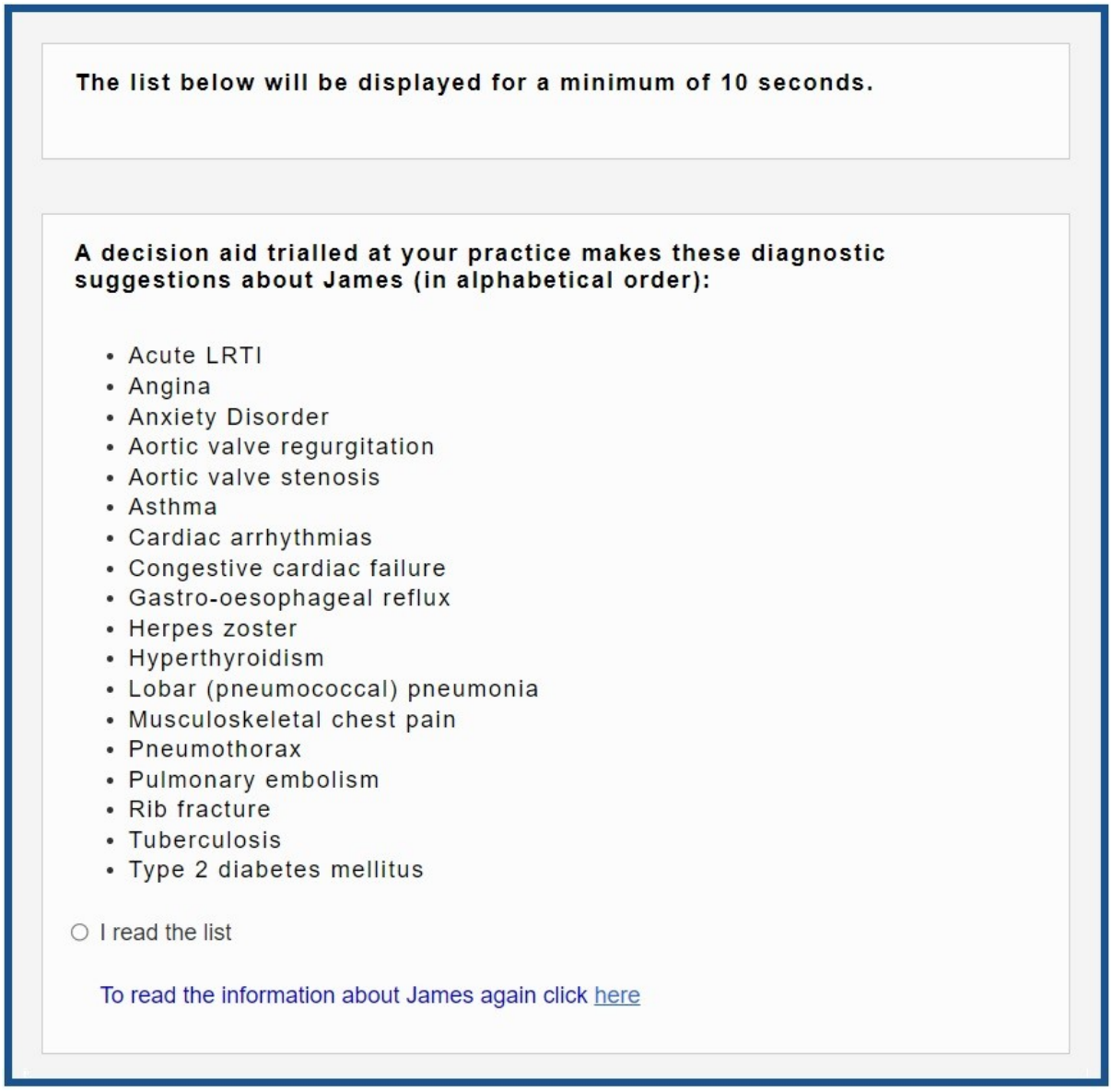


Figure 2.2. The decision aid: a list of diagnostic suggestions (James scenario).

In the second step, GPs were presented with additional information about the patient including physical examination results. They were then asked to update their certainty about their initial diagnosis on a 0-10 scale, as before. On the next screen, GPs wrote in free text the next most important question that they would like to ask the patient. Subsequently, they provided their final diagnosis and certainty. Finally, they were asked to indicate whether they would order investigations for the patient and – if yes – what investigations they would order.

The study flow did not allow GPs to return to previous screens and, therefore, they were not able to change their responses. To keep the patient information accessible

to GPs as the study progressed, we provided hyperlinks containing the initial patient description and the additional patient information. After completing the two patient scenarios, GPs were asked to complete the AOT scale. For each statement, GPs were asked to indicate their agreement on a 5-point scale, ranging from “Completely disagree” to “Completely agree”.

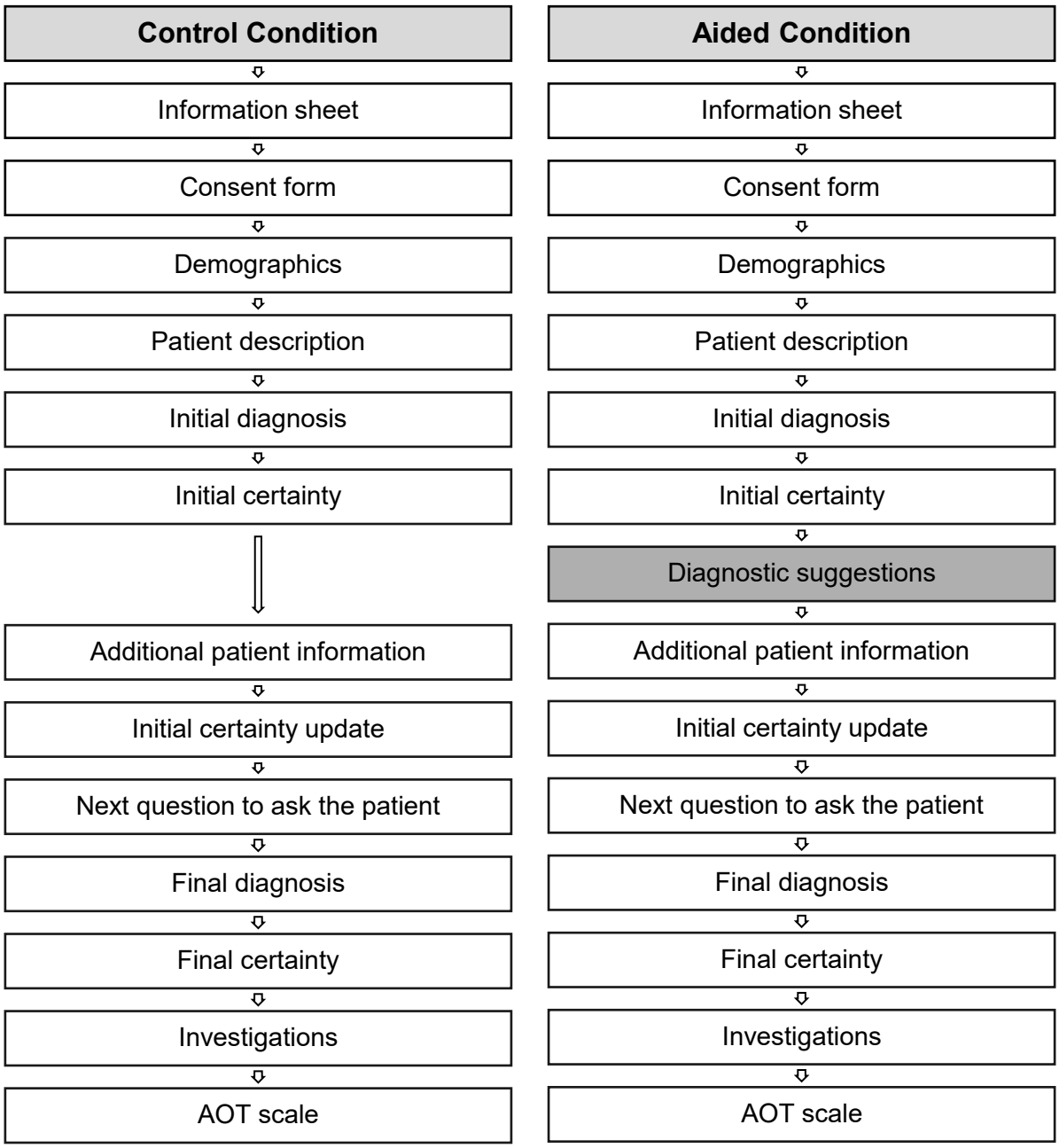


Figure 2.3. Flow of activity

2.5. Analyses

2.5.1. Dependent variables

Based on the hypotheses above, the dependent variables were:

1. **Initial certainty**, a continuous variable measured on a 0-10 scale.
2. **Change in initial certainty** (H1), a continuous variable computed by subtracting the second measurement of certainty about the initial diagnosis from the first.
3. **Change in diagnosis** (H2), a dichotomous variable indicating whether the initial and final diagnoses were different (change, coded as 1, versus no change, coded as 0).
4. **Final certainty**, a continuous variable measured on a 0-10 scale.
5. **AOT score** (H3), a continuous variable.
6. **Number of investigations**, a count variable.

2.5.2. Independent variables

The following independent variables were used in the analyses:

1. Condition, coded as 0 for Control and 1 for Aided.
2. GP gender, coded as 0 for females and 1 for males.
3. Years of experience, a continuous variable.
4. The order of aid presentation, coded as 1 to indicate that suggestions were presented in the first scenario and 0 to indicate that they were presented in the second scenario.
5. Scenario, coded as 0 for James (chest pain) and 1 for Derek (breathlessness).

2.5.3. Preparing the data for analysis

As mentioned above, the main predictor “Condition” had two levels (scenarios assessed without diagnostic suggestions versus scenarios assessed with diagnostic suggestions). To make the data more easily analysable, GPs’ responses were

simplified according to supervisors' suggestions by identifying synonyms or similar conditions and classifying them under a single category. Specifically, in the chest pain scenario (James), Lower and Upper Respiratory Tract Infections were classified under "RTI" (Respiratory Tract Infection). Costochondritis and Musculoskeletal chest pain were classified under "MSK" (Musculoskeletal). Angina, Coronary Artery Disease and Ischaemic Heart Disease (IHD) were classified under "Cardiac/Angina/IHD". In the breathlessness scenario (Derek), carcinoma, malignancy and cancer were all classified under "Lung Cancer".

To determine whether there was a change in diagnosis, we set the following two criteria:

1. In case of a single initial diagnosis, any change in diagnosis, including a switch to a different diagnosis or an addition of a new diagnosis, counted as a change.
2. In case of multiple initial diagnoses, any change in the diagnostic set, such as the addition or removal of diagnoses, counted as a change.

2.5.4. Statistical tests

To analyse the data, we run multilevel regressions with random intercept by participant: linear for continuous dependent variables, logistic for dichotomous dependent variables, and Poisson for the count variable. The main predictor was condition (suggestions versus no suggestions), on which we regressed all the dependent variables in simple models (Step 1). Each model was subsequently extended from simple to multiple one to include the following variables and interactions:

1. Condition, patient scenario and interaction (Step 2)
2. Condition, patient scenario, initial certainty and interactions (Step 3)
3. Condition, patient scenario, initial certainty, gender, GP experience, AOT score and interactions (Step 4).

2.5.5. Data exclusion

We excluded incomplete responses (e.g., GPs who completed only one scenario or closed their browser before completing the scenarios). We also excluded non-analysable responses (e.g., responses such as “uncertain” or “need more information”, provided as either initial or final diagnosis).

2.6. Results

We recruited 217 GPs, however, 21 of them did not complete their participation in full and their responses were removed. Two GPs who did not provide an analysable diagnosis (i.e., “uncertain” and “need more information”) were also excluded. The remaining 194 GPs were included in the analysis. Of those 194 GPs, 101 were males (52%). Only two were trainees. The average professional experience of the sample was 12.63 years (SD 9.10), ranging from 0 to 41 years (median 10 years). Participation lasted on average 10 minutes and 40 seconds. Table 2.1. presents a summary of the results across the two conditions (Aided: with diagnostic suggestions and Control: without diagnostic suggestions).

Dependent variable	Control condition Mean (SD)	Aided condition Mean (SD)
Initial certainty	5.68 (1.62)	5.91 (1.53)
Second measurement of initial certainty	4.96 (2.17)	4.92 (2.12)
Change in initial certainty	-0.72 (1.63)	-0.99 (1.72)
Change in diagnosis	46.90%	50.00%
Number of investigations	3.15 (2.49)	3.18 (2.41)
Certainty about the final diagnosis	5.62 (1.75)	5.71 (1.66)

Table 2.1. Results by condition (with and without diagnostic suggestions).
Change in initial certainty refers to the difference between the first and second measurements of initial certainty and was always negative (reduction).

2.6.1. Change in initial certainty (H1)

GPs' change in certainty from the first to the second measurement was found to be in the expected direction (i.e., reduction rather than an increase) across conditions and scenarios. We expected a reduction in GPs' certainty in their initial diagnosis, as the additional patient information was not consistent with the initial patient description and, as a result, GPs were expected to change their initial diagnosis after obtaining further patient information (also see Materials above). In both conditions, we observed a reduction in initial certainty after additional information was received: GPs reduced their certainty by an average of 0.85 units (SD 1.68) on the 0-10 scale. Reduction in certainty was greater when diagnostic suggestions were provided (0.99 versus 0.72). However, when we regressed the dependent variable on condition using simple linear regression, the difference between control and experimental condition was not statistically significant (b = 0.273, [-0.06, 0.60], p = 0.105).

When we subsequently added patient scenario and initial certainty to the regression model (in two different steps), the analysis revealed that both were significantly associated with change in certainty. Specifically, we observed a smaller drop in certainty for patient Derek than patient James (means -0.56 versus -1.15, respectively, b = -0.599, [-0.92, -0.28], p < 0.001). There was also a positive association between initial certainty and change in initial certainty, suggesting that GPs with higher initial certainty reduced it more at second measurement (b = 0.175, [0.07, 0.28], p = 0.001). There were no interactions.

In addition to the regression models above, we run separate regression models for each patient scenario. There was a larger reduction in certainty in patient James (chest pain) than patient Derek (breathlessness). This reduction was slightly larger when diagnostic suggestions were provided (-1.23 versus -1.07), but not significantly so (b = 0.16, [-0.42, 0.74], p = 0.583). In patient Derek (breathlessness), the reduction in certainty when diagnostic suggestions were provided was more than twice as large than when no suggestions were provided (-0.76 versus -0.35). This difference was statistically significant (b = 0.41, [0.09, 0.73], p = 0.013).

Linear regression of change in initial certainty							
Step		Estimate	Std. Error	t	Sig.	95% CI	
						Lower Bound	Upper Bound
1	Intercept	.716495	.120392	5.951	.000	.479788	.953202
	Condition	.273196	.167822	1.628	.105	-.057804	.604196
2	Intercept	1.009658	.142858	7.068	.000	.728712	1.290604
	Condition	.285540	.162653	1.756	.081	-.035277	.606357
	Scenario	-.598671	.162653	-3.681	.000	-.919488	-.277854
3	Intercept	.053681	.327567	.164	.870	-.590553	.697915
	Condition	.246511	.159953	1.541	.125	-.068978	.561999
	Scenario	-.670795	.161050	-4.165	.000	-.988408	-.353183
	Certainty 1	.174670	.054034	3.233	.001	.068368	.280971
4	Intercept	-1.424264	1.216557	-1.171	.243	-3.822773	.974245
	Condition	.207827	.156882	1.325	.187	-.101625	.517278
	Scenario	-.711134	.157981	-4.501	.000	-1.022716	-.399552
	Certainty 1	.183404	.054800	3.347	.001	.075596	.291211
	Gender	.140938	.174926	.806	.421	-.204151	.486028
	Experience	.006757	.009628	.702	.484	-.012237	.025750
	AOT score	.290213	.275378	1.054	.293	-.253013	.833438
Dependent Variable: Change in initial certainty							

Table 2.2. Regression model of change in initial Certainty

Linear regression of change in initial certainty (James)							
Step		Estimate	Std. Error	t	Sig.	95% CI	
						Lower Bound	Upper Bound
1	Intercept	1.070707	.204820	5.228	.000	.666720	1.474694
	Condition	.160872	.292693	.550	.583	-.416434	.738178
Dependent Variable: Change initial certainty							

Table 2.3. Regression model of change in initial certainty (James)

Linear regression of change in initial certainty (Derek)							
Step		Estimate	Std. Error	t	Sig.	95% CI	
						Lower Bound	Upper Bound
1	Intercept	.347368	.117475	2.957	.003	.115661	.579076
	Condition	.410207	.164448	2.494	.013	.085851	.734564
Dependent Variable: Change in initial certainty							

Table 2.4. Regression model of change in initial certainty (Derek)

To test whether the order in which the Aided scenario was seen (first or second) affected GPs' change in initial certainty, we included the variable "Condition order" in the regression model. Change in initial certainty was indeed associated with the order in which the Aided scenario was seen. Specifically, when diagnostic suggestions were provided in the scenario that was seen first, there was a larger reduction in certainty across both scenarios (b = 0.423 [0.08, 0.76], p = 0.015), suggesting a possible spillover effect to the scenario seen second (Table 2.5.).

Linear regression of change in initial certainty							
Step		Estimate	Std. Error	t	Sig.	95% CI	
						Lower Bound	Upper Bound
5	Intercept	-1.306991	1.201477	-1.088	.278	-3.675851	1.061870
	Condition	.209952	.156991	1.337	.183	-.099716	.519621
	Patient	-.707085	.158074	-4.473	.000	-1.018851	-.395318
	Certainty 1	.173252	.054409	3.184	.002	.066210	.280293
	Gender	.120481	.172720	.698	.486	-.220271	.461232
	Experience	.005974	.009501	.629	.530	-.012769	.024718
	AOT score	.233600	.272563	.857	.393	-.304093	.771292
	Condition order	.422462	.172669	2.447	.015	.081818	.763106
Dependent Variable: Change in initial certainty							

Table 2.5. Regression model including variable "Condition order"

Table 2.6. presents the first and the second measurements of initial certainty, as well as reduction in GPs' certainty in their initial diagnosis by patient scenario and condition.

Condition	Scenario	Initial certainty Mean (SD)	Second measurement of initial certainty Mean (SD)	Change in certainty Mean (SD)
Control	James (chest pain)	5.44 (1.71)	4.37 (2.47)	-1.07 (2.01)
	Derek (breathlessness)	5.92 (1.49)	5.57 (1.59)	-0.35 (0.98)
	Overall	5.68 (1.62)	4.96 (2.17)	-0.72 (1.63)
Aided	James (chest pain)	5.73 (1.53)	4.49 (2.38)	-1.23 (2.06)
	Derek (breathlessness)	6.08 (1.52)	5.32 (1.76)	-0.76 (1.29)
	Overall	5.91 (1.53)	4.92 (2.12)	-0.99 (1.27)
Overall	James and Derek	5.79 (1.58)	4.94 (2.14)	-0.85 (1.68)

Table 2.6. Measurements of initial certainty and change in initial certainty across scenarios and conditions.
Negative values indicate reduction in certainty.

2.6.2. Change in diagnosis (H2)

We collected 388 initial diagnoses and 388 final diagnoses. Overall, GPs changed their diagnosis in 48.45% of responses (188/388). In line with our hypothesis (H2), when diagnostic suggestions were received, GPs changed diagnosis more frequently than when they were not received (97/194 versus 91/194, i.e., 50% versus 46.90%). To test the effects of condition on change in diagnosis, we run simple logistic regression and found no significant differences (OR = 1.13, [0.76, 1.69], p = 0.543).

When patient scenario and initial certainty were included in the model in different steps, they were both significantly associated with change in diagnosis (OR = 0.250, [0.16, 0.383], $p < 0.001$ and OR = 0.820, [0.71, 0.95], p = 0.008 respectively). Specifically, GPs with higher initial certainty changed diagnosis less frequently. Also, patient James (chest pain) had approximately double the number of diagnostic changes than patient Derek (breathlessness) (126/194 versus 62/194, i.e., 64.90% versus 32.00%). There were no interactions.

Change in diagnosis was also positively associated with change in certainty: GPs who reduced their certainty in their initial diagnosis more were more likely to change diagnosis in the end (OR = 2.332, [1.87, 2.91], p < 0.001).

Logistic regression of change in diagnosis										
Step		Coefficient	Std. Error	t	Sig.	95% CI		Exp(Coefficient)	95% CI for Exp(Coefficient)	
						Lower	Upper		Lower	Upper
1	Intercept	-.124	1.735	-.071	.943	-3.535	3.288	.883	.029	26.777
	Condition	.124	.2033	.609	.543	-.276	.524	1.132	.759	1.688
2	Intercept	.537	1.6461	.326	.744	-2.699	3.774	1.711	.067	43.539
	Condition	.171	.2162	.789	.430	-.254	.596	1.186	.775	1.814
	Scenario	-1.385	.2162	-6.405	.000	-1.810	-.960	.250	.164	.383
3	Intercept	1.634	1.6842	.970	.333	-1.678	4.945	5.122	.187	140.469
	Condition	.218	.2192	.994	.321	-.213	.649	1.244	.808	1.913
	Scenario	-1.333	.2188	-6.092	.000	-1.763	-.902	.264	.172	.406
	Certainty 1	-.198	.0740	-2.674	.008	-.343	-.052	.820	.709	.949
4	Intercept	-.041	2.2737	-.018	.986	-4.511	4.430	.960	.011	83.942
	Condition	.230	.2244	1.026	.305	-.211	.672	1.259	.810	1.957
	Scenario	-1.390	.2260	-6.149	.000	-1.834	-.945	.249	.160	.389
	Certainty 1	-.177	.0754	-2.347	.019	-.325	-.029	.838	.722	.972
	Gender	.594	.2346	2.534	.012	.133	1.056	1.812	1.142	2.874
	Experience	.035	.0131	2.686	.008	.009	.061	1.036	1.009	1.063
	AOT score	.212	.3698	.574	.566	-.515	.939	1.236	.598	2.558
Dependent variable: Change in diagnosis										

Table 2.7. Regression model of change in diagnosis

To determine whether the order in which diagnostic suggestions were provided (i.e., either in the first patient scenario or in the second) affected the frequency of diagnostic changes, we included the relevant variable in the model. The results showed that when diagnostic suggestions were provided in the scenario seen first, there were more diagnostic changes overall (OR = 1.653, [1.04, 2.62], p = 0.033), suggesting again a possible spillover effect to the scenario seen last (Table 2.8).

Logistic regression of change in diagnosis										
Step		Coefficient	Std. Error	t	Sig.	95% CI		Exp(Coefficient)	95% CI for Exp(Coefficient)	
						Lower	Upper		Lower	Upper
5	Intercept	.094	2.2752	.041	.967	-4.380	4.567	1.098	.013	96.299
	Condition	.233	.2260	1.031	.303	-.211	.677	1.262	.810	1.969
	Scenario	-1.403	.2279	-6.157	.000	-1.851	-.955	.246	.157	.385
	Certainty 1	-.191	.0762	-2.508	.013	-.341	-.041	.826	.711	.960
	Gender	.574	.2352	2.442	.015	.112	1.037	1.776	1.118	2.821
	Experience	.034	.0131	2.643	.009	.009	.060	1.035	1.009	1.062
	AOT score	.143	.3731	.383	.702	-.591	.877	1.154	.554	2.403
	Condition order	.503	.2353	2.136	.033	.040	.965	1.653	1.041	2.625
Dependent variable: Change in diagnosis										

Table 2.8. Regression model including Condition order.

Table 2.9. presents the change in GPs' certainty in their initial diagnosis across the two patient scenarios:

Scenario (main symptom)	Control condition % (proportion)	Aided condition % (proportion)	Overall % (proportion)
James (chest pain)	64.65 (64/99)	65.26 (62/95)	64.95 (126/194)
Derek (breathlessness)	28.42 (27/95)	35.35 (35/99)	31.96 (62/194)
Overall	46.90 (91/194)	50.00 (97/194)	48.45 (188/388)

Table 2.9. Frequency of diagnostic changes across patient scenarios.

In addition to the regression models above, we run separate logistic regression models for each patient scenario (Tables 2.10. and 2.11.). The chest pain scenario produced more frequent diagnostic changes than the breathlessness scenario (64.95% versus 31.96%). Within patient scenarios, the provision of diagnostic suggestions resulted in almost equal diagnostic changes in the chest pain scenario with no significant differences (64.65% with suggestions versus 65.26% without suggestions, OR = 1.027, [0.57, 1.87], p = 0.929). In the breathlessness scenario, the provision of diagnostic suggestions led to more frequent diagnostic changes

(28.42% versus 35.35%), but this difference was not significant either (OR = 1.377, [0.74, 2.54], p = 0.305).

Logistic regression of change in diagnosis (James)										
Step		Coefficient	Std. Error	t	Sig.	95% CI		Exp(Coefficient)	95% CI for Exp(Coefficient)	
						Lower	Upper		Lower	Upper
1	Intercept	.604	1.6714	.361	.718	-2.693	3.900	1.829	.068	49.415
	Condition	.027	.3026	.090	.929	-.570	.624	1.027	.566	1.866
Dependent variable: Change in diagnosis										

Table 2.10. Regression model of change in diagnosis (James).

Logistic regression of change in diagnosis (Derek)										
Step		Coefficient	Std. Error	t	Sig.	95% CI		Exp(Coefficient)	95% CI for Exp(Coefficient)	
						Lower	Upper		Lower	Upper
1	Intercept	-.924	1.6320	-.566	.572	-4.143	2.295	.397	.016	9.927
	Condition	.320	.3114	1.028	.305	-.294	.934	1.377	.745	2.545
Dependent variable: Change in diagnosis										

Table 2.11. Regression model of change in diagnosis (Derek).

Thus, we observed a similar pattern of results for both dependent variables.

2.6.3. Exploratory analyses

2.6.3.1. Actively Open-minded Thinking

The mean AOT score across GPs was 4.03 (SD 0.32, range 3.00 - 4.82, median 4) on the 1-5 scale, indicating that GPs were generally high in AOT. AOT was included as a covariate in the regression models described above. The regression models showed that there was no association with change in certainty (b = 0.29, [-0.25, 0.83], p = 0.296), or change in diagnosis (OR = 1.23, [0.59, 2.57], p = 0.571).

2.6.3.2. Initial diagnosis

Across the two patient scenarios, GPs generally responded as expected, i.e., they gave the diagnoses that the initial patient descriptions were designed to elicit (see Table 2.12.).

Scenario	Diagnosis	Proportion (%)
Chest pain (James)	Pleurisy	83/194 (42.80)
	Musculoskeletal chest pain	64/194 (33.00)
	Respiratory Tract Infection	22/194 (11.30)
	Pulmonary Embolism	6/194 (3.20)
	Pneumothorax	4/194 (2.10)
Breathlessness (Derek)	COPD	145/194 (74.70)
	Angina/Ischaemic Heart Disease	30/194 (15.00)
	Lung Cancer	10/194 (5.20)

Table 2.12. Most frequent initial diagnoses across the two patient scenarios.

2.6.3.3. Initial certainty

Given that the initial patient description was designed to suggest one or two diagnoses as more likely, GPs were expected to start with relatively high certainty. However, on average, GPs' certainty in their initial diagnosis was 5.79 (SD 1.58) on the 0-10 scale. Each patient scenario produced different levels of initial certainty. Specifically, in the chest pain scenario, GPs started with 5.58 certainty on average (SD 1.63) while in the breathlessness scenario, GPs' level of initial certainty was 6.00 (SD 1.50). This difference was significant ($t = -3.64$, $p < 0.001$).

As discussed above, initial certainty was included in all regression models as a covariate and found to be significantly associated with both change in certainty and change in diagnosis. It was also found to be negatively associated with AOT scores (b = -0.59, [-1.18, -0.003], p = 0.049) which suggests that GPs high in AOT expressed lower levels of certainty in their initial diagnosis.

Finally, initial certainty was slightly different between conditions (5.90, SD 1.53 with diagnostic suggestions and 5.67, SD 1.62 with no suggestions). This difference was borderline significant (t = -1.975, p = 0.0497). However, due to the temporal order, it was not possible that the diagnostic suggestions affected GPs' certainty in their initial diagnosis since they were provided after the certainty measurement.

2.6.3.4. Final diagnosis

GPs provided their final diagnosis after they saw the additional patient information. The frequency of final diagnoses is presented in Table 2.13.

Scenario	Diagnosis	Proportion (%)
Chest pain (James)	Pleurisy	53/194 (27.30)
	Musculoskeletal chest pain	19/194 (9.80)
	Respiratory Tract Infection	38/194 (19.60)
	Pulmonary Embolism	28/194 (14.40)
	Pneumothorax	9/194 (4.60)
	Pneumonia	19/194 (9.80)
Breathlessness (Derek)	COPD	104/194 (53.60)
	Angina/Ischaemic Heart Disease	41/194 (21.10)

	Lung Cancer	22/194 (11.30)

Table 2.13. Most frequent final diagnoses across the two patient scenarios.

2.6.3.5. Final certainty

Final certainty refers to GPs' certainty about their final diagnosis. Across patient scenarios and conditions, the average final certainty was moderate (see Table 2.14.)[12]. There were no significant differences in final certainty between the two patient scenarios (b = 0.06, [-0.16, 0.28], p = 0.600), or the two conditions (b = 0.09, [-0.13, 0.32], p = 0.417). Final certainty was positively associated with initial certainty (b = 0.67, [0.59, 0.76], p < 0.001) and negatively associated with diagnostic change (b = -0.55, [-0.80, -0.29], p < 0.001), suggesting that GPs who changed diagnosis were less certain in their new diagnosis than GPs who did not change diagnosis.

Condition	Scenario	Initial certainty Mean (SD)	Second measure of initial certainty Mean (SD)	Final certainty Mean (SD)
Control	James (chest pain)	5.44 (1.71)	4.37 (2.47)	5.55 (1.92)
	Derek (breathlessness)	5.92 (1.49)	5.57 (1.59)	5.68 (1.57)
	Overall	5.68 (1.62)	4.96 (2.17)	5.62 (1.75)
Aided	James (chest pain)	5.73 (1.53)	4.49 (2.38)	5.71 (1.82)
	Derek (breathlessness)	6.08 (1.52)	5.32 (1.76)	5.71 (1.51)
	Overall	5.91 (1.53)	4.92 (2.12)	5.71 (1.66)
Overall	James and Derek	5.79 (1.58)	4.94 (2.14)	5.66 (1.71)

Table 2.14. Certainty measurements across conditions and scenarios.

[12] Compared to Kostopoulou, Porat, *et al.*, (2017), there was a 2.15 absolute difference in final certainty on the 0-10 scale (Control condition: 5.62 versus 7.61 and Aided condition: 5.71 versus 8.01), showing that GPs in the present study were less certain about their final diagnosis. This is possibly explained by the fact that in Kostopoulou et al. (2017) GPs used the actual decision aid (as opposed of a non-interactive list of diagnostic suggestions) and assessed patient scenarios with richer information. Furthermore, in Kostopoulou et al. (2017), the difference in physicians' certainty between conditions was significant (Control: 7.61 versus Aided: 8.01, b = 0.39, [0.12, 0.67], p < 0.005).

2.6.3.6. Number of investigations

On average, GPs ordered 3.16 investigations across the two patient scenarios (SD 2.44). As Table 2.1. shows, there was a small difference between conditions, which was not significant (3.18 versus 3.15, IRR = 0.992, [0.89, 1.11], p = 0.886). When scenario was included in the model, it was found to be significantly associated with number of investigations (IRR = 0.430, [0.38, 0.49], p < 0.001). There was no interaction between condition and scenario. When initial certainty was added to the model, it was found to be associated with number of investigations, suggesting that GPs with higher initial certainty ordered fewer investigations compared to GPs with lower initial certainty (IRR = 0.964, [0.93, 1.00], p = 0.042). There was no interaction with scenario or condition. Finally, GPs with higher AOT scores ordered on average more investigations (IRR = 1.431, [1.20, 1.71], p < 0.001). There was no association with the rest of the variables.

2.6.3.7. Gender

Gender was found to be associated only with change in diagnosis: female GPs were more likely to change diagnoses than their male counterparts (OR = 1.81, [1.13, 2.88], p = 0.013).

2.6.3.8. Experience

There was a positive association between years of experience in general practice and change in diagnosis (OR = 1.036, [1.01, 1.06], p = 0.008). There was no other association.

2.7. Discussion

The present study explored the effects of a decision aid on GPs' certainty about their initial diagnosis, and on their inclination to change diagnosis when encountering new information that could not be entirely accounted by their initial diagnosis. Although

findings were in the expected direction, the evidence that the provision of early diagnostic suggestions can reduce physicians' initial certainty that the study provides is weak: differences between control experimental condition were not statistically significant, except in the case of one scenario. There was no evidence that this reduction in initial certainty makes a change in diagnosis more likely. AOT score was not found to be a good predictor for either change in certainty or change in diagnosis.

Despite the fact that the main hypotheses were not confirmed, the study revealed some interesting associations. GPs with higher initial certainty were less likely to change diagnosis. This is in line with Thompson and colleagues (2011, 2013), who tested the effects of answer fluency (i.e., the speed with which a response is given) and perceptual fluency (i.e., the ease with which information is processed) on people's certainty in their decisions (i.e., their feeling of rightness) and found that more confident decision makers were less likely to change their mind (Thompson, Prowse Turner and Pennycook, 2011; Thompson, Ackerman, *et al.*, 2013).

Another interesting finding from the exploratory analyses was that female GPs were more likely to change diagnosis than their male counterparts. A subsequent search in the literature was conducted for evidence on the effects of gender on changing one's mind without finding any studies that reported similar results. This difference could have been explained by lower initial certainty in female GPs, as previous studies have shown that males tend to be overconfident compared to females (Barber and Odean, 2001; De Paola, Gioia and Scoppa, 2014). However, in the present study, this gender effect cannot be explained by lower initial certainty, or greater reduction in certainty, as initial certainty ratings were identical across male and female GPs (i.e., 5.79 versus 5.79).

Finally, in terms of investigations that GPs ordered in each patient scenario, the results showed that GPs with higher AOT scores were more likely to order more investigations, as we would expect. The same was true for GPs with lower certainty in their final diagnosis. However, the decision aid did not have any effect on the number of investigations that GPs ordered for the patients.

We found some evidence that the provision of diagnostic suggestions at the early stages of the diagnostic process can reduce physicians' certainty in their initial diagnosis, but no evidence that it can lead to more frequent diagnostic changes. The choice of scenarios used in the study, as well as potential spillover effects, may have compromised the ability to detect an effect on diagnostic changes. It is also possible that the patient scenarios did not work as we expected in terms of diagnostic certainty. They were designed to induce high initial certainty by providing strong signals of certain diseases. The level of initial certainty was expected to reduce at a later stage when additional patient information was presented. However, the average initial certainty was rather moderate (i.e., 5.79).

It is possible that the initial patient description was not sufficient to induce high initial certainty, given the limited information. The reason for providing only limited information in the initial patient description was our intention to capture GPs' very first, intuitive response to the problem (i.e., the first thoughts that come to mind when they face a patient case). The additional patient information contained signals of other possible diseases, inducing, thus, uncertainty about the most likely diagnosis. The decision aid was expected to make GPs switch from intuitive to analytical mode of thinking, by reducing initial certainty, given that there were other possible diagnoses to consider. The decision aid did make GPs less certain about their initial diagnosis. However, the reduction in certainty did not make GPs change diagnosis.

It is also possible that GPs interpreted the symptoms differently from what we were expecting, given the variability in responses. Furthermore, it is also possible that GPs approached the patient scenarios with an analytical mindset from the start, given that they knew that they were being studied. For these reasons, any effects of the decision aid might have weakened.

Finally, the diagnostic suggestions themselves might have had a negative effect on GPs (also see Appendix 8: Piloting for Study 1). Based on the limited patient description, some of the suggestions seemed very unlikely and this might have been a distraction to GPs (common and uncommon diagnoses mixed in the list, as mentioned by GPs during piloting). In addition, the fact that the list of suggestions was relatively lengthy (i.e., 18 diagnoses for James and 23 for Derek) might have

made GPs more confused, or pay less attention to it, as a response to cognitive load (also see Schwartz, 2005). Compared to Kostopoulou and colleagues' studies (2015, 2017) where these diagnostic suggestions were presented to GPs and improved their diagnostic accuracy, the present study was different in several ways:

1. It used only two scenarios, whereas in Kostopoulou and colleagues (2015) study GPs assessed nine patient scenarios.
2. It used scenarios with limited patient history and without sufficient information for a definite diagnosis, whereas in Kostopoulou and colleagues studies the scenarios were much richer in information, so the researchers were able to measure the accuracy of GPs' diagnoses.
3. It used pre-fixed patient information and did not allow GPs to request further information in order to reach a conclusion.
4. It focused on GPs' certainty in their initial diagnosis (and not to their overall certainty after the intervention of the aid), which is not necessarily the only mechanism that can explain the effect of the decision aid.

The present study shed some light on the psychological mechanisms that could underlie the provision of early diagnostic suggestions as a decision aid, and how it might reduce physicians' certainty in their initial diagnostic hypotheses. In contrast to previous studies in the field (e.g., Friedman, 2005 and Kostopoulou, 2017), where certainty and diagnosis were measured only after physicians had all the information about the patient and after were presented with the decision aid, this study attempted to capture GPs' diagnosis and certainty over the course of the diagnostic task (i.e., after a brief patient description, after new information and after final diagnosis). Likewise, diagnosis was given before and after the use of the decision aid. This was due to the fact that, in the presence of new information, a leading diagnosis could change during the diagnostic process. Furthermore, in contrast to the present study, previous studies measured physicians' diagnostic accuracy and how decision aids improve this. In the present study, we focused more on how physicians' diagnostic hypotheses and certainty levels change during the consultation. The results of this study are, therefore, not directly comparable with the results from previous studies.

Decision aids have been found to improve diagnostic accuracy in primary care (Kostopoulou, Rosen, *et al.*, 2015; Kostopoulou, Porat, *et al.*, 2017). The present study did not provide strong evidence that GPs improved performance could be due to a reduction in certainty. A study with different scenarios and possibly with a list of fewer diagnostic suggestions for GPs to consider could yield different results.

Future research could explore different psychological mechanisms that have not been captures in the present study, to test whether diagnostic suggestions can make GPs less biased in the way they search for and evaluate patient information. For instance, it could explore the effects of diagnostic suggestions on confirmation bias or premature closure (also see Chapter 1: Introduction). In a similar online task with hypothetical clinical scenarios, GPs could have the opportunity to seek patient information themselves (e.g., by choosing from a list of available items of information about the patient) before they reach a diagnostic conclusion. This would help them have a more complete view of the patient and possibly express greater certainty in their diagnostic judgements. Such a study would be more realistic, given that GPs would be able to actively search for information they need to make a diagnosis. Such a study would also give us insights on what kind of strategy GPs use when they search for clinical information (e.g., hypothesis-confirming or hypothesis-testing) and how extensive this search would be (e.g., do they obtain all the available information to make a diagnosis or reach a conclusion prematurely?). It would also be useful to include a management plan, something that was not asked in the present study, which could include investigations, referral to a specialist and prescribing medications. This would possibly help the diagnostic process flow more naturally.

Finally, given that the two patient scenarios used in the present study did not produce the expected results (e.g., the participants possibly evaluated the clinical information differently and provided different diagnoses), it would be interesting to measure how clinical cues are interpreted. In a similar study, after a brief patient information where GPs would provide an initial diagnosis, they could be presented with a set of neutral (i.e., non-diagnostic) and open to interpretation clinical cues (e.g., patient symptoms) and asked to rate how much each of them supports their initial diagnosis. In such a study, the main hypothesis could be that diagnostic

suggestions would make GPs evaluate neutral patient information more fairly, rather than as supportive of their initial diagnostic hypotheses.

Summary

- 194 GPs assessed two online patient scenarios, one with and the other without suggestions
- Diagnostic suggestions led to greater change in initial certainty in one scenario only
- Physicians with higher initial certainty reduced it more after additional patient information
- Physicians with higher initial certainty changed diagnosis less frequently
- Physicians with higher AOT score had lower initial certainty and ordered on average more investigations.

CHAPTER 3: Effects of diagnostic suggestions on physicians' confidence, information search and evaluation (Study 2)

3.1. Introduction

In Study 1 we tested whether the provision of diagnostic suggestions after a brief patient description can reduce diagnostic certainty and lead to more frequent changes in diagnosis when encountering subsequent information that does not confirm an initial diagnostic hypothesis. The results showed that when GPs received diagnostic suggestions, their certainty about their initial diagnosis significantly dropped but this happened only in one of the two patient scenarios used. In the present study, we addressed several limitations of Study 1 and tried to increase our ability to detect an effect of providing diagnostic suggestions early. Specifically, in this study we used different patient scenarios, where the patient description contained stronger signals with greater diagnostic value, in order to induce higher initial certainty. Each scenario had two different versions, one suggesting a serious and the other suggesting a less serious disease. This was to allow for greater generalisability, while ensuring that the effects are not limited by the severity of suggested diagnosis. We also reduced the number of diagnostic suggestions and made it equal for both scenarios. In addition, physicians were able to request further information about the patient by choosing from a list of clinical cues provided. Finally, diagnostic suggestions were always provided in the scenario seen last, to avoid transferring their effect (if any) to the other scenario. The rest of the adjustments are discussed below.

In addition to physicians' certainty and frequency of diagnostic changes, we also looked into their information search (number of cues selected) before providing a final diagnosis, and information interpretation, namely, the way they evaluated neutral clinical cues as more or less supportive of their initial diagnostic hypothesis.

3.2. Research questions and objectives

In the present study, we aimed to explore two further psychological mechanisms that could explain the effect of early diagnostic suggestions: information search and information evaluation. Specifically, the study aimed to test whether the provision of diagnostic suggestions can lead GPs to obtain more patient information (cues) before reaching a diagnostic conclusion, and to interpret neutral information as less supportive of their own diagnosis. The neutral patient information could not help to differentiate between two or more competing diagnoses. It also aimed to seek further evidence that diagnostic suggestions can make GPs less certain in their diagnostic judgements, as partly found in Study 1. Finally, it aimed to test whether the effects of diagnostic suggestions depend on GPs' gender, experience and/or thinking style.

3.3. Hypotheses

We hypothesised that when GPs receive diagnostic suggestions, they would:

H_1: request more information about the patient before providing a final diagnosis,

H_2: rate the neutral patient information (cues) as less supportive of their initial diagnosis[13],

H_3: increase their certainty in their initial diagnosis less after acquiring the additional patient information, and

H_4: change their initial diagnosis more frequently than when they do not receive diagnostic suggestions.

[13] This is the main hypothesis. It has been listed second due to the temporal order of the task: GPs first sought (H1) and then evaluated (H2) the neutral patient information.

3.4. Methods

3.4.1. Participants

Participants were fully qualified GPs and GP trainees, currently practising in England.

3.4.2. Sample size

We powered the study for the main hypothesis that when GPs receive diagnostic suggestions, they will evaluate neutral information as less supportive of their initial diagnosis (H2). In previous studies where GPs diagnosed different clinical scenarios, Kostopoulou and colleagues measured very small intraclass correlations (less than 0.01; Woolley, Kostopoulou and Delaney (2016), Kostopoulou, Porat and colleagues (2017)). We therefore assumed independence of responses. For the sample size calculation, we used the G*Power software (version 3.1.9.4). For a multiple linear regression (fixed model, R^2 deviation from 0) with two predictors (i.e., experimental condition and severity of initial diagnosis), we would need 485 responses to detect a small effect ($f^2 = 0.02$) with 80% power and $\alpha = 0.05$. Thus, we would need 243 participants, each responding to two scenarios. We increased this to 250, assuming that not everyone would provide complete and analysable responses.

3.4.3. Recruitment

Participants were recruited through the National Institute for Health and Care Research (NIHR) Clinical Research Network (CRN) of North West London. We decided to involve CRN North West London (www.local.nihr.ac.uk/lcrn/north-west-london) for recruitment for two reasons. Firstly, we did not want to invite the same GPs who took part in Study 1, using the database. Secondly, we wanted to ensure that we could reach the required sample size (i.e., 250 GPs). Without CRN's support, this would not be feasible.

CRN representatives distributed an invitation e-mail (Appendix 6) to potential participants. The invitation e-mail contained information about the study and a link to the study website. GPs were offered a £10 Amazon voucher as a token of appreciation for their participation. This value was determined based on the estimated time to complete the study (i.e., approximately 15 minutes). Recruitment was closed as soon as the required sample size was reached.

The study was approved by the Health Research Authority (REC: 19/HRA/5931). It was also pre-registered on the Open Science Framework (OSF) and is available at https://osf.io/4ksb7.

3.4.4. Materials

Two patient scenarios, previously developed by Kostopoulou and colleagues (2009, 2017), were adapted for the purposes of this study. We constructed two versions of each scenario, suggesting two different diagnoses. In one version, both patient characteristics and the initial information suggested a benign and common diagnosis, while in the other version, they suggested a serious and less common diagnosis. We will refer to this initial information as the “steer”. Each scenario also contained some neutral information (identical in both versions), which did not help to differentiate between the two diagnoses (Figures 3.1. and 3.2.). The AOT scale used in Study 1 was also used in the present study.

<table>
<tr><th colspan="2">Scenario 1: Chest pain</th></tr>
<tr><th>Version 1: Cardiac steer</th><th>Version 2: Musculoskeletal steer</th></tr>
<tr><td>Patient name: David H.

Age: 61 BMI: 27
Smoking history: Ex-smoker
Past medical history: Osteoarthritis of the knee (2016)
Last consultation: URTI a year ago.

David says: "Hello doctor. I've been getting a pain in my chest for a week now. It gets worse when I work in my garden. Nothing else seems to affect it. I am a keen gardener, but I don't do any other exercise. I don't walk fast and avoid using stairs as I get knee pain. This pain in my chest usually goes away when I take a rest, doctor, but I thought it'd be good to come and see you."</td><td>Patient name: David H.

Age: 51 BMI: 23
Smoking history: Never smoked
Past medical history: Osteoarthritis of the knee (2016)
Last consultation: URTI a year ago.

David says: "Hello doctor. I've been getting a pain in my chest for a week now. I was helping my daughter move house and when I lifted the washing machine, I felt the pain coming on. I must have pulled a muscle in my chest, but my wife got a bit worried and asked me to come and see you."</td></tr>
<tr><th colspan="2">Additional neutral information</th></tr>
<tr><td colspan="2">Resting ECG results: It is normal.
Self-treatment: David says that he has not taken any pain relief so far. He does not like taking tablets.
General physical examination: David's respiratory rate is 16 breaths per minute. His pulse is 80 beats per minute, regular and with a normal character. There is no wheeze in the chest and his heart sounds are normal. There is no peripheral oedema and the JVP is not raised.
Family history of any significant illness: There has not been any significant illness in David's family, but his uncle had a myocardial infarction at age 70.
Pain intensity: David says that it is not really intense, but it does make him want to stop what he is doing.
Chest pain during the night: David says that it occasionally disturbs him at night.
Other symptoms: David does not report any other symptoms apart from the chest pain.</td></tr>
</table>

Figure 3.1. Chest pain scenario.

Version 1: Based on the initial description, most GPs were expected to think of a cardiac chest pain (i.e., angina), given the patient's risk factors (age, smoking history, BMI) and the fact that the pain stops when resting. **Version 2:** Based on the initial description, GPs were expected to consider musculoskeletal chest pain as the most likely diagnosis, given that the patient has no risk factors and the pain started after a plausible chest injury (lifting a washing machine). **Additional information:** Based on this information, GPs were expected not to change their initial diagnosis, given that this information had minimal diagnostic value and did not help to differentiate between the possible diagnoses.

Scenario 2: Constipation	
Version 1: Cancer steer	**Version 2: IBS steer**
Patient name: Barbara H. **Age:** 61 **BMI:** 23.5 **Smoking history**: Never smoked **Past medical history:** IBS diagnosis in 2015. Colonoscopy and coeliac screen normal. **Last consultation:** Seven months ago, about loose stools, abdominal pain, and bloating. You ordered a full blood count. The results were: Hb 11.2 g/dL* (12-15.5) MCV 73 fL* (76-100) WCC 8.6 x 10^9 (4.0-11.0) Neutrophils 6 x 10^9 (2.0-7.5) Lymphocytes 2.1 x 10^9 (1.3-3.5) Monocytes 0.3 x 10^9 (0.2-0.8) Eosinophils 0.2 x 10^9 (0-0.4) Platelets 280 x 10^9 (150-450) Barbara was advised to get in touch about her results, but she did not. **Barbara says**: *"Hello doctor. I'm having trouble with my bowels again. Not unusual for me, as you know. This time it's constipation. I've had hard stools for six weeks now, and it's very uncomfortable."*	**Patient name: Barbara H.** **Age:** 61 **BMI:** 23.5 **Smoking history**: Never smoked **Past medical history:** IBS diagnosis in 2015. Colonoscopy and coeliac screen normal. **Last consultation:** Seven months ago, about loose stools, abdominal pain, and bloating. You ordered a full blood count. The results were: Hb 13.6 g/dL (12-15.5) MCV 88 fL (76-100) WCC 8.7 x 10^9 (4.0-11.0) Neutrophils 6.2 x 10^9 (2.0-7.5) Lymphocytes 2.8 x 10^9 (1.3-3.5) Monocytes 0.5 x 10^9 (0.2-0.8) Eosinophils 0.4 x 10^9 (0-0.4) Platelets 356 x 10^9 (150-400) **Barbara says**: *"Hello doctor. I'm having trouble with my bowels again. Not unusual for me, as you know. This time it's constipation. I've had hard stools for six weeks now, and it's very uncomfortable. Other than that, I feel OK. I don't feel tired or anything. Sometimes my tummy gets a bit painful, but it feels better after I've been to the loo."*
Additional neutral information	
• **Mucus in stools:** Barbara occasionally has mucus in her stools. • **Blood in stools:** Barbara has not had any blood in her stools. • **Sensation of bowel emptying:** Barbara reports that after she passed stools, she feels that she has emptied her bowel completely. • **Ferritin levels (current):** The result is 58µg/L (27-330). • **Digital rectal examination results:** The perianal region feels normal. The sphincter tone is normal. You feel no lumps, blood or faeces. • **Previous episodes of bowel problems:** Barbara occasionally experiences episodes of either loose or hard stools, which may last for a few weeks. Her last bout of bowel trouble was seven months ago, when she had loose stools for a while. • **Family history of any significant illness:** There is no family history of any significant illness.	

Figure 3.2. Constipation scenario.

Version 1: At this stage, GPs were expected to consider colorectal cancer as the most likely diagnosis for Barbara, given her low level of haemoglobin. **Version 2**: Given this initial description, GPs were expected to consider IBS as the most likely diagnosis. **Additional information**: Based on this information, GPs were expected not to change their initial diagnosis.

3.4.5. Design

The design of the present study was arrived at via a highly iterative process: we explored (and piloted) several research questions, study designs, patient cases, and procedures; a full account is available in Appendix 8.

The study employed a within-participant design with one independent variable ("Condition": decision aid versus no aid)[14]. Each participant saw one version of each patient scenario (the version seen was allocated at random). The two scenarios were presented in a random order, but the aid (diagnostic suggestions) was always presented in the scenario seen second. This was to preclude the experimental condition (aided) from influencing GPs' reasoning in the control condition (unaided). The software randomiser in Qualtrics ensured that scenario versions (i.e., the steers) were evenly presented across conditions, so that each scenario version was seen with and without the diagnostic suggestions an equal number of times.

The scenario versions contained diagnostic information to steer participants towards a diagnosis, so that they quickly develop a leaning that would potentially influence the perceived support of the subsequent information for their diagnostic hypothesis, while ensuring that the effects were not diagnosis-specific.

3.4.6. Procedure

As in Study 1, the patient scenarios were administered using Qualtrics (Qualtrics.com). Potential participants were sent an invitation e-mail, containing a brief description of the study, as well as a hyperlink to access the study site. Upon accessing the site, participants read the information sheet and indicated their consent (Appendices 4 and 5 respectively). They were then asked to indicate their gender (male, female or other) and their professional status (fully qualified GP or GP trainee). Fully qualified GPs were asked to provide their year of GP training completion.

[14] We opted for a within-participant design to keep it consistent with Study 1. Also, a between-participants design would require twice the number of participants. It would not be practical to recruit such a high number of GPs.

GPs were then presented with two patient scenarios in a random order. The first scenario was assessed without diagnostic suggestions (Control condition) and the second always with suggestions (Aided condition). Each scenario was presented in two steps. Initially, GPs read a short patient description (age, BMI, smoking and medical history, last consultation and presenting problem). They were then asked to provide their initial diagnosis in free text and indicate their level of certainty in their diagnosis (Figure 3.3.)[15].

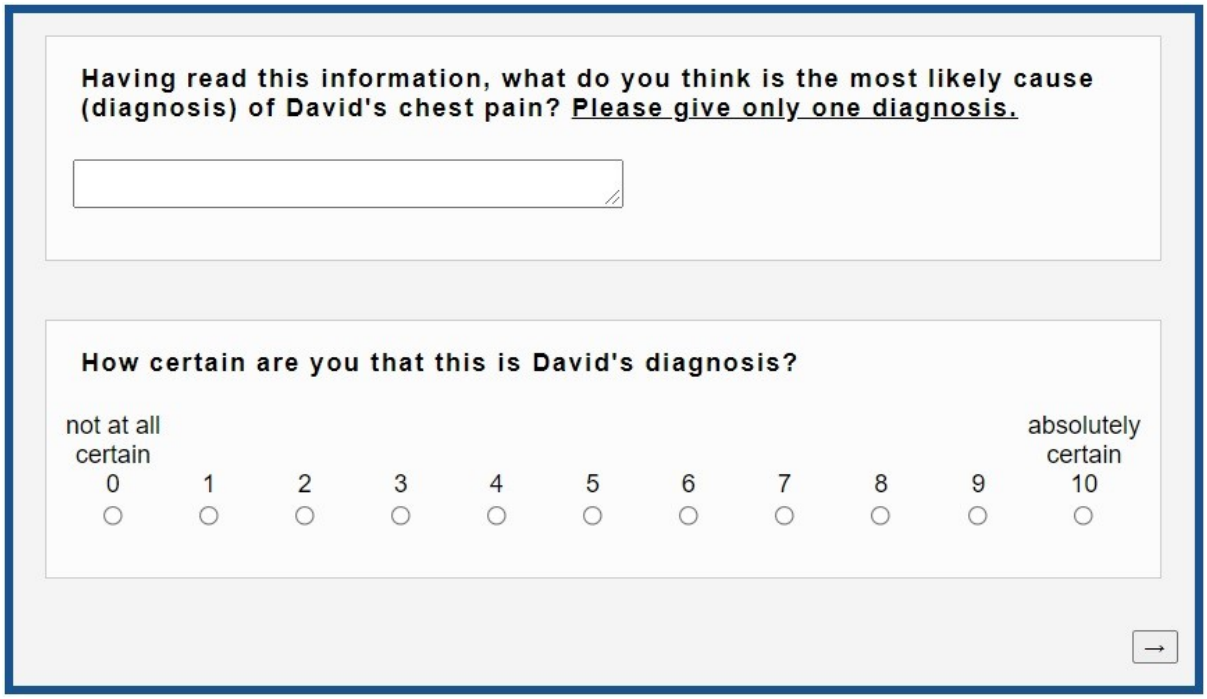


Figure 3.3. Initial diagnosis and certainty scale (chest pain scenario).

Subsequently, and in the Aided condition only, GPs were presented with a list of 12 diagnostic suggestions (Figure 3.4.). All GPs were then given the opportunity to request additional items of information about the patient, from a list (Figure 3.5.). The list comprised seven items (corresponding to the "neutral information" in Figures 3.1. and 3.2.) and GPs could request as many of these as they wished. Each time that a GP requested an item (or "cue"), the answer was revealed on the next screen (Figure 3.6.). After requesting a cue and reading the relevant information, GPs were asked to rate how much the cue supported their initial diagnosis.

[15] The same scale was used in Study 1.

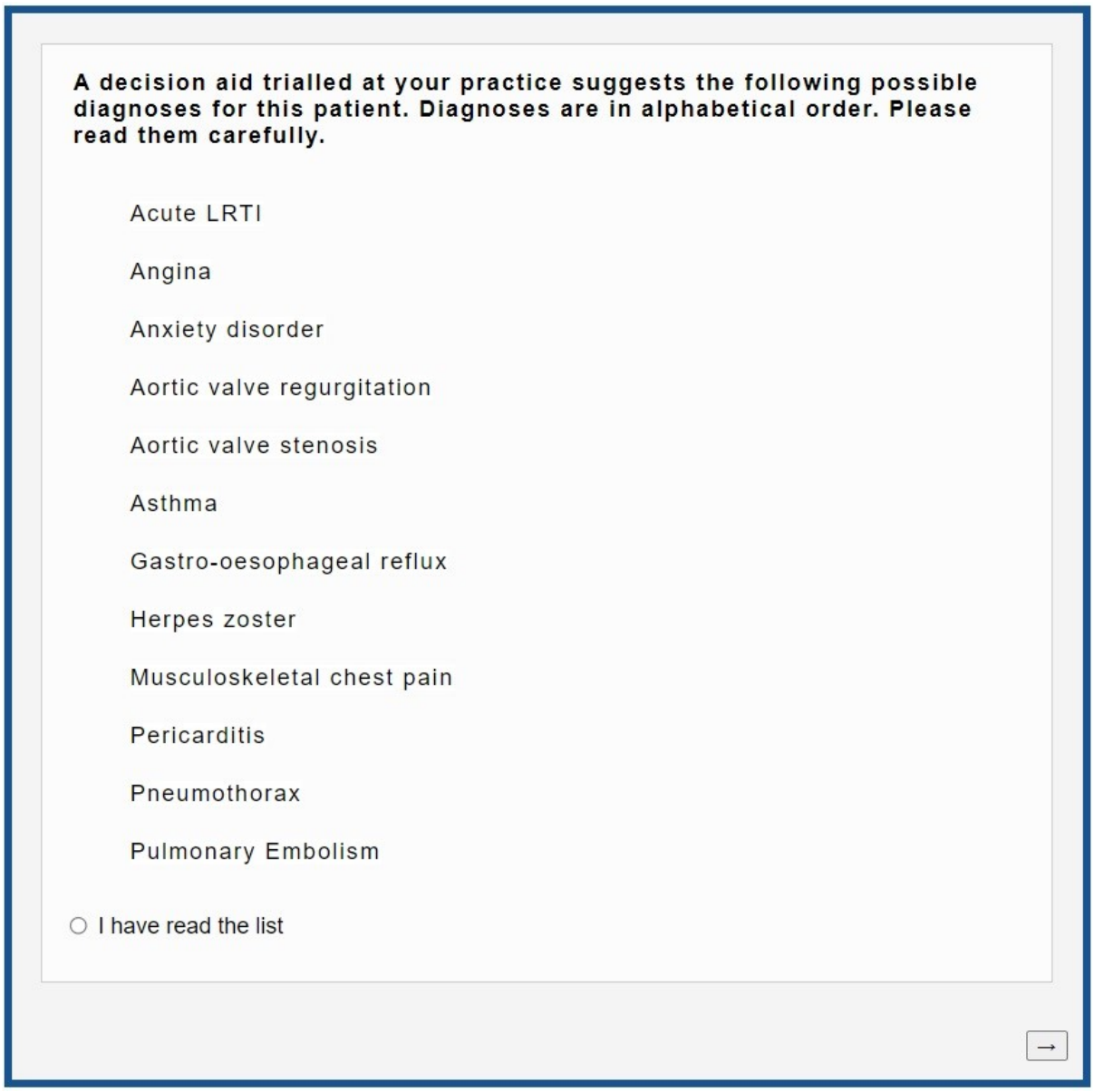


Figure 3.4. The decision aid (chest pain scenario).

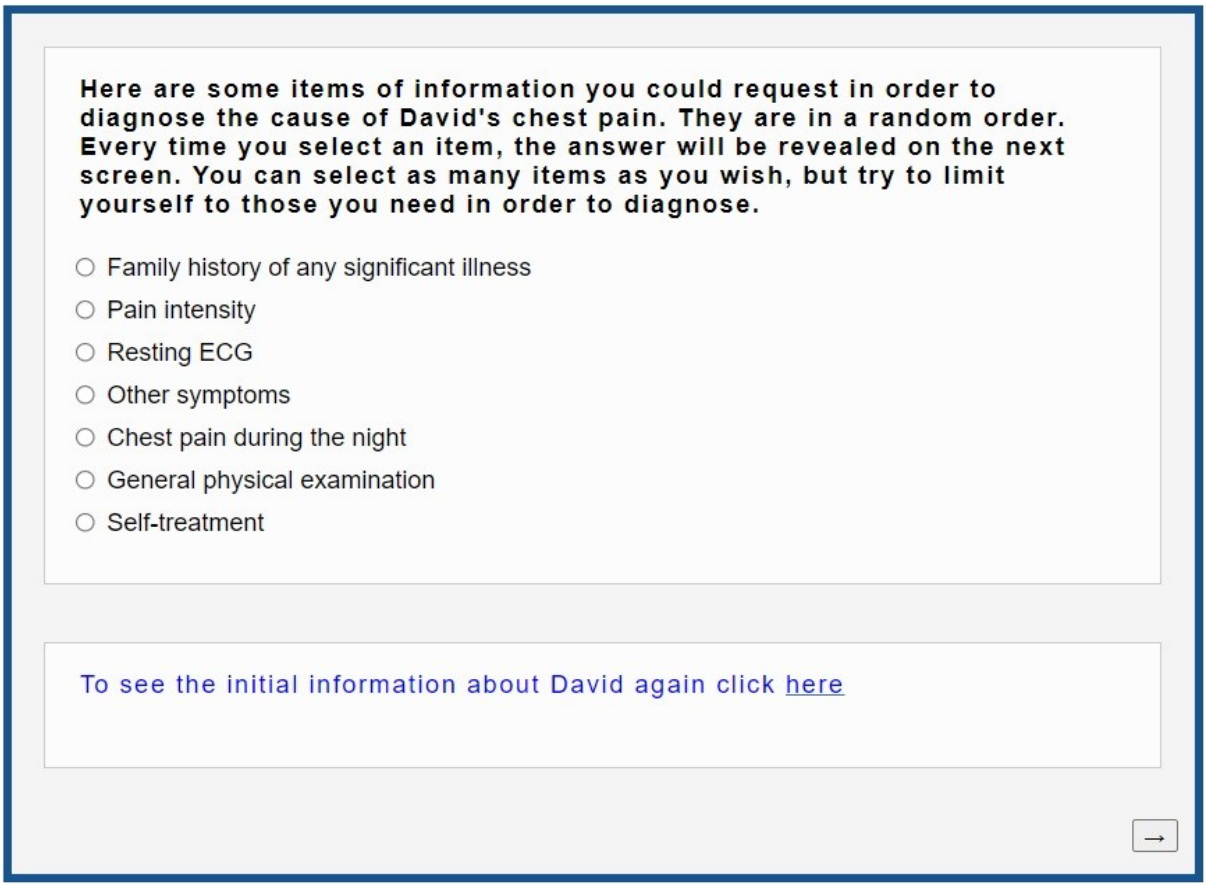


Figure 3.5. Additional cues (chest pain scenario).

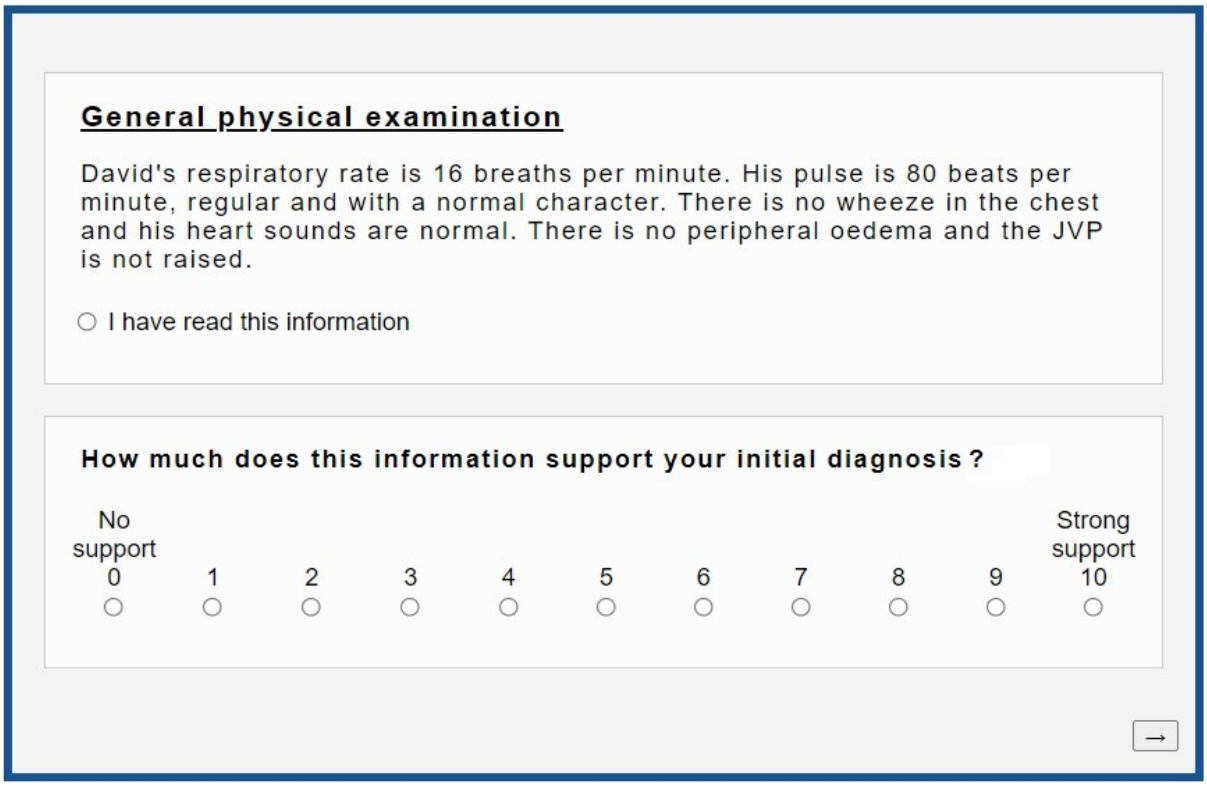


Figure 3.6. "General physical examination" cue information (chest pain scenario).

After each cue evaluation, GPs were asked to indicate whether they wished to a) provide their final diagnosis (“Have you settled on a diagnosis at this stage?”) or b) request more information about the patient. When GPs opted to request more information, they were offered the remaining cues and invited to make a selection. When GPs indicated that they had settled on a diagnosis – or when they had requested all of the available cues – they were asked to a) indicate their certainty in their initial diagnosis, b) provide their final diagnosis (in free text), and c) rate their certainty in their final diagnosis (Figure 3.7.). They were given the opportunity to review the patient case (the initial patient description plus any cues that they had requested) before providing these responses (“Before you provide your final diagnosis, would you like to see the patient information again?”). They were also asked to list the differential diagnoses that they were considering, if any.

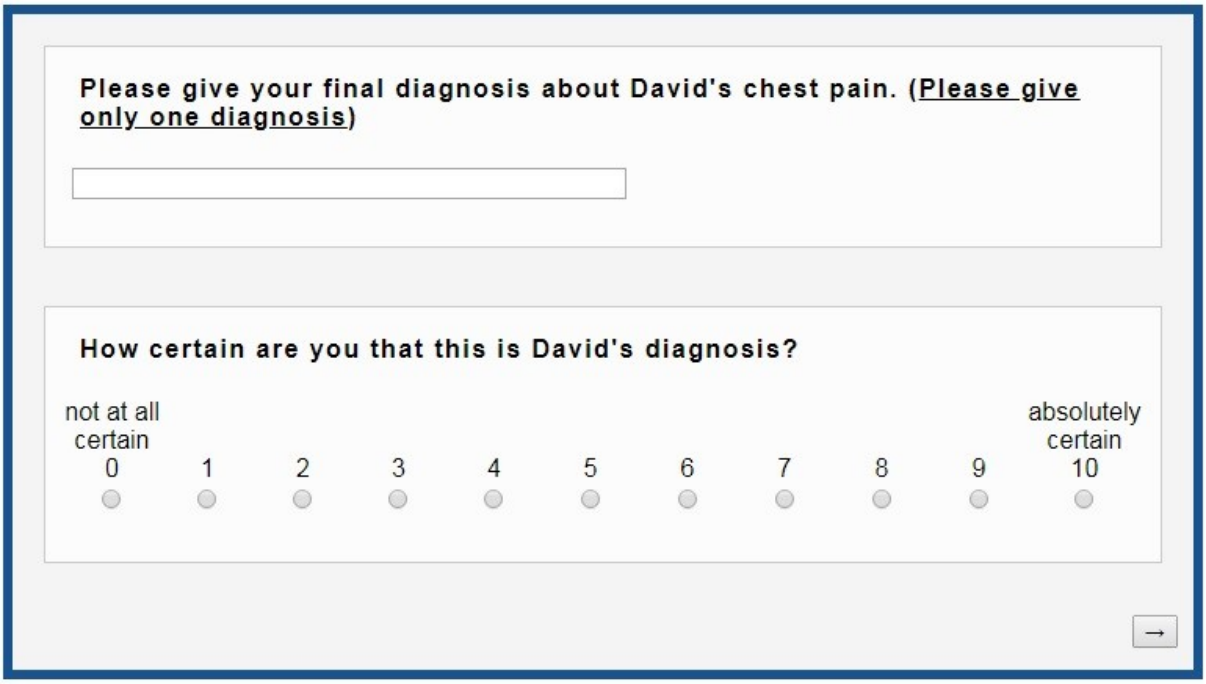


Figure 3.7. Initial diagnosis and certainty scale (chest pain scenario).

At the end of each patient scenario, GPs were asked how they would manage the patient. There were provided with five different options and could select as many as they wished (Figure 3.8.).

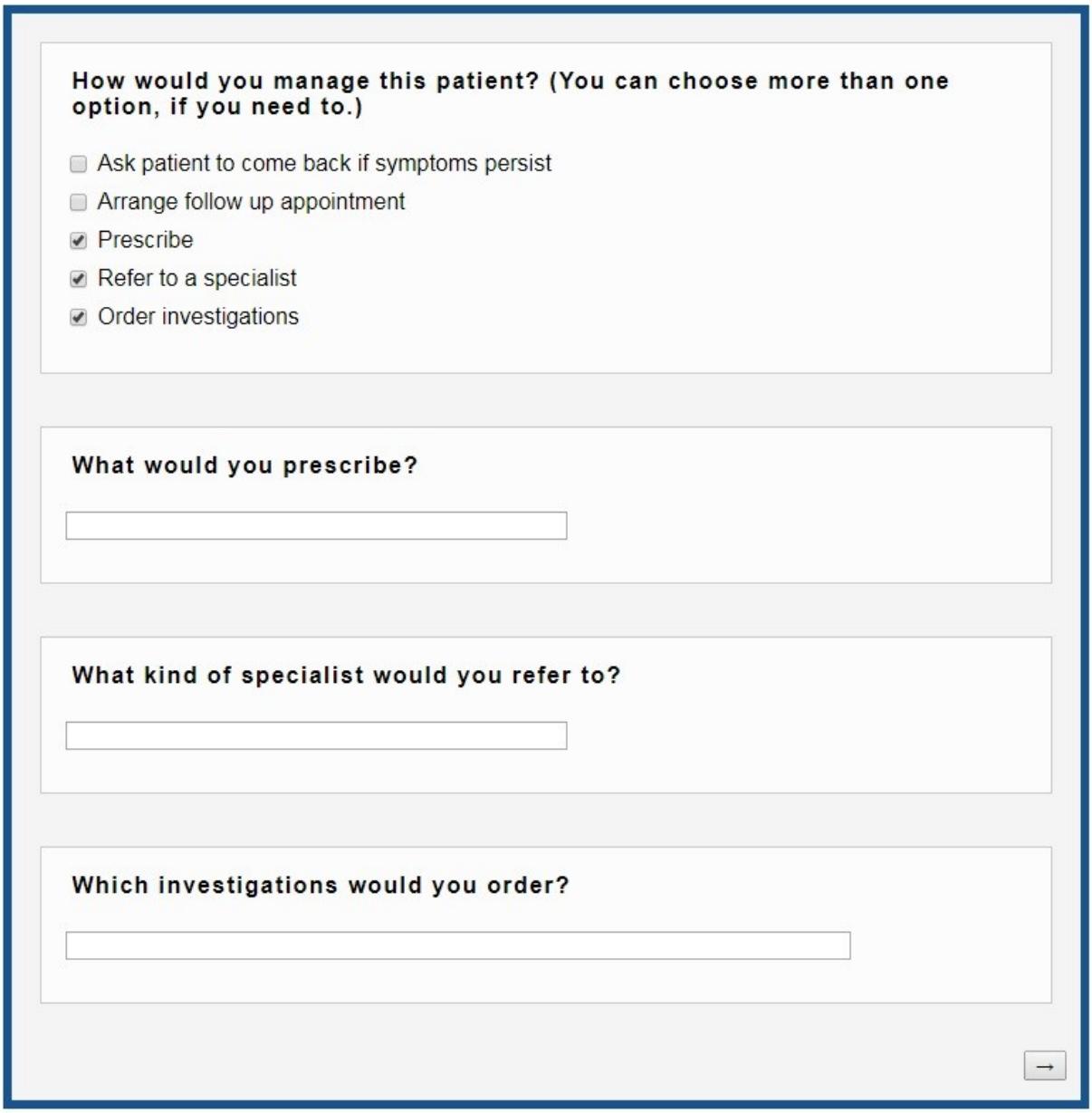


Figure 3.8. Patient management.

After both scenarios were completed, participants were asked to complete the Actively Open-Minded Thinking (AOT) scale (Baron, 2019). Upon study completion, participants were offered a £10 Amazon voucher for their time. We estimated that it would take on average 15 minutes to complete the study. The study procedure is graphically presented below (Figure 3.9.) and can also be found in Appendix 7.

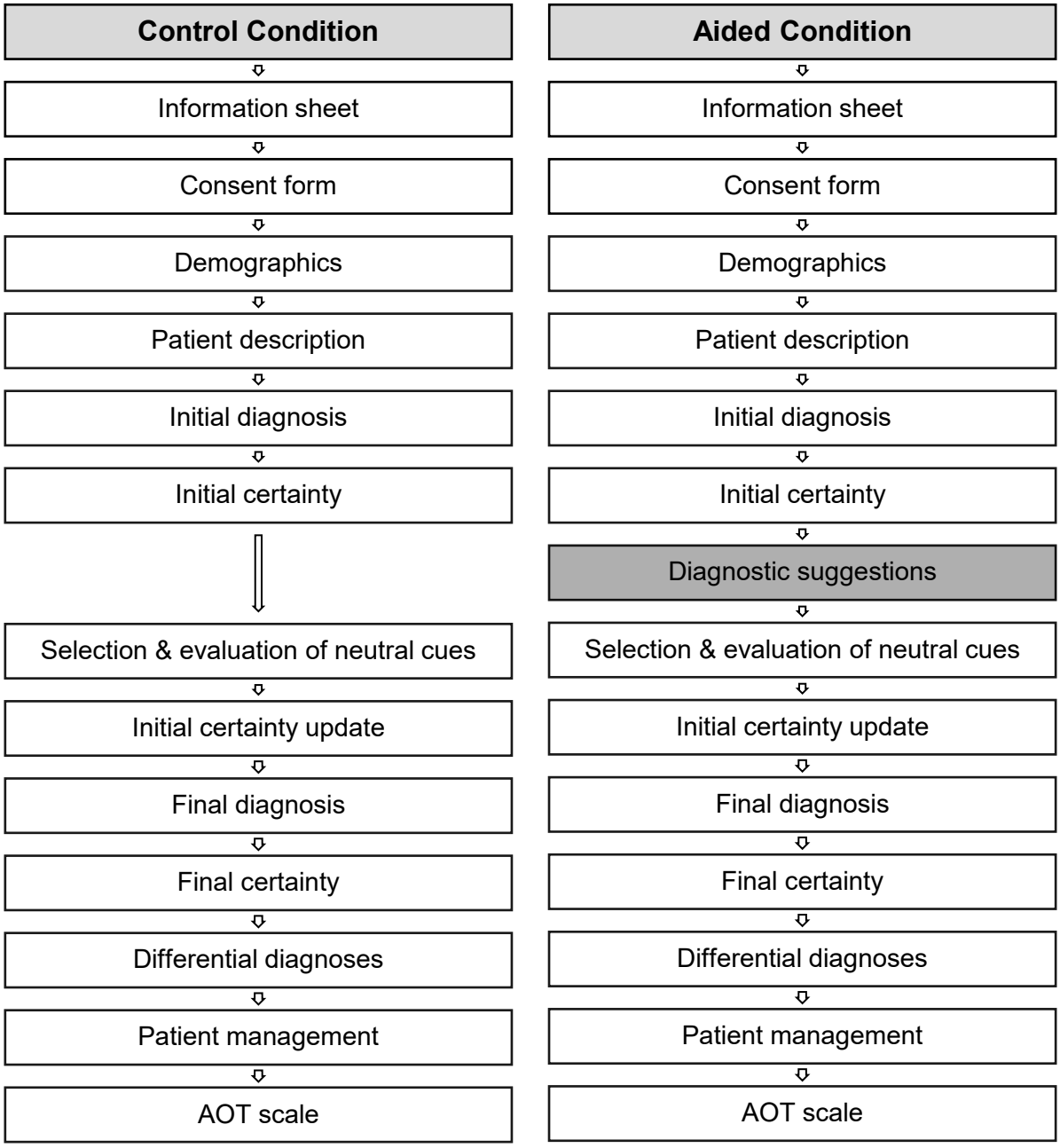


Figure 3.9. Flow of activity.

3.5. Analyses

3.5.1. Dependent variables

Based on the hypotheses above (section 3.3.), the dependent variables were:

1. **Initial certainty**, a continuous variable measured on a 0-10 scale.
2. **Number of cues** requested (H1), a count variable.

3. **Perceived cue support** (H2), i.e., the average support rating of all cues requested, a continuous variable measured on a 0-10 scale.
4. **Change in initial certainty** after obtaining additional cues (H3), a continuous variable computed by subtracting the second measurement of certainty from the first.
5. **Change in diagnosis** (H4), a dichotomous variable indicating whether the initial and the final diagnoses differed, coded as 0 for no change and 1 for change.
6. **Final certainty**, a continuous variable measured on a 0-10 scale.
7. **AOT score**, a continuous variable measured on a 5-point scale with verbal anchors from completely disagree (1) to completely agree (5).
8. **Number of differential diagnoses,** a count variable.
9. **Appropriateness of management,** a dichotomous variable indicating whether the management options were appropriate for the patient, coded as 0 for not appropriate management and 1 for appropriate management[16].

3.5.2. Independent variables

The following independent variables were used in the analyses:

1. **Condition**, coded as 0 for Control and 1 for Aided.
2. **Severity of initial diagnosis**, coded as 0 for the less serious diagnoses (i.e., musculoskeletal and IBS) and 1 for the more serious diagnoses (i.e., cardiac and cancer).
3. **GPs' gender**, coded as 0 for females, 1 for males and 2 for 'other'.
4. **Years of experience**, a continuous variable computed based on the GPs' year of qualification.

[16] Appropriateness of patient management was evaluated according to suggestions of Dr Kavleen Arora, GP and clinical research fellow in the Decision Making research group at Imperial College London.

3.5.3. Preparing the data for analysis

The independent variable “Condition”, which was the main predictor, could take two values: 0 for Control (without diagnostic suggestions) and 1 for Aided (with diagnostic suggestions). To make the diagnoses more easily analysable, some responses were simplified. Specifically, in the chest pain scenario (David), diagnoses related to cardiac causes (e.g., angina, pericarditis, ischaemic heart disease) were classified under “cardiac”. Likewise, diagnoses related to musculoskeletal causes (e.g., costochondritis, musculoskeletal chest pain, pulled muscle) were classified under “MSK”. In the constipation scenario (Barbara), diagnoses related to cancer (e.g., colon cancer, bowel cancer, lower gastrointestinal malignancy) were classified under “cancer”. IBS diagnosis stood on its own as a category. Likewise, the rest of the diagnoses were not classified into a category but used as they were listed (e.g., diverticular disease, stricture). These classifications were based on advice from the clinical supervisor (Professor Brendan Delaney).

We also measured changes in diagnoses within each patient scenario. If GPs’ initial diagnosis was different from their final diagnosis, then this would count as a diagnostic change. In addition to this criterion, if GPs provided a single initial diagnosis and multiple final diagnoses, then this would also count as a change.

3.5.4. Statistical tests

In separate regression models with random intercept (linear for continuous dependent variables, logistic for dichotomous dependent variables and Poisson for count variables), we regressed the dependent variables on the main predictor only (condition: diagnostic suggestions versus no suggestions), and subsequently on the second predictor (severity: serious versus less serious initial diagnoses). In all regression models, we subsequently included initial certainty as a covariate. AOT scores, GP experience (covariates) and gender (explanatory variable) were finally included (in one step) in all regression models.

3.5.5. Data exclusion

We excluded GPs that did not provide an initial diagnosis and those that provided a non-analysable response (e.g., 'uncertain') or multiple initial diagnoses. The latter were excluded because we could not be certain if they were evaluating the neutral cues in relation to a single initial diagnosis and which specific diagnosis out of those they gave. We also excluded GPs that completed only one scenario or provided incomplete data (e.g., closed the browser without finishing a scenario).

3.6. Results

273 GPs were recruited. 15 of them did not complete the task and their responses were excluded. Another 10 GPs were excluded due to non-analysable responses (e.g., multiple initial diagnoses). The remaining 248 GPs' responses were included in the analysis. Of those 248 GPs, 115 were male (46.40%) and 113 were female (53.60%). There were 223 fully-qualified GPs (89.90%), the remaining being trainees in general practice. The average clinical experience post-qualification was 10.44 years (SD 9.06), ranging from 0 to 44 years (median 9 years). Participation lasted on average 16 minutes. Table 3.1. presents the mean and SD for each dependent variable, per condition (Control and Aided).

Dependent variable	Control condition Mean (SD)	Aided condition Mean (SD)
Initial certainty	6.11 (1.87)	6.11 (1.79)
Second measurement of certainty	6.50 (2.30)	6.53 (2.05)
Perceived support of neutral cues	5.46 (2.54)	5.82 (2.00)
Number of cues requested	4.16 (2.06)	4.37 (1.87)
Change in initial certainty	0.40 (2.16)	0.42 (1.87)
Change in diagnosis	19.00%	15.70%
Certainty about final diagnosis	7.06 (1.58)	6.73 (1.80)

Number of differential diagnoses	1.32 (1.18)	1.25 (1.08)
Appropriateness of management	8.00%	13.00%

Table 3.1. Results by condition (with and without the decision aid).
Initial certainty refers to GPs' certainty about their initial diagnosis. The second measurement of certainty refers to the certainty measured after GPs received additional patient information and before they provided their final diagnosis. Change in initial certainty refers to the difference between the second and the first measurement of initial certainty and was positive in both conditions (increase).

3.6.1. Initial and final diagnoses

Since the initial patient description in each scenario version (the steer) was strongly suggestive, most GPs were expected to provide the steer diagnosis initially. On average, GPs provided the expected diagnosis in 78.83% of the cases (Chest pain scenario: Musculoskeletal version = 84.68%, Cardiac version = 97.58%, Constipation scenario: IBS version = 70.97%, Colorectal cancer version = 62.10%). In the chest pain scenario, most participants gave either a musculoskeletal or a cardiac diagnosis, while in the constipation scenario there was greater variability between responses. Table 3.2. presents the diagnoses that GPs provided in each patient scenario after the brief patient history and presenting complaint.

Scenario	Version	Diagnosis	Proportion (%)
Chest pain scenario	**Musculoskeletal chest pain**	Musculoskeletal chest pain	84.70
		Cardiac-related diagnosis	15.30
	Cardiac chest pain	Musculoskeletal chest pain	1.60
		Cardiac-related diagnosis	97.60
		Gastro-oesophageal reflux disease	0.80
Constipation scenario	**Irritable bowel syndrome (IBS)**	Irritable bowel syndrome	71.00
		Cancer	18.50
		Simple constipation	7.30
		Nutritional disease	1.60

		Diverticular disease	0.80
		Stricture	0.80
	Colorectal cancer	Irritable bowel syndrome	25.80
		Cancer	62.10
		Simple constipation	3.20
		Nutritional disease	0.80
		Diverticular disease	4.80
		Haemorrhoid	0.80
		Dehydration	0.80
		Inflammatory bowel disease (IBD)	0.80
		Side effect of iron replacement therapy	0.80

Table 3.2. Proportion of initial diagnoses across patient scenarios.

As previously mentioned, GPs were not expected to change diagnosis. In line with our hypothesis, most GPs maintained their initial diagnosis throughout the task (82.40%, 337/409). As with the initial diagnosis, most participants gave either a musculoskeletal or a cardiac diagnosis at the end of the chest pain scenario, while in the constipation scenario, there was more variability amongst the final diagnoses. Table 3.3. presents GPs' final diagnoses across patient scenarios:

Scenario	Version	Diagnosis	Proportion (%)
Chest pain scenario	**Muscular pain**	MSK	79.80
		Cardiac diagnosis	16.90
		Non-cardiac diagnosis	2.40
		Multiple diagnoses	0.80
	Cardiac pain	MSK	5.60
		Cardiac diagnosis	91.10

		Gastroesophageal reflux disease (GORD)	1.60
		GORD or Cardiac	0.80
		Asthma	0.80
Constipation scenario	**Irritable bowel syndrome (IBS)**	IBS	73.40
		Cancer	14.50
		Simple constipation	3.20
		IBD and constipation	0.80
		Diverticular disease	4.80
		IBD	0.80
		IBS or cancer	0.80
		Diet	0.80
		Stricture	0.80
	Colorectal cancer	IBS	43.50
		Cancer	43.50
		IBD	2.40
		IBS or IBD	0.80
		Diverticular disease	5.60
		Simple constipation	1.60
		Haemorrhoid	0.80
		Uncertain	1.60

Table 3.3. Final diagnoses across patient scenarios.

3.6.2. Number of cues requested (H_1)

On average, GPs requested 4.26 cues per scenario (SD 1.97). In line with our hypothesis, GPs requested more cues when they received diagnostic suggestions than when they did not (4.37 versus 4.16). To test whether this difference was

significant, we regressed the dependent variable (number of cues requested) on the main predictor (condition), using simple Poisson regression. The results showed that this difference was not significant (IRR = 0.95, [0.87, 1.04], p = 0.250).

We subsequently extended the model by regressing the dependent variable on the second predictor (severity of the initial diagnosis) and found that there was a significant association (IRR = 1.162, [1.06, 1.27], p = 0.001). This finding seems somewhat counterintuitive, as it suggests that GPs who provided a serious and uncommon diagnosis (e.g., angina, bowel cancer) assessed less information about the patient before reaching a diagnostic conclusion than those who provided a common diagnosis (*cf*. Kostopoulou, Mousoulis and Delaney, 2009). Assuming that GPs would be more familiar with common clinical cases than uncommon ones, one would expect that they would seek more information about the latter. However, it is also possible that GPs who provided a common diagnosis engaged in a more extensive search for information, in an attempt to exclude the possibility of a more serious disease for the sake of safety. There was no significant interaction between condition and severity.

To test the effects of initial certainty on the number of cues requested, we added initial certainty to the model. It was found to be significantly associated with the number of cues requested, suggesting that GPs with higher initial certainty accessed a smaller amount of patient information before reaching a conclusion (IRR = 0.916, [0.89, 0.94], p < 0.001). There was no significant interaction between initial certainty and condition or severity.

Poisson regressions of number of cues requested

Step		Coefficient	Std. Error	t	Sig.	95% CI		Exp(Coefficient)	95% CI for Exp(Coefficient)	
						Lower	Upper		Lower	Upper
1	Intercept	1.470	.0330	44.50	.000	1.405	1.535	4.348	4.075	4.639
	Condition	-.050	.0435	-1.152	.250	-.136	.035	.951	.873	1.036
2	Intercept	1.393	.0411	33.92	.000	1.312	1.473	4.026	3.714	4.364
	Condition	-.050	.0435	-1.157	.248	-.136	.035	.951	.873	1.036
	Severity	.150	.0466	3.225	.001	.059	.242	1.162	1.060	1.274

3	Intercept	1.949	.0880	22.15	.000	1.776	2.122	7.021	5.907	8.346
	Condition	-.053	.0435	-1.219	.223	-.139	.032	.948	.871	1.033
	Severity	.095	.0460	2.072	.039	.005	.186	1.100	1.005	1.204
	Certainty 1	-.088	.0127	-6.933	.000	-.113	-.063	.916	.893	.939
4	Intercept	.653	.2415	2.704	.007	.178	1.127	1.921	1.195	3.088
	Condition	-.051	.0434	-1.168	.243	-.136	.035	.951	.873	1.035
	Severity	.053	.0446	1.180	.238	-.035	.140	1.054	.966	1.151
	Certainty 1	-.083	.0121	-6.884	.000	-.107	-.059	.920	.899	.942
	Gender	.072	.0440	1.629	.104	-.015	.158	1.074	.985	1.171
	Experience	.002	.0024	.823	.411	-.003	.007	1.002	.997	1.007
	AOT score	.314	.0572	5.486	.000	.201	.426	1.368	1.223	1.531
Dependent Variable: Number of cues requested										

Table 3.4. Regression models of number of cues requested.

The scenario version in which GPs requested fewer cues was the cardiac steer of the chest pain scenario. This appeared to be the most straightforward scenario version, where GP provided the expected diagnosis in 97.60% of instances and had the highest initial and final certainty (i.e., 7.15 and 7.36 respectively). This scenario version also produced the smallest change in initial certainty (-0.06), the lowest rate of diagnostic change (6.45%, 8/124) and the lowest number of differential diagnoses – i.e., approximately one diagnosis. Table 3.5. presents the number of cues that GPs requested in each version of the patient scenarios:

Patient scenario	Control condition Mean (SD)	Aided condition Mean (SD)	Overall Mean (SD)
Chest pain (musculoskeletal steer)	4.76 (2.03)	4.61 (1.92)	4.69 (1.97)
Chest pain (cardiac steer)	3.53 (2.19)	3.85 (1.83)	3.69 (2.02)
Constipation (IBS steer)	3.97 (1.76)	4.66 (1.73)	4.31 (1.77)
Constipation (cancer steer)	4.37 (2.06)	4.35 (1.93)	4.36 (1.99)

Overall	4.16 (2.06)	4.37 (1.87)	4.26 (1.97)

Table 3.5. Number of cues requested across patient scenarios.

In the chest pain scenario, the three most requested cues for both steers were “General physical examination” (MSK steer: 86.30%, Cardiac steer: 62.10%), “Resting ECG results” (MSK steer: 75%, Cardiac steer: 73.40%) and “Other symptoms” (MSK steer: 77.40%, Cardiac steer: 66.10%). In the constipation scenario, the most requested cues for both steers were “Blood in stools” (IBS steer: 85.50%, cancer steer: 83.90%), “Digital rectal examination” (IBS steer: 79.80%, cancer steer: 82.30%) and “Ferritin levels” (IBS steer: 67.70%, cancer steer: 75%).

3.6.3. Perceived cue support (H_2)

Across scenarios and conditions, the average perceived cue support rating was 6.11 (SD 1.97). In contrast to our hypothesis, GPs who received diagnostic suggestions rated the cues as more supportive of their initial diagnosis (5.82 versus 5.46, $p = 0.037$), as the simple linear regression model revealed. When severity of initial diagnosis and initial certainty were added to the regression model sequentially, both were found to be significantly associated with perceived cue support ($p < 0.001$, see Table 3.6.): GPs with higher initial certainty, as well as GPs who initially provided common and less serious diagnoses (i.e., MSK, IBS), rated the cues as more supportive of their initial diagnosis. However, as shown in Table 3.6., there was a significant interaction between severity and condition ($b = 0.753, 0.04, 1.46$, $p = 0.038$), suggesting that the effect of severity differed by condition (i.e., condition did not produce any differences in perceived cue support for less serious diagnoses whereas for serious diagnoses, diagnostic suggestions resulted in higher rates). This is not toward the expected diagnosis, given that we hypothesised that the provision of diagnostic suggestions would make GPs less biased in their interpretations irrespective of the severity of initial diagnosis.

Linear regression of perceived cue support							
Step		Estimate	Std. Error	t	Sig.	95% CI	
						Lower Bound	Upper Bound
1	Intercept	5.458669	.145121	37.615	.000	5.173539	5.743800
	Condition	.358596	.171448	2.092	.037	0.020910	.696283
2	Intercept	6.486556	.185037	35.055	.000	6.122995	6.850117
	Condition	-.012338	.244129	-.051	.960	-.492448	.467771
	Severity	-2.055773	.257456	-7.985	.000	-2.561701	-1.549845
	Condition*Severity	.752381	.362268	2.077	.038	.040449	1.464313
3	Intercept	3.525225	.319550	11.032	.000	2.897368	4.153082
	Condition	-.092384	.219840	-.420	.675	-.524720	.339953
	Severity	-2.521554	.235438	-10.710	.000	-2.984200	-2.058908
	Condition*Severity	.914917	.326169	2.805	.005	.273930	1.555904
	Certainty 1	.522882	.048155	10.858	.000	.428267	.617498
4	Intercept	3.504605	1.005478	3.486	.001	1.525696	5.483513
	Condition	-.105374	.220127	-.479	.632	-.538280	.327532
	Severity	-2.525189	.237398	-10.637	.000	-2.991701	-2.058677
	Condition*Severity	.940719	.326965	2.877	.004	.298156	1.583281
	Certainty 1	.529959	.048443	10.940	.000	.434776	.625142
	Gender	-.275530	.190567	-1.446	.150	-.650899	.099840
	Experience	.004771	.010545	.452	.651	-.016001	.025542
	AOT score	.014623	.238539	.061	.951	-.455186	.484431
Dependent Variable: Perceived cue support							

Table 3.6. Regression model of perceived cue support.

Tables 3.7.-3.10. present the perceived support of each cue across the scenario versions:

Clinical cue	Control condition Mean (SD)	Aided condition Mean (SD)	Overall rating Mean (SD)
General physical examination	6.00 (2.15)	6.67 (1.84)	6.34 (2.02)
Family history	5.56 (2.53)	5.59 (2.26)	5.57 (2.39)
Pain intensity	5.42 (2.41)	5.86 (2.23)	5.65 (2.31)

Self-treatment	4.59 (3.04)	5.31 (2.82)	4.91 (2.94)
Chest pain during the night	6.00 (2.69)	5.09 (2.12)	5.54 (2.44)
Resting ECG results	6.60 (2.28)	6.40 (1.75)	6.51 (2.03)
Other symptoms	6.76 (2.03)	6.43 (1.93)	6.60 (1.98)
Overall	**6.24 (1.94)**	**6.14 (1.82)**	**6.19 (1.88)**

Table 3.7. Mean cue ratings in chest pain scenario (David), MSK version.

Clinical cue	Control condition Mean (SD)	Aided condition Mean (SD)	Overall rating Mean (SD)
General physical examination	4.30 (2.53)	6.20 (SD 2.14)	5.29 (2.51)
Family history	4.72 (2.72)	6.39 (1.69)	5.49 (2.46)
Pain intensity	6.05 (2.82)	6.62 (1.50)	6.38 (2.15)
Self-treatment	2.94 (2.17)	4.28 (2.40)	3.65 (2.36)
Chest pain during the night	4.29 (2.79)	5.33 (1.83)	4.77 (2.43)
Resting ECG results	4.05 (2.86)	5.67 (2.44)	4.96 (2.74)
Other symptoms	3.95 (2.85)	6.30 (1.92)	5.21 (2.66)
Overall	**4.94 (2.55)**	**6.03 (1.67)**	**5.48 (2.21)**

Table 3.8. Mean cue ratings in chest pain scenario (David), cardiac version.

Clinical cue	Control condition Mean (SD)	Aided condition Mean (SD)	Overall rating Mean (SD)
Digital rectal examination	6.64 (2.52)	6.38 (2.12)	6.51 (2.31)
Family history	5.83 (2.92)	6.18 (2.40)	6.03 (2.62)
Mucus in stools	5.24 (2.90)	5.59 (2.46)	5.44 (2.64)

Blood in stools	6.60 (2.46)	6.26 (1.94)	5.50 (2.60)
Ferritin levels (current)	6.37 (3.02)	6.31 (2.32)	5.11 (2.93)
Previous episodes of bowel problems	6.56 (2.93)	5.69 (2.51)	5.41 (2.78)
Sensation of bowel emptying	5.60 (3.29)	6.45 (1.73)	5.12 (2.87)
Overall	**6.48 (2.29)**	**6.27 (1.83)**	**6.37 (2.07)**

Table 3.9. Mean cue ratings in constipation scenario (Barbara), IBS version.

Clinical cue	Control condition Mean (SD)	Aided condition Mean (SD)	Overall rating Mean (SD)
Digital rectal examination	3.79 (2.69)	4.76 (2.50)	4.25 (2.63)
Family history	3.00 (2.74)	4.42 (2.56)	3.73 (2.72)
Mucus in stools	3.75 (2.05)	4.52 (2.00)	4.18 (2.04)
Blood in stools	4.48 (2.85)	4.63 (2.32)	4.56 (2.58)
Ferritin levels (current)	3.72 (3.00)	4.30 (2.51)	4.01 (2.77)
Previous episodes of bowel problems	4.42 (2.72)	5.10 (2.58)	4.73 (2.66)
Sensation of bowel emptying	4.13 (3.08)	4.27 (2.74)	4.20 (2.90)
Overall	**4.17 (2.65)**	**4.83 (2.30)**	**4.50 (2.49)**

Table 3.10. Mean cue ratings in constipation scenario (Barbara), colorectal cancer version.

3.6.4. Change in initial certainty (H_3)

On average, GPs significantly increased certainty in their initial diagnosis after receiving additional patient information by 0.41 units (SD 2.02) on the 0-10 scale ($t = -4.51$, $p < 0.001$). We expected an increase rather than a reduction, as the additional patient information was consistent with both steers. As Table 3.1. shows, increase in

initial certainty was almost identical with and without diagnostic suggestions (0.42 versus 0.40 respectively), which suggests that the decision aid did not lead to a smaller increase in initial certainty (b = 0.028, [-0.31, 0.37], p = 0.872), as the simple logistic regression revealed. In two scenario versions (cardiac steer in chest pain scenario and colorectal steer in constipation scenario), there was an overall reduction (rather than an increase) in GPs' initial certainty.

When we regressed severity of GPs' initial diagnosis on change in initial certainty, the analysis revealed a significant association, suggesting that GPs' who provided a serious initial diagnosis had a reduction in certainty, compared to those who provided a less serious initial diagnosis who showed an increase (b = -2.18, [-2.64, -1.73], p < 0.001). However, the effect of severity differed by condition, as there was a significant interaction between condition and severity (b = 1.078, [0.43, 1.72], p = 0.001), suggesting that change in certainty (either reduction or increase) was smaller when diagnostic suggestions were provided. When initial certainty was included in the model as a covariate, it was found to be significantly associated with increase in initial certainty (b = -0.499, [-0.63, -0.37], p < 0.001), suggesting that GPs with higher initial certainty changed their certainty less after receiving additional patient information. However, the effect of initial certainty depended on whether GPs' initial diagnosis was serious or not (i.e., stronger in more serious diagnoses), as there was a significant interaction between initial certainty and severity of initial diagnosis (b = 0.332, [0.16, 0.51], p < 0.001).

Table 3.11. presents the change in GPs' initial certainty in each patient scenario. For a comparison between the first and second measures of certainty in the initial diagnosis, see table 3.12.

Patient scenario	Control condition Mean (SD)	Aided condition Mean (SD)	Overall Mean (SD)
Chest pain (musculoskeletal)	1.32 (1.98)	0.55 (1.93)	0.94 (1.99)
Chest pain (cardiac)	-0.35 (1.61)	0.23 (1.49)	-0.06 (1.57)

Constipation (irritable bowel syndrome)	1.03 (1.89)	1.06 (1.74)	1.05 (1.81)
Constipation (colorectal cancer)	-0.42 (2.51)	-0.15 (2.10)	-0.28 (2.31)
Overall	0.40 (2.16)	0.42 (1.87)	0.41 (2.02)

Table 3.11. Change in initial certainty across patient scenarios.
Negative values indicate reduction in certainty.

Condition	Scenario	Initial certainty Mean (SD)	Second measure of initial certainty Mean (SD)	Change in certainty Mean (SD)
Control	Chest pain (MSK steer)	5.69 (1.67)	7.02 (2.00)	1.32 (1.98)
	Chest pain (cardiac steer)	7.23 (1.63)	6.87 (1.89)	-0.35 (1.61)
	Constipation (IBS steer)	5.94 (1.74)	6.97 (2.10)	1.03(1.89)
	Constipation (cancer steer)	5.58 (1.99)	5.16 (2.66)	-0.42 (2.51)
	Overall	**6.11 (1.87)**	**6.50 (2.30)**	**0.40 (2.16)**
Aided	Chest pain (MSK steer)	6.13 (1.72)	6.68 (2.09)	0.55 (1.93)
	Chest pain (cardiac steer)	7.08 (1.56)	7.31 (1.36)	0.23 (1.49)
	Constipation (IBS steer)	5.66 (1.46)	6.73 (2.01)	1.06 (1.74)
	Constipation (cancer steer)	5.56 (1.63)	5.42 (2.21)	-0.15 (2.10)
	Overall	**6.11 (1.70)**	**6.53 (2.05)**	**0.42 (1.87)**
Overall	**All steers**	**6.11 (1.79)**	**6.52 (2.18)**	**0.41 (2.02)**

Table 3.12. Measurements of initial certainty.
Negative values indicate reduction in certainty.

The regression model for change in initial diagnosis is presented below:

Linear regression of change in initial certainty							
Step		Estimate	Std. Error	t	Sig.	95% CI	
						Lower Bound	Upper Bound
1	Intercept	.395161	.128464	3.076	.002	.142754	.647568
	Condition	.028226	.174759	.162	.872	-.315982	.372434
2	Intercept	1.488003	.164265	9.059	.000	1.165256	1.810750
	Condition	-.506142	.231011	-2.191	.029	-.960344	-.051941
	Severity	-2.185683	.232256	-9.411	.000	-2.642020	-1.729346
	Condition*Severity	1.077672	.328272	3.283	.001	.432677	1.722667
3	Intercept	4.333611	.410962	10.545	.000	3.526118	5.141103
	Condition	-.441447	.218093	-2.024	.044	-.870253	-.012641
	Severity	-3.938540	.584745	-6.735	.000	-5.087479	-2.789602
	Condition*Severity	.989880	.309673	3.197	.001	.381423	1.598338
	Certainty 1	-.499282	.066799	-7.474	.000	-.630540	-.368024
	Severity*Certainty 1	.331752	.089293	3.715	.000	.156300	.507204
4	Intercept	3.673168	.879061	4.179	.000	1.943358	5.402979
	Condition	-.450548	.218246	-2.064	.040	-.879657	-.021439
	Severity	-3.968192	.585259	-6.780	.000	-5.118161	-2.818224
	Condition*Severity	1.008904	.310345	3.251	.001	.399118	1.618689
	Certainty 1	-.502696	.066983	-7.505	.000	-.634317	-.371075
	Severity*Certainty 1	.340883	.089735	3.799	.000	.164561	.517206
	Gender	.016083	.158622	.101	.919	-.296361	.328528
	Experience	.013923	.008769	1.588	.114	-.003350	.031195
	AOT score	.131354	.200025	.657	.512	-.262579	.525286
Dependent Variable: Change in initial certainty							

Table 3.13. Regression model of change in initial certainty.

3.6.5. Change in diagnosis (H_4)

We collected 496 initial and 496 final diagnoses. On average, GPs changed diagnosis in 17.34% of instances (86/496). In contrast to our hypothesis, there were more diagnostic changes when GPs did not receive diagnostic suggestions than when they did (47/248 versus 39/248). The simple logistic regression showed that this difference was not significant (OR = 0.822, [0.52, 1.31], p = 0.407).

When severity of initial diagnosis was subsequently added to the model, it showed a significant association: GPs who provided a more serious initial diagnosis (i.e., angina, cancer) were more likely to change it compared to GPs who provided a less serious diagnosis (OR = 1.905, [1.18, 3.07], p = 0.008). There was no significant interaction between predictors in the model. We also regressed initial certainty on the dependent variable and found a significant association: GPs with higher certainty were less likely to change their diagnosis (OR = 0.680, [0.59, 0.78], p < 0.001). There was no significant interaction between predictors and covariates.

Table 3.14. presents the frequency of diagnostic changes in each patient scenario:

Patient scenario	Control condition % (proportion)	Aided condition % (proportion)	Overall % (proportion)
Chest pain (musculoskeletal)	11.29 (7/62)	14.52 (9/62)	12.90 (16/124)
Chest pain (cardiac)	9.68 (6/62)	3.22 (2/62)	6.45 (8/124)
Constipation (irritable bowel syndrome)	14.52 (9/62)	22.58 (14/62)	18.55 (23/124)
Constipation (colorectal cancer)	40.32 (25/62)	24.19 (15/62)	32.26 (40/124)
Overall	19.00 (47/248)	15.70 (39/248)	17.34 (86/496)

Table 3.14. Frequency of diagnostic changes across patient scenarios.

The regression model for change in diagnosis is presented below:

Logistic regression of change in diagnosis										
Step		**Coefficient**	**Std. Error**	**t**	**Sig.**	**95% CI**		**Exp(Coefficient)**	**95% CI for Exp(Coefficient)**	
						Lower	**Upper**		**Lower**	**Upper**
1	Intercept	-1.452	1.3237	-1.097	.273	-4.053	1.149	.234	.017	3.154
	Condition	-.197	.2367	-.830	.407	-.662	.269	.822	.516	1.308
2	Intercept	-1.804	1.3250	-1.362	.174	-4.408	.799	.165	.012	2.223
	Condition	-.201	.2382	-.843	.400	-.669	.267	.818	.512	1.307

	Severity	.645	.2433	2.650	.008	.167	1.123	1.905	1.181	3.073
3	Intercept	.293	1.3281	.221	.825	-2.316	2.903	1.341	.099	18.222
	Condition	-.203	.2470	-.823	.411	-.689	.282	.816	.502	1.326
	Severity	.914	.2577	3.547	.000	.408	1.420	2.495	1.504	4.139
	Certainty 1	-.386	.0701	-5.504	.000	-.523	-.248	.680	.593	.780
4	Intercept	.196	1.8373	.107	.915	-3.414	3.806	1.217	.033	44.983
	Condition	-.204	.2471	-.826	.409	-.689	.281	.815	.502	1.325
	Severity	.921	.2608	3.531	.000	.408	1.433	2.511	1.504	4.192
	Certainty 1	-.385	.0703	-5.469	.000	-.523	-.246	.681	.593	.782
	Gender	-.028	.2499	-.112	.911	-.519	.463	.972	.595	1.589
	Experience	.005	.0132	.340	.734	-.021	.030	1.005	.979	1.031
	AOT score	.014	.3159	.043	.966	-.607	.634	1.014	.545	1.886
Dependent variable: Change in diagnosis										

Table 3.15. Regression model of change in diagnosis.

3.6.6. Exploratory analyses

3.6.6.1. Initial certainty

On average, GPs' certainty in their initial diagnosis was 6.11 (SD 1.79) on the 0-10 scale. As in Study 1, GPs provided their initial certainty before they saw diagnostic suggestions and, therefore, it was not possible for the suggestions to have any effect on it. The scenario version that produced the highest initial certainty was the cardiac version of the chest pain scenario (M = 7.15, SD 1.59).

As discussed above, initial certainty was included as a covariate in all the regression models and found to impact the dependent variables, in the expected direction. Specifically, GPs with higher initial certainty a) rated the neutral clinical cues as more supportive of their initial diagnosis, b) requested fewer clinical cues before providing their final diagnosis and c) changed diagnosis less frequently. These findings suggest that, irrespective of the provision of diagnostic suggestions, initial certainty was the main driver in the process.

The initial certainty ratings are presented on the table below.

Patient scenario	Control condition Mean (SD)	Aided condition Mean (SD)	Overall Mean (SD)
Chest pain (MSK steer)	5.69 (1.67)	6.13 (1.72)	5.91 (1.71)
Chest pain (cardiac steer)	7.23 (1.63)	7.08 (1.56)	7.15 (1.59)
Constipation (IBS steer)	5.94 (1.74)	5.66 (1.46)	5.80 (1.61)
Constipation (cancer steer)	5.58 (1.99)	5.56 (1.63)	5.57 (1.81)
Overall	6.11 (1.87)	6.11 (1.70)	6.11 (1.79)

Table 3.16. Initial certainty across patient scenarios and conditions.

3.6.6.2. Final certainty

Final certainty refers to GPs' certainty in their final diagnosis. Across scenarios and conditions, GPs' average final certainty was 6.89 (SD 1.70). There was a significant difference between conditions, suggesting that when GPs received diagnostic suggestions (versus no suggestions), they were less certain in their final diagnosis (6.73, SD 1.80 versus 7.06, SD 1.58, b = -0.326, [-0.58, -0.07], p = 0.011). Initial certainty was also included in the model as a covariate and found to be positively associated with final certainty (b = 0.494, [0.42, 0.57], p < 0.001). There was no interaction between predictors and covariates.

Table 3.17. presents GPs' final certainty across patient scenarios, with and without diagnostic suggestions. For a comparison between the three different measures of certainty, see table 3.18.

Patient scenario	Control condition Mean (SD)	Aided condition Mean (SD)	Overall Mean (SD)
Chest pain (MSK steer)	7.24 (1.63)	6.65 (2.00)	6.94 (1.84)
Chest pain (cardiac steer)	7.18 (1.48)	7.55 (1.14)	7.36 (1.33)

Constipation (IBS steer)	7.21 (1.65)	6.85 (1.61)	7.03 (1.63)
Constipation (cancer steer)	6.60 (1.51)	5.87 (1.95)	6.23 (1.77)
Overall	7.06 (1.58)	6.73 (1.80)	6.89 (1.70)

Table 3.17. Final certainty across patient scenarios and conditions.

Condition	Scenario	Initial certainty Mean (SD)	Second measure of initial certainty Mean (SD)	Final certainty Mean (SD)
Control	Chest pain (MSK steer)	5.69 (1.67)	7.02 (2.00)	7.24 (1.63)
	Chest pain (cardiac steer)	7.23 (1.63)	6.87 (1.89)	7.18 (1.48)
	Constipation (IBS steer)	5.94 (1.74)	6.97 (2.10)	7.21 (1.65)
	Constipation (cancer steer)	5.58 (1.99)	5.16 (2.66)	6.60 (1.51)
	Overall	**6.11 (1.87)**	**6.50 (2.30)**	**7.06 (1.58)**
Aided	Chest pain (MSK steer)	6.13 (1.72)	6.68 (2.09)	6.65 (2.00)
	Chest pain (cardiac steer)	7.08 (1.56)	7.31 (1.36)	7.55 (1.14)
	Constipation (IBS steer)	5.66 (1.46)	6.73 (2.01)	6.85 (1.61)
	Constipation (cancer steer)	5.56 (1.63)	5.42 (2.21)	5.87 (1.95)
	Overall	**6.11 (1.70)**	**6.53 (2.05)**	**6.73 (1.80)**
Overall	**All steers**	**6.11 (1.79)**	**6.52 (2.18)**	**6.89 (1.70)**

Table 3.18. Certainty measurements across conditions and scenarios.

3.6.6.3. Number of differential diagnoses

The number of differential diagnoses refers to the number of alternative diagnostic hypotheses that GPs recorded for each scenario version, after providing their final diagnosis. On average, GPs recorded 1.28 differential diagnoses per scenario version (SD 1.13, range from 0 to 7 diagnoses, mode 1). Across conditions, the means were similar (Control: 1.32 versus Aided: 1.25) with no significant differences

(IRR = 1.06, [0.91, 1.23], p = 0.441, *cf*. Sibbald *et al.*, 2021). Table 3.19. presents the number of differential diagnoses that GPs provided for each patient scenario across the two conditions.

Patient scenario	Control condition Mean (SD)	Aided condition Mean (SD)	Overall Mean (SD)
Chest pain (musculoskeletal)	1.08 (0.87)	1.00 (0.96)	1.04 (0.91)
Chest pain (cardiac)	0.92 (0.97)	1.02 (0.93)	0.97 (0.95)
Constipation (irritable bowel syndrome)	1.71 (1.41)	1.24 (1.03)	1.48 (1.25)
Constipation (colorectal cancer)	1.58 (1.23)	1.73 (1.24)	1.65 (1.24)
Overall	1.32 (1.18)	1.25 (1.08)	1.28 (1.13)

Table 3.19. Number of differential diagnoses that GPs recorded after providing their final diagnosis.

As in the previous regression models, we tested for severity of initial diagnosis and controlled for initial certainty. We found no association between number of differential diagnoses and severity of initial diagnosis (IRR = 0.927, [0.80, 1.08], p = 0.319) and an association between the number of differential diagnoses and Initial certainty (IRR = 0.949, [0.91, 0.99], p = 0.014), suggesting that GPs with higher initial certainty recorded fewer differential diagnoses. There were no significant interactions.

3.6.6.4. Appropriateness of management

On average, GPs management was deemed appropriate in 11% of the cases (54/496). There was a small difference between conditions (Control: 8% versus Aided: 13.00%) which was not significant. A separate logistic regression model was used to test associations with predictors (condition, severity) and covariates. No significant associations were found.

3.6.6.5. Gender

Gender was found to be associated with only one dependent variable: initial certainty. Male GPs were more certain in their initial diagnosis than their female counterparts (6.32, SD 1.82 versus 5.92, SD 1.74, b = 0.38, [0.03, 0.72], p = 0.032).

3.6.6.6. Experience

There was a negative association between GPs' years of experience and number of differential diagnoses (IRR = 0.99, [0.98, 1.00], p = 0.020), suggesting that more experienced GPs recorded fewer differential diagnoses. Experience was not significantly associated with any of the other dependent variables.

3.6.6.7. Actively Open-minded Thinking

The mean AOT score across GPs was 3.91 (SD 0.40) on the 1-5 scale. The AOT score was included as a covariate in all the regression models and was found to be associated with the number of cues that GPs requested before reaching a final diagnosis (IRR = 1.368, [1.22, 1.53] p < 0.001), suggesting that GPs with higher (versus lower) AOT scores engaged in a more extensive information search before reaching a conclusion. Likewise, GPs with higher AOT scores were more likely to record a greater number of differential diagnoses (IRR = 1.247, [1.03, 1.51], p = 0.023). AOT scores bore no significant association to any of the other dependent variables.

3.7. Discussion

In this study we tested whether providing diagnostic suggestions at the early stages of the diagnostic process (i.e., after a brief patient history) might lead GPs to a) engage in a more extensive search for information about the patient; b) evaluate subsequent patient information more fairly; c) question their initial diagnosis, i.e., have a smaller increase in diagnostic certainty; and/or d) change their diagnosis

more frequently. Our main hypothesis was that the decision aid would have a de-biasing effect on GPs. As in Study 1, we hypothesised that providing diagnostic suggestions early would interrupt GPs' intuitive responses and reduce certainty in their judgements, given that there are other possibilities to consider (the diagnostic suggestions from the aid). In turn, this would make them more cautious about their diagnostic judgements, less biased in the interpretation of neutral clinical information and more likely to assess a greater amount of information before settling on a final diagnosis. However, the study failed to detect an effect of the diagnostic suggestions on any of the main outcome variables.

There was an effect of the diagnostic suggestions on GPs' certainty in their final diagnosis. That is, GPs who received diagnostic suggestions were less certain in their final diagnosis than those who did not receive suggestions, possibly due to their exposure to alternative hypotheses. This was a finding that we did not predict, as our focus was on certainty in the initial diagnosis, before and after the provision of diagnostic suggestions. We focused on certainty in the initial (rather than the final) diagnosis because the main hypothesis of this thesis is that early intervention (diagnostic suggestions) targets initial, fluent, intuitive, and possibly more certain responses (also see Chapter 1). Our objective, therefore, was to measure the evolution of these early responses over the course of a patient case, in the presence (versus absence) of diagnostic suggestions.

We also found some effect of the severity of GPs' initial diagnoses on the variables of interest. Specifically, we found that GPs who provided a common and less serious initial diagnosis perceived the neutral cues as more supportive of their diagnosis, had a larger increase in their certainty and changed their diagnosis less frequently. These were towards the expected direction, given that the familiarity of common diseases would possibly influence GPs' interpretation of the neutral cues, finding more evidence for their initial (common) diagnosis. Also, since the additional patient information was not inconsistent with the initial patient description, a larger increase in initial certainty about a common disease would be reasonable. In terms of diagnostic change, since less serious diagnoses are more common in General Practice and given that the additional patient information had minimal diagnostic value, GPs were expected to not change their initial diagnosis.

Based on the results, the only driver of the process (i.e., information search, information evaluation and changing of mind) was initial certainty which was found to be significantly associated with all the variables of interest. Specifically, GPs with higher certainty in their initial diagnosis requested less information before reaching a final diagnosis, rated the neutral patient information as more supportive of their initial diagnosis, recorded fewer differential diagnoses about the patients, and changed diagnosis less frequently. This finding suggests that a change in physicians' certainty in their initial diagnostic hypotheses could have a ripple effect on the diagnostic process.

The study failed to detect an effect of the decision aid on GPs' information search and information interpretation. This could not be explained by the idiosyncrasies of the patient scenarios used in this study, as most of them worked as expected. Addressing the limitations from Study 1, we included stronger signals in the initial patient description to reduce variability between responses and induce higher initial certainty. Most of the GPs accepted the steer of the initial patient description and remained committed to it. However, there was no difference between Aided and Control condition. Taking onboard GPs' feedback from Study 1, we also reduced the number of diagnostic suggestions in the study. This was to avoid any distractions caused by the fact that the decision aid suggested common and uncommon diagnoses together. This manipulation did not help us to detect an effect either.

One possible explanation is that GPs have grown insensitive to decision aids and are quick to dismiss them. GPs from Study 1 reported that during real consultations, "pop-up windows" and alerts often appear on their computer screens and "disrupt" the diagnostic process. Some also reported that the list of diagnostic suggestions reminded them of those pop-ups windows and alerts. If indeed GPs have become accustomed to dismissing these, having seen them so many times, then the chances of detecting an effect for the aid may have been reduced.

In Study 1, we found that when diagnostic suggestions were provided in the scenario seen first, there was a larger reduction in certainty and more diagnostic changes in the scenario seen second, suggesting a possible spillover effect. To avoid such an effect in the present study, we provided diagnostic suggestions always in the

scenario seen second. However, this did not help us to detect an effect of the suggestions on GPs' information search and information interpretation.

Another limitation of the study was the online task itself. We tried to keep it as close to reality as possible. However, we are aware that the diagnostic process was narrowed down in an unrealistic way. Contrary to Study 1, where the additional patient information was fixed and the same for all participants, in the present study we allowed for more active information search by asking GPs to choose the information they needed (from a list of seven clinical cues) in order to make a final diagnosis. However, in real life, GPs would take a more complete patient history, and ask more open and/or different questions, before reaching a diagnostic conclusion. Such constraints were necessary though, in order to reduce variability between responses and ensure that the data are meaningful for this study.

Finally, as GPs were aware that they were being studied and their responses were being recorded, it is possible that they tried to perform "well". This implicit demand could have made GPs switch to an analytical mode of thinking from the beginning of the task. For example, they may have concentrated very hard (trying to find "the correct answer") and/or been very cautious (expecting to be "tricked"). This may have been reflected in their initial ratings of certainty, which were quite low (M = 6), considering that the initial patient description was highly diagnostic. If GPs indeed approached the patient scenarios analytically in the first place, then they could have been free from any bias that would affect their information search and information interpretation. Despite the fact that we did not explicitly mention the purpose of the study to GPs, we were not able to control the potential effects of demand characteristics on their decision making and their impact on the measured outcomes.

Both Studies 1 and 2 failed to detect an effect of the diagnostic suggestions on the psychological mechanisms explored. It would be therefore interesting to see why the decision aid did not work as expected. It is possible that GPs used the decision aid as a tool to validate their own diagnostic hypotheses: if GPs' own diagnoses were included in the list of suggestions (which was indeed the case for the majority of GPs who took part in the study), then this may have reassured them that their diagnosis was "correct". A future study could test this hypothesis. A potential study could use

a) a design where the decision aid does not include GPs' own, initial diagnosis (and therefore cannot serve as a validation), or b) eye tracking methodology to explore GPs' attentional processes when interacting with the aid (specifically, number and length of fixations on their own diagnosis versus the others on the list).

Summary

- 248 GPs assessed two online patient scenarios, one with and the other without suggestions
- Diagnostic suggestions led to more biased evaluation of neutral patient information and to lower certainty in physicians' final diagnosis
- Physicians with higher initial certainty requested fewer clinical cues before reaching a conclusion, were more biased in their evaluations, changed diagnosis less frequently, and recorded fewer differential diagnoses
- Physicians who provided a less serious (versus serious) initial diagnosis requested more cues, were more biased in their evaluations and changed diagnosis less frequently
- Physicians with higher AOT score requested more cues and recorded more differential diagnoses

CHAPTER 4: Removing GPs' own diagnoses from the list of suggestions – a follow up (Study 3)

4.1. Introduction

Study 2 was an attempt to test the effects of diagnostic suggestions on GPs' diagnostic certainty, information search, information evaluation and diagnostic change. The study provided evidence that it was rather GPs' initial certainty, and not the aid's diagnostic suggestions, that predicted a) the amount of clinical information that GPs would seek before reaching a diagnostic conclusion, b) the evaluation of this information as more or less supportive of their diagnosis, and c) the likelihood of diagnostic change when encountering new information. Diagnostic suggestions were found to have some effect only on GPs' final diagnostic certainty, namely, when GPs received diagnostic suggestions, they were less certain about their final diagnosis than when they did not receive suggestions. This could have been for a number of reasons, as discussed above (see section 3.7. Discussion).

4.2. Research questions and objectives

The present study tested one of the potential reasons that the provision of diagnostic suggestions did not produce any effect in Study 2. Specifically, we tested the possibility that GPs used the suggestions to validate their own initial diagnosis rather than to question it. We hypothesised that the inclusion of the GPs' own diagnoses in the list of suggestions could have been a confirmatory sign that would reassure them that they are following the correct diagnostic path. If this were the case, then the provision of diagnostic suggestions would not serve its purpose, which was to enable GPs to think more about their diagnosis by reminding them that there are alternative explanations. The present study is, thus, different from Study 2 in two ways: firstly, GPs were presented with diagnostic suggestions that did not include the most likely diagnosis for each patient scenario; secondly, they saw only one patient scenario, and always with diagnostic suggestions. GPs' responses to scenarios seen without diagnostic suggestions in Study 2 were used as controls.

For each scenario version (steer), we removed the most likely diagnosis from the list of diagnostic suggestions. Specifically, in the chest pain scenario, when the steer pointed towards MSK as the most likely diagnosis, the list of diagnostic suggestions did not include MSK; when it pointed towards a cardiac cause, the list did not include angina. In the constipation scenario, when the steer pointed towards IBS, IBS was not included in the list of diagnostic suggestions; and when it pointed towards colorectal cancer, this diagnosis was not included in the list. This way we ensured that GPs who provided the most likely diagnosis according to the steer did not see this diagnosis in the list of suggestions and thus, could not use the list to validate their initial belief. However, the way the suggestions were introduced did not suggest that GPs should exclude the most likely diagnosis from consideration. Specifically, GPs were told: “A decision aid trialled at your practice suggests that you also consider the following possible diagnoses for this patient.” That is, we added the word “also” to ensure that GPs would not think that their diagnosis was not a likely one. Only GPs who accepted the steer (i.e., those who initially provided the diagnosis indicated by the steer) were included in the analysis, as we could not predict what other diagnoses GPs might provide, so that we remove them from the list.

The aims of this study were the same as in Study 2. That is, to investigate whether providing diagnostic suggestions early in the diagnostic process can affect a) certainty in an initial diagnosis, b) extent of information search c) information evaluation and d) likelihood of diagnostic change. The present study also aimed to test whether removing the most likely diagnosis from the list of diagnostic suggestions can yield any different results compared to Study 2. As in Studies 1 and 2, the effect of individual differences (gender, years of experience, thinking style) were also explored.

4.3. Hypotheses

In comparison to both Control and Aided responses from Study 2, we expected that GPs who saw diagnostic suggestions that did not include their own initial diagnosis (referred to henceforth as “Aided 2 condition”) would:

H_1: request more information about the patient before providing a final diagnosis,

H_2: rate the neutral cues as less supportive of their initial diagnosis,

H_3: increase their certainty about their initial diagnosis less after acquiring additional patient information,

H_4: change their initial diagnosis more frequently.

Following the results from Study 2, we further hypothesised that:

H_5: GPs’ certainty about their initial diagnosis would be positively associated with their perceived cue support, and negatively associated with the number of cues requested and frequency of diagnostic changes,

H_6: GPs in this study would have lower final certainty.

4.4. Methods

4.4.1. Participants

As in the previous studies, participants in the present study were fully qualified GPs and GP trainees, currently practising in England.

4.4.2. Sample size

Study 2 was powered for the main hypothesis that GPs would evaluate the neutral cues as less supportive of their initial diagnosis when they diagnose with the decision aid than without (see Chapter 3, Hypothesis 1). We estimated that we needed at least 243 participants and increased this to 250 participants, assuming

that not all responses would be analysable. In the present study, we recruited the same number of participants, so we could use the responses from Study 2 for comparison.

4.4.3. Recruitment

As in Study 2, participants were recruited through the North West London Clinical Research Network. An invitation e-mail was circulated among participating GP practices. The invitation e-mail contained information about the study and a link to the study website. GPs were offered a £10 Amazon voucher for their participation.

The study was approved by the Health Research Authority (REC: 20/HRA/3867).

4.4.4. Materials

The present study used the same patient scenarios and steers as Study 2. As before, each scenario had two versions (the steers). One version steered GPs towards a benign and common diagnosis (i.e., MSK in the chest pain scenario and IBS in the constipation scenario), while the other version steered GPs towards a serious and less common diagnosis (i.e., angina in the chest pain scenario and colorectal cancer in the constipation scenario).

The AOT scale used in Studies 1 and 2 was also used in this study.

4.4.5. Design

Each GP saw only one version of only one scenario, allocated at random. The randomiser in Qualtrics ensured that each scenario version (MSK, angina, IBS, colorectal cancer) was evenly allocated, so that all scenario versions were seen by the same number of participants. All study participants received the list of suggested diagnoses for the scenario version allocated to them. Their responses were

compared with both the Aided and Control responses from Study 2 in separate, between-subject analyses.

4.4.6. Procedure

As in Study 2, the scenarios were administered online using Qualtrics (Qualtrics.com). The procedure for the participating GPs was almost identical to Study 2, with the exception that each GP saw only one patient scenario where the diagnostic suggestions were introduced as "A decision aid trialled at your practice suggests that you also consider the following possible diagnoses for this patient. Diagnoses are in alphabetical order. Please read them carefully"[17]. As in Study 2, certainty in diagnosis and perceived cue support were measured on the same Visual Analogue Scales (VAS), both ranging from 0 ("Not at all certain"/"No support") to 10 ("Absolutely certain"/"Strong support").

After GPs completed the AOT scale, they also answered three questions about the decision aid:

1. Did you find the diagnostic suggestions from the decision aid helpful?
2. Would you like to have a similar decision aid in your practice?
3. Would you trust the diagnostic suggestions of such a decision aid?

GPs indicated their responses on a 4-point scale ("Definitely yes" to "Definitely not").

4.5. Analyses

We used the data from both this study and Study 2 to form three conditions for analysis purposes:

1. **Control** (scenario versions seen without diagnostic suggestions in Study 2).

[17] In Study 1, respondents read "A decision aid trialled at your practice makes these diagnostic suggestions about the patient (in alphabetical order)." In Study 2, they read "A decision aid trialled at your practice suggests the following possible diagnoses for this patient. Diagnoses are in alphabetical order. Please read them carefully."

2. **Aided 1** (scenario versions seen with diagnostic suggestions in Study 2).
3. **Aided 2** (scenario versions seen with diagnostic suggestions, where the most likely diagnosis was removed from the list).

Control and Aided 1 responses were elicited from the same 250 GPs, i.e., they were not independent. Aided 2 responses were elicited from entirely different GPs.

4.5.1. Dependent variables

Similar to Study 2, and based on the hypotheses above, the dependent variables were:

1. **Number of cues** requested (H1), a count variable.
2. **Perceived cue support** (H2), a continuous variable measured on a 0-10 scale.
3. **Change in initial certainty** after obtaining additional cues (H3), a continuous variable (i.e., the difference between the second and first measurements of initial certainty).
4. **Change in diagnosis** (H4), a dichotomous variable (change versus no change), coded as 0 for no change and 1 for change.
5. **Certainty about the initial diagnosis** (H5), a continuous variable measured on a 0-10 scale.
6. **Certainty about the final diagnosis** (H6), a continuous variable measured on a 0-10 scale.
7. **AOT score**, a continuous variable measured on a scale with verbal anchors from completely disagree (1) to completely agree (5).
8. **Number of differential diagnoses,** a count variable.
9. **Appropriateness of management,** a dichotomous variable, coded as 0 for inappropriate management and 1 for appropriate management[18].

[18] As in Study 2, appropriateness of patient management was evaluated according to suggestions of Dr Kavleen Arora.

4.5.2. Independent variables

As in Study 2, the following independent variables were used in the analyses:

1. **Condition**, coded as 0 for Control, 1 for Aided 1 and 2 for Aided 2.
2. **Severity of initial diagnosis**, coded as 0 for less serious and 1 for more serious diagnoses.
3. **GPs' gender,** coded as 0 for females, 1 for males and 2 for other.
4. **GPs' years of experience**, a continuous variable.

4.5.3. Preparing the data for analysis

The independent variable "Condition" could take three values: 0 for Control responses from Study 2, 1 for Aided responses from Study 2 (Aided 1 condition), and 2 for Study 3 responses (Aided 2 condition). Control and Aided 1 responses were not independent, since they came from the same participants. Therefore, two separate comparisons were conducted: Control versus Aided 2 and Aided 1 versus Aided 2 (the comparison between Control and Aided 1 responses was redundant, since it was completed in Study 2, Chapter 3). To standardise GPs' diagnoses and measure diagnostic changes, the same rules were followed as in Study 2 (see Chapter 3: Preparing the data for analysis).

4.5.4. Statistical tests

We conducted linear regression analyses for continuous dependent variables, logistic for dichotomous dependent variables, and Poisson for count variables. The main predictor was always condition, on which we regressed all the dependent variables using simple regression. All the models were subsequently extended to multiple regression ones by including the following additional independent variables and interactions in steps:

4. Condition, severity of initial diagnosis and interactions
5. Condition, severity of initial diagnosis, initial certainty and interactions

6. Condition, severity of initial diagnosis, initial certainty, gender, GP experience, AOT score and interactions.

We conducted separate regression analyses to model the association of Control versus Aided 2 responses with the dependent variables, and the association of Aided 1 versus Aided 2 responses with the dependent variables. In all regression models we included initial certainty as a covariate and severity of initial diagnosis as an explanatory variable. We also tested for interactions between predictors, explanatory variables and covariates. Finally, we tested for associations between the dependent variables and individual differences by including gender as an explanatory variable and GPs' experience and thinking style (i.e., AOT score) as covariates in the regression models. A direct comparison that would include all three conditions was not possible (e.g., a one-way ANOVA), given that the dataset included both dependent and independent responses from two different studies/datasets.

4.5.5. Data exclusion

We excluded GPs who did not provide an initial diagnosis or provided more than one initial diagnosis. We also excluded GPs who provided a non-analysable response as their initial diagnosis (e.g., 'uncertain' or "I need more information"). This kind of responses were excluded as there was no way to measure perceived support relative to multiple initial diagnoses, or no diagnosis. GPs who did not complete their participation to the end were also excluded. Finally, for the reasons described earlier, GPs who did not provide the expected diagnosis according to the steers were not included in the analyses. GPs in the Control and Aided 1 conditions of Study 2 who did not provide the expected diagnoses were also excluded from the analyses.

4.6. Results

258 GPs were originally recruited. Eight of them were excluded due to incomplete data. Another two GPs were excluded due to non-analysable responses (e.g.,

multiple initial diagnoses). From the remaining 248 GPs, 58 GPs did not provide the expected diagnoses and were also excluded. The final sample consisted of 190 GPs: 74 males (38.90%) and 172 GPs fully-qualified in general practice (90.50%), with an average experience post-qualification of 10.33 years (SD 8.74), ranging from 0 to 36 years (median 7 years). Participation lasted on average 13 minutes.

Following exclusion of GPs who did not provide the expected diagnoses in Study 2, a total of 200 Control responses and 191 Aided 1 responses were included in the analyses. Table 4.1. presents a summary of the results across the three conditions.

Dependent variable	Control Mean (SD)	Aided 1 Mean (SD)	Aided 2 Mean (SD)
Initial certainty	6.38 (1.77)	6.35 (1.66)	6.13 (1.76)
Second measurement of certainty	6.64 (2.27)	6.86 (1.84)	6.42 (2.47)
Perceived cue support	5.50 (2.60)	5.90 (1.99)	5.57 (2.39)
Number of cues requested	4.12 (2.10)	4.38 (1.85)	4.42 (1.96)
Change in initial certainty	0.27 (2.12)	0.51 (1.78)	0.29 (2.24)
Change in diagnosis	18%	9.40%	18.90%
Certainty about the final diagnosis	7.19 (1.50)	6.97 (1.67)	6.95 (1.82)
Number of differential diagnoses	1.26 (1.15)	1.25 (1.07)	1.34 (1.11)
Appropriateness of management	10%	16%	14%

Table 4.1. Results across conditions.
Change in initial certainty was positive in all conditions (increase).

For the descriptive analysis of initial and final diagnoses (reported in 4.6.1.) and GPs' attitudes towards the decision aid (reported in 4.6.7), we used the total number of participants (instead of only GPs who provided the expected diagnosis) for comparison reasons (also see 4.6.1. for details).

4.6.1. Initial and final diagnoses

As previously explained, for analysis purposes we excluded GPs who did not provide the expected diagnosis. However, in order to make a more meaningful comparison with Study 2 responses, we checked the proportion of GPs who did not accept the steer and provided a different (than the expected) diagnosis. For this reason, and contrary to the analyses described above, the data reported in this section include all GPs' responses (i.e., those who provided and those who did not provide the expected diagnosis).

On average, GPs provided the expected diagnosis in 76.60% of the cases (190/248). Steer acceptance in Study 2 was 78.83% (391/496). Steer acceptance per scenario in the present study was:

- 82.80% (53/64) for chest pain (MSK steer),
- 95.20% (59/62) for chest pain (cardiac steer),
- 68.30% (41/60) for constipation (IBS steer) and
- 61.30% (38/62) for constipation (colorectal cancer steer).

Table 4.2. summarises the initial diagnoses that GPs provided in each patient scenario and version.

Scenario	Version	Diagnosis	Proportion (%)
Chest pain scenario	**Muscular pain (MSK)**	MSK	82.80
		Cardiac diagnosis	17.20
	Cardiac pain	MSK	3.20
		Cardiac diagnosis	95.20
		Bronchospasm	1.60
Constipation scenario	**Irritable bowel syndrome (IBS)**	IBS	68.30
		Cancer	11.70
		Simple constipation	10.00
		Dehydration	1.70

		Diverticular disease	8.30
	Colorectal cancer	IBS	24.20
		Cancer	61.30
		Simple constipation	3.20
		Anaemia	3.20
		Diverticular disease	3.20
		Polyp	1.60
		Inflammatory bowel disease (IBD)	3.20

Table 4.2. Proportion of initial diagnoses across patient scenarios.

Final diagnosis was not expected to be different from initial diagnosis, given that the intermediate patient information (i.e., the neutral clinical cues) had minimal diagnostic value. In line with our hypothesis, most GPs (84.50%) maintained their initial diagnosis throughout the diagnostic task (see also Change in diagnosis above). Table 4.3. presents the GPs' final diagnoses across patient scenarios.

Scenario	Version	Diagnosis	Proportion (%)
Chest pain scenario	**Muscular pain**	MSK	86.60
		Cardiac diagnosis	8.90
		Non-cardiac diagnosis	1.30
		Gastritis	0.60
		Uncertain	2.50
	Cardiac pain	MSK	2.80
		Cardiac diagnosis	94.40
		Gastroesophageal reflux disease (GORD)	1.70
		GORD or Cardiac	0.6

		Asthma	0.60
Constipation scenario	**Irritable bowel syndrome (IBS)**	IBS	90.70
		Cancer	2.30
		Simple constipation	1.60
		Anxiety	0.80
		Diverticular disease	2.30
		IBD	1.60
		IBS or cancer	0.80
	Colorectal cancer	IBS	27.00
		Cancer	59.10
		IBD	2.60
		IBS or cancer	0.90
		Diverticular disease	6.10
		IBS or IBD	0.90
		Uncertain	3.50

Table 4.3. GPs' final diagnoses across patient scenarios.

4.6.2. Number of cues requested (H_1)

The number of cues that GPs requested refers to the amount of information that GPs collected before settling on a final diagnosis. The minimum number of cues that GPs could have requested was one and the maximum was seven. On average, GPs requested 4.30 cues per patient scenario across the three conditions (SD 1.98). In line with our hypothesis, GPs in Aided 2 condition requested on average more cues than in the other two conditions, especially in more serious diseases like angina and cancer (Table 4.1.). To test whether the effect of diagnostic suggestions on the number of cues requested was significant, we regressed the dependent variable on condition using simple Poisson regression. We found no significant differences between conditions (Control versus Aided 2: IRR = 0.933, [0.85, 1.03], p = 0.157 and

Aided 1 versus Aided 2: IRR = 0.991, [0.90, 1.09], p = 0.857). We subsequently included severity of initial diagnosis in the model, and we found no association (Control versus Aided 2: IRR = 1.079, [0.98, 1.19], p = 0.123 and Aided 1 versus Aided 2: IRR = 1.097, [1.00, 1.21], p = 0.057), with no interactions.

Initial certainty was subsequently included in the models as a covariate and was found to be associated with the number of cues that GPs requested, suggesting that GPs with higher initial certainty reached a diagnostic conclusion based on fewer pieces of information (Control versus Aided 1: IRR = 0.902, [0.88, 0.93], p < 0.001 and Aided 1 versus Aided 2: IRR = 0.913, [0.89, 0.94], p < 0.001). There were no interactions (see Tables 4.4. and 4.5.).

Poisson regression of number of cues requested (Control versus Aided 2)											
Step		**B**	**Std. Error**	**95% Wald CI**		**Hypothesis Test**			**Exp(B)**	**95% Wald CI for Exp(B)**	
				Lower	**Upper**	**Wald χ^2**	**df**	**Sig.**		**Lower**	**Upper**
1	Intercept	1.485	.0345	1.418	1.553	1850.65	1	.000	4.416	4.127	4.725
	Condition	-.069	.0490	-.165	.027	1.998	1	.157	.933	.848	1.027
2	Intercept	1.447	.0426	1.364	1.531	1155.01	1	.000	4.252	3.912	4.622
	Condition	-.070	.0490	-.166	.026	2.023	1	.155	.933	.847	1.027
	Severity	.076	.0491	-.020	.172	2.379	1	.123	1.079	.980	1.187
3	Intercept	2.094	.0911	1.915	2.272	528.248	1	.000	8.113	6.787	9.699
	Condition	-.044	.0492	-.140	.052	.797	1	.372	.957	.869	1.054
	Severity	.010	.0497	-.087	.108	.043	1	.837	1.010	.917	1.114
	Certainty 1	-.103	.0132	-.129	-.077	60.468	1	.000	.902	.879	.926
4	Intercept	1.319	.2883	.754	1.884	20.917	1	.000	3.738	2.124	6.578
	Condition	-.060	.0503	-.158	.039	1.412	1	.235	.942	.853	1.040
	Severity	.000	.0507	-.099	.100	.000	1	.995	1.000	.906	1.105
	Certainty 1	-.095	.0137	-.122	-.068	48.229	1	.000	.909	.885	.934
	Gender	-.061	.0520	-.163	.041	1.371	1	.242	.941	.850	1.042
	Experience	.000	.0029	-.006	.005	.003	1	.954	1.000	.994	1.005
	AOT score	.196	.0672	.065	.328	8.541	1	.003	1.217	1.067	1.388
Dependent Variable: Number of cues requested											

Table 4.4. Regression model of number of cues requested (Control versus Aided 2).

Poisson regression of number of cues requested (Aided 1 versus Aided 2)											
Step		B	Std. Error	95% Wald CI		Hypothesis Test			Exp(B)	95% Wald CI for Exp(B)	
				Lower	Upper	Wald χ^2	df	Sig.		Lower	Upper
1	Intercept	1.485	.0345	1.418	1.553	1850.64	1	.000	4.416	4.127	4.725
	Condition	-.009	.0489	-.105	.087	.033	1	.857	.991	.901	1.091
2	Intercept	1.439	.0427	1.355	1.522	1137.106	1	.000	4.215	3.877	4.583
	Condition	-.009	.0489	-.105	.087	.035	1	.853	.991	.900	1.091
	Severity	.093	.0489	-.003	.189	3.612	1	.057	1.097	.997	1.208
3	Intercept	2.009	.0948	1.823	2.195	449.033	1	.000	7.456	6.192	8.978
	Condition	.013	.0490	-.083	.109	.073	1	.787	1.013	.921	1.116
	Severity	.045	.0494	-.052	.142	.825	1	.364	1.046	.949	1.152
	Certainty 1	-.091	.0139	-.119	-.064	43.078	1	.000	.913	.888	.938
4	Intercept	1.490	.2978	.906	2.073	25.019	1	.000	4.436	2.474	7.951
	Condition	-.003	.0502	-.101	.095	.004	1	.953	.997	.904	1.100
	Severity	.040	.0503	-.058	.139	.646	1	.422	1.041	.944	1.149
	Certainty 1	-.085	.0145	-.113	-.056	33.999	1	.000	.919	.893	.945
	Gender	-.062	.0525	-.165	.041	1.384	1	.239	.940	.848	1.042
	Experience	.000	.0029	-.005	.006	.008	1	.928	1.000	.995	1.006
	AOT score	.132	.0694	-.004	.268	3.602	1	.058	1.141	.996	1.307
Dependent Variable: Number of cues requested											

Table 4.5. Regression model of number of cues requested (Aided 1 versus Aided 2).

Table 4.6. presents the number of cues that GPs requested for each patient scenario across the three conditions.

Patient scenario	Control Mean (SD)	Aided 1 Mean (SD)	Aided 2 Mean (SD)	Overall Mean (SD)
Chest pain (MSK steer)	4.76 (2.12)	4.84 (1.84)	4.48 (2.13)	3.58 (2.81)
Chest pain (cardiac steer)	3.53 (2.19)	3.73 (1.77)	4.32 (2.10)	2.68 (2.39)
Constipation (IBS steer)	4.00 (1.64)	4.67 (1.69)	4.39 (1.83)	3.44 (2.05)

Constipation (cancer steer)	4.31 (2.19)	4.42 (1.94)	4.50 (1.69)	3.49 (2.62)
Overall	4.12 (2.10)	4.38 (1.85)	4.42 (1.96)	4.30 (1.98)

Table 4.6. Number of cues requested across scenarios and conditions.

In the chest pain scenario for both MSK and cardiac steers, the three most requested cues were General physical examination (MSK steer: 80.30%, Cardiac steer: 64.40%), Resting ECG results (MSK steer: 70.10%, Cardiac steer: 64.40%) and Family history (MSK steer: 63.70%, Cardiac steer: 56.10%). In the constipation scenario for both IBS and colorectal cancer steers, the most requested cues were Blood in stools (IBS steer: 76.70%, cancer steer: 73%), Digital rectal examination (IBS steer: 74.40%, Cancer steer: 73%) and Ferritin levels (IBS steer: 69.80%, Cancer steer: 73.90%).

4.6.3. Perceived cue support (H_2)

Table 4.1. above presents the mean cue ratings (SD) per condition. Across conditions and scenarios, the average perceived cue support was 5.65 (SD 2.35) on the 0-10 scale. As expected, GPs who received diagnostic suggestions that did not include their own diagnosis (i.e., Aided 2 condition) perceived the neutral cues as less supportive of their initial diagnosis compared to GPs who received suggestions that included their own diagnosis (i.e., Aided 1 condition), but not compared to the Control responses. To test the effects of condition on perceived cue support, we regressed the dependent variable on condition, using simple linear regression. The regression analyses did not detect any significant differences between conditions (Control versus Aided 2: b = 0.037, [-0.21, 0.29], p = 0.768, Aided 1 versus Aided 2: b = -0.33, [-0.77, 0.11], p = 0.144).

Severity of initial diagnosis was subsequently included as an explanatory variable in both regression models to test whether perceived support depended on how serious the initial diagnosis was. The first regression model showed that perceived support was associated with severity of initial diagnosis (Control versus Aided 2: b = -2.42, [-

2.85, -1.98], p < 0.001) with no interaction, while the other model showed no association (Aided 1 versus Aided 2: b = -0.438, [-1.70, 0.83], p = 0.497) with an interaction between severity and condition (b = -0.941, [-1.74, -0.14], p = 0.021). Specifically, in the first model, GPs who provided less serious initial diagnoses rated the subsequent patient information as more supportive of their initial diagnoses. In the second model, GPs who provided serious initial diagnoses found more support for it when they received diagnostic suggestions that included their own diagnosis (Aided 1 condition) than when they received suggestions that did not include their own diagnosis (Aided 2 condition). For less serious initial diagnoses, condition did not make any difference in perceived cue support.

Initial certainty was also included in both models as a covariate and was found to be positively associated with perceived cue support (Control versus Aided 1: b = 0.456, [0.34, 0.57], p < 0.001 and Aided 1 versus Aided 2: b = 0.441, [0.33, 0.55], p < 0.001), suggesting that GPs with higher certainty in their initial diagnosis found more support for it when assessing the neutral patient information. There were no interactions (see Tables 4.7. and 4.8.).

Linear regression of perceived cue support (Control versus Aided 2)

Step		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% CI for B		R^2
		B	Std. Error	Beta			Lower Bound	Upper Bound	
1	(Constant)	5.496	.177		31.067	.000	5.148	5.843	0.000224
	Condition	.037	.127	.015	.295	.768	-.212	.287	
2	(Constant)	6.716	.191		35.110	.000	6.340	7.092	0.235
	Condition	.044	.111	.018	.397	.692	-.174	.262	
	Severity	-2.416	.222	-.484	-10.884	.000	-2.853	-1.980	
3	(Constant)	3.964	.403		9.825	.000	3.171	4.757	0.334
	Condition	.102	.104	.041	.978	.329	-.103	.306	
	Severity	-2.724	.211	-.546	-12.899	.000	-3.140	-2.309	
	Certainty 1	.456	.060	.323	7.608	.000	.338	.574	
4	(Constant)	2.967	1.207		2.459	.014	.594	5.340	0.343
	Condition	.049	.106	.020	.462	.645	-.160	.258	
	Severity	-2.689	.215	-.541	-12.489	.000	-3.113	-2.266	

	Certainty 1	.470	.062	.334	7.607	.000	.349	.592	
	Gender	-.140	.216	-.028	-.648	.517	-.566	.285	
	Experience	.000	.012	.000	-.011	.991	-.024	.024	
	AOT score	.242	.277	.038	.876	.381	-.301	.786	
Dependent Variable: Perceived cue support									

Table 4.7. Regression model of perceived cue support (Control versus Aided 2).

Linear regression of perceived cue support (Aided 1 versus Aided 2)									
Step		**Unstandardized Coefficients**		**Standardized Coefficients**	**t**	**Sig.**	**95.0% CI for B**		**R^2**
		B	**Std. Error**	**Beta**			**Lower Bound**	**Upper Bound**	
1	(Constant)	6.229	.356		17.498	.000	5.529	6.929	0.006
	Condition	-.330	.225	-.075	-1.462	.144	-.773	.114	
2	(Constant)	6.445	.458		14.058	.000	5.544	7.347	0.194
	Condition	.155	.290	.035	.533	.594	-.416	.726	
	Severity	-.438	.643	-.099	-.680	.497	-1.702	.827	
	Condition*Severity	-.941	.407	-.355	-2.313	.021	-1.742	-.141	
3	(Constant)	3.711	.549		6.763	.000	2.632	4.790	0.308
	Condition	.226	.270	.051	.837	.403	-.304	.756	
	Severity	-.761	.598	-.173	-1.274	.204	-1.937	.414	
	Condition*Severity	-.885	.378	-.334	-2.344	.020	-1.627	-.143	
	Certainty 1	.441	.056	.344	7.887	.000	.331	.551	
4	(Constant)	1.766	1.188		1.487	.138	-.570	4.102	0.329
	Condition	.098	.271	.022	.361	.718	-.436	.632	
	Severity	-.726	.596	-.167	-1.218	.224	-1.898	.447	
	Condition*Severity	-.871	.381	-.327	-2.283	.023	-1.621	-.121	
	Certainty 1	.472	.057	.371	8.232	.000	.359	.584	
	Gender	-.286	.197	-.065	-1.452	.147	-.674	.101	
	Experience	-.006	.011	-.023	-.531	.596	-.028	.016	
	AOT score	.519	.256	.089	2.027	.043	.015	1.023	
Dependent Variable: Perceived cue support									

Table 4.8. Regression model of perceived cue support (Aided 1 versus Aided 2).

There was high variability in perceived cue support across scenarios. The following tables present the mean ratings for each cue across conditions and patient scenarios:

Clinical cue	Control Mean (SD)	Aided 1 Mean (SD)	Aided 2 Mean (SD)	Overall rating Mean (SD)
General physical examination	6.19 (2.10)	6.92 (1.70)	5.23 (2.58)	6.23 (2.17)
Family history	5.47 (2.68)	5.82 (2.15)	5.27 (1.98)	5.53 (2.30)
Pain intensity	4.96 (2.35)	6.11 (2.09)	5.14 (2.07)	5.47 (2.21)
Self-treatment	4.66 (3.04)	5.67 (2.61)	6.02 (2.68)	5.53 (2.81)
Chest pain during the night	5.86 (2.76)	5.33 (2.07)	6.24 (2.38)	5.82 (2.42)
Resting ECG results	6.84 (2.14)	6.69 (1.70)	5.32 (2.81)	6.36 (2.30)
Other symptoms	6.80 (2.10)	6.69 (1.87)	6.89 (1.36)	6.77 (1.93)
Overall	**6.32 (1.97)**	**6.49 (1.51)**	**6.28 (1.83)**	**6.36 (1.78)**

Table 4.9. Mean cue ratings in chest pain scenario, MSK steer.

Clinical cue	Control Mean (SD)	Aided 1 Mean (SD)	Aided 2 Mean (SD)	Overall rating Mean (SD)
General physical examination	4.30 (2.53)	6.14 (2.21)	5.36 (2.49)	5.27 (2.51)
Family history	5.72 (2.75)	6.35 (1.74)	6.16 (1.95)	5.66 (2.35)
Pain intensity	6.05 (2.82)	6.73 (1.40)	4.70 (2.64)	5.80 (2.47)
Self-treatment	2.94 (2.17)	3.93 (2.40)	5.43 (2.65)	4.62 (2.69)
Chest pain during the night	4.29 (2.79)	5.14 (1.82)	4.56 (2.73)	4.61 (2.57)
Resting ECG results	4.05 (2.86)	5.61 (2.47)	3.59 (2.78)	4.60 (2.80)
Other symptoms	3.95 (2.85)	6.32 (1.95)	5.50 (2.37)	5.21 (2.64)

Overall	**4.94 (2.55)**	**6.01 (1.70)**	**5.25 (2.10)**	**5.40 (2.19)**

Table 4.10. Mean cue ratings in chest pain scenario, cardiac steer.

Clinical cue	Control Mean (SD)	Aided 1 Mean (SD)	Aided 2 Mean (SD)	Overall rating Mean (SD)
Digital rectal examination	7.53 (1.38)	6.89 (1.56)	7.41 (1.82)	7.25 (1.57)
Family history	6.52 (2.58)	6.21 (2.42)	6.83 (2.04)	6.40 (2.42)
Mucus in stools	5.94 (2.46)	5.50 (2.55)	7.24 (2.05)	6.37 (2.42)
Blood in stools	7.51 (1.45)	6.71 (1.35)	7.23 (1.87)	7.14 (1.54)
Ferritin levels (current)	7.73 (1.15)	7.00 (1.68)	7.42 (1.91)	7.36 (1.64)
Previous episodes of bowel problems	7.91 (1.63)	6.24 (1.82)	7.38 (1.80)	7.25 (1.85)
Sensation of bowel emptying	7.28 (2.92)	6.83 (1.43)	7.61 (1.87)	7.26 (1.70)
Overall	**7.29 (1.36)**	**6.74 (1.48)**	**7.36 (1.38)**	**7.13 (1.43)**

Table 4.11. Mean cue ratings in constipation scenario, IBS version.

Clinical cue	Control Mean (SD)	Aided 1 Mean (SD)	Aided 2 Mean (SD)	Overall rating Mean (SD)
Digital rectal examination	2.79 (2.31)	3.93 (2.73)	4.36 (3.37)	3.58 (2.81)
Family history	1.83 (1.82)	3.75 (2.52)	0.67 (1.15)	2.68 (2.39)
Mucus in stools	3.58 (2.15)	4.07 (1.68)	3.10 (2.15)	3.44 (2.05)
Blood in stools	3.28 (2.74)	3.85 (2.26)	3.19 (3.04)	3.49 (2.62)
Ferritin levels (current)	1.92 (2.43)	3.14 (2.21)	3.19 (2.53)	2.79 (2.44)
Previous episodes of	3.71 (2.56)	4.60 (2.23)	3.05 (2.35)	3.77 (2.44)

bowel problems				
Sensation of bowel emptying	3.00 (2.55)	3.40 (2.72)	3.43 (2.50)	3.29 (2.55)
Overall	**3.16 (2.57)**	**3.99 (2.26)**	**3.16 (2.28)**	**3.44 (2.39)**

Table 4.12. Mean cue ratings in constipation scenario, colorectal cancer version.

4.6.4. Change in initial certainty (H_3)

Change in certainty was the difference between the second and the first measurements of certainty about the initial diagnosis. Certainty was first measured after GPs made their initial diagnosis. It was measured again after they had requested additional cues, just before providing a final diagnosis. On average, GPs increased their certainty by 0.35 units (SD 2.06) on the 0-10 response scale. As Table 4.15. shows, overall GPs increased their certainty in an initial diagnosis in all three conditions. However, in two scenario versions (cardiac and cancer steers), there was an overall reduction, rather than an increase in certainty. As expected, GPs who received diagnostic suggestions that did not include their own diagnosis (i.e., Aided 2 condition) increased their certainty in their initial diagnosis less than GPs who received suggestions that included their own diagnosis (Aided 1 condition) and similar to the Control condition (Table 4.1.).

To test the effects of condition on change in initial certainty we regressed the dependent variable on condition using simple linear regression. We found no differences (Control versus Aided 2: b = 0.015, [-0.20, 0.23], p = 0.893 and Aided 1 versus Aided 2: b = -0.213, [-0.62, 0.19], p = 0.305). To test whether change in initial certainty differed by severity of initial diagnosis, we subsequently added severity as an explanatory variable in the regression models. Both models showed that there was a significant association (Control versus Aided 2: b = -2.055, [-2.44, -1.67], p < 0.001 and Aided 1 versus Aided 2: b = -1.50, [-1.88, -1.21], p < 0.001), suggesting that GPs' who provided a serious initial diagnosis had a reduction in certainty, compared to those who provided a less serious initial diagnosis who showed an

increase. There was no interaction between condition and severity in any of the regression models (see Tables 4.13. and 4.14.).

Initial certainty was also included in the models as a covariate and found to be associated with change in certainty after GPs acquired further patient information: GPs with higher initial certainty had a smaller increase in it after obtaining further patient information (Control versus Aided 1: b = -0.562, [-0.76, -0.37], p < 0.001 and Aided 1 versus Aided 2: b = 0.791, [-1.15, -0.43], p < 0.001). However, in both regression models, there was a significant interaction between severity of initial diagnosis and initial certainty, suggesting that (in line with Study 2) the effect of initial certainty on change in certainty was stronger in more serious diagnoses (Control versus Aided 1: b = 0.518, [0.31, 0.73], p < 0.001 and Aided 1 versus Aided 2: 0.416, [0.20, 0.63], p < 0.001).

Linear Regression of change in initial certainty (Control versus Aided 2)									
Step		**Unstandardized Coefficients**		**Standardized Coefficients**	**t**	**Sig.**	**95.0% CI for B**		**R^2**
		B	**Std. Error**	**Beta**			**Lower Bound**	**Upper Bound**	
1	(Constant)	.265	.154		1.718	.087	-.038	.568	0.000 047
	Condition	.015	.111	.007	.135	.893	-.202	.232	
2	(Constant)	1.303	.168		7.755	.000	.973	1.633	0.223
	Condition	.021	.098	.009	.211	.833	-.171	.212	
	Severity	-2.055	.195	-.472	-10.540	.000	-2.439	-1.672	
3	(Constant)	4.579	.503		9.099	.000	3.589	5.568	0.308
	Condition	.009	.093	.004	.096	.924	-.173	.191	
	Severity	-5.094	.693	-1.170	-7.351	.000	-6.456	-3.731	
	Certainty 1	-.552	.081	-.448	-6.855	.000	-.711	-.394	
	Severity*Certainty 1	.518	.107	.843	4.820	.000	.307	.729	
4	(Constant)	3.622	1.127		3.214	.001	1.406	5.838	0.305
	Condition	-.034	.095	-.016	-.356	.722	-.221	.153	
	Severity	-4.789	.708	-1.111	-6.768	.000	-6.180	-3.397	
	Certainty 1	-.524	.083	-.430	-6.285	.000	-.688	-.360	
	Severity*Certainty 1	.468	.110	.768	4.268	.000	.253	.684	
	Gender	.085	.193	.019	.438	.662	-.295	.464	

	Experience	.006	.011	.024	.544	.587	-.016	.028	
	AOT score	.177	.247	.032	.717	.474	-.308	.662	
Dependent Variable: Change in initial certainty									

Table 4.13. Regression model of change in initial certainty (Control versus Aided 2).

Linear Regression of change in initial certainty (Aided 1 versus Aided 2)									
Step		**Unstandardized Coefficients**		**Standardized Coefficients**	**t**	**Sig.**	**95.0% CI for B**		**R^2**
		B	**Std. Error**	**Beta**			**Lower Bound**	**Upper Bound**	
1	(Constant)	.721	.328		2.200	.028	.077	1.365	0.003
	Condition	-.213	.207	-.053	-1.027	.305	-.621	.195	
2	(Constant)	1.479	.320		4.623	.000	.850	2.108	0.140
	Condition	-.209	.193	-.052	-1.084	.279	-.588	.170	
	Severity	-1.500	.193	-.371	-7.778	.000	-1.880	-1.121	
3	(Constant)	4.897	.566		8.645	.000	3.783	6.010	0.247
	Condition	-.274	.181	-.068	-1.509	.132	-.630	.083	
	Severity	-3.907	.691	-.966	-5.652	.000	-5.266	-2.548	
	Certainty 1	-.557	.079	-.472	-7.054	.000	-.712	-.402	
	Severity*Certainty 1	.416	.107	.717	3.882	.000	.205	.627	
4	(Constant)	3.284	1.133		2.899	.004	1.056	5.512	0.249
	Condition	-.367	.185	-.092	-1.984	.048	-.730	-.003	
	Severity	-3.641	.703	-.913	-5.176	.000	-5.024	-2.258	
	Certainty 1	-.530	.082	-.455	-6.501	.000	-.691	-.370	
	Severity*Certainty 1	.377	.109	.656	3.448	.001	.162	.592	
	Gender	.039	.191	.010	.207	.836	-.336	.415	
	Experience	.014	.011	.060	1.304	.193	-.007	.035	
	AOT score	.349	.248	.065	1.406	.161	-.139	.837	
Dependent Variable: Change in initial certainty									

Table 4.14. Regression model of change in initial certainty (Aided 1 versus Aided 2).

Table 4.15. presents the difference in certainty across conditions and patient scenarios.

Patient scenario	Control Mean (SD)	Aided 1 Mean (SD)	Aided 2 Mean (SD)	Overall Mean (SD)
Chest pain (MSK steer)	1.43 (2.06)	0.76 (1.93)	0.94 (2.08)	1.05 (2.03)
Chest pain (cardiac steer)	-0.35 (1.61)	0.17 (1.50)	0.12 (1.61)	-0.03 (1.58)
Constipation (IBS steer)	1.40 (1.37)	1.51 (1.42)	1.59 (1.55)	1.50 (1.44)
Constipation (cancer steer)	-1.67 (1.85)	-0.45 (1.75)	1.71 (2.53)	-1.28 (2.14)
Overall	0.27 (2.12)	0.51 (1.78)	0.29 (2.24)	0.35 (2.06)

Table 4.15. Changes in certainty across conditions and patient scenarios.

4.6.5. Change in diagnosis (H_4)

Change in diagnosis was a dichotomous variable with possible values of 0 (No change) and 1 (Change). To qualify as a diagnostic change, the final diagnosis should have been different from the initial diagnosis. Similar to Study 2, GPs changed their diagnosis in 15.50% of the responses (i.e., 90/581). As expected, GPs in the Aided 2 condition changed their diagnosis more frequently than both in the Control and Aided 1 conditions, i.e., 18.90% (36/190) versus 18% (36/200) versus 9.40% (18/191) respectively. In two separate logistic regression models, we tested whether changes in diagnosis were associated with Condition. The first regression model revealed no significant difference between Control and Aided 2 conditions (OR = 1.032, [0.80, 1.33], p = 0.810), whereas the second regression model revealed a significant difference between Aided 1 and Aided 2 conditions (OR = 2.247, [1.23, 4.12], p = 0.009): diagnostic suggestions that did not contain GPs' initial diagnoses led to more frequent changes in diagnosis, compared to suggestions that included GPs' initial diagnoses.

When severity of initial diagnosis was added to the models, a significant positive relationship with change in diagnosis was detected in the first model, suggesting that the frequency of change was different for serious versus less serious initial

diagnoses: more serious initial diagnoses changed more frequently (Control versus Aided 2: OR = 1.932, [1.14, 3.28], p = 0.015). There was no interaction between condition and severity in any of the regression models.

We subsequently regressed initial certainty on change in diagnosis and found a significant association in both models: in line with Studies 1 and 2, GPs with higher certainty in their initial diagnosis were less likely to change it (Control versus Aided 1: OR = 0.632, [0.53, 0.73], p < 0.001 and Aided 1 versus Aided 2: OR = 0.639, [0.53, 0.76], p < 0.001). There were no interactions (see Tables 4.16. and 4.17).

Logistic regression of diagnostic change (Control versus Aided 2)										
Step		B	S.E.	Wald	df	Sig.	Exp(B)	95% CI.for Exp(B)		Nagelkerke R^2
								Lower	Upper	
1	Condition	.031	.131	.058	1	.810	1.032	.799	1.333	0.000242
	Constant	-1.516	.184	67.876	1	.000	.220			
2	Condition	.030	.132	.053	1	.819	1.031	.796	1.334	0.026
	Severity	.658	.270	5.954	1	.015	1.932	1.138	3.278	
	Constant	-1.883	.248	57.701	1	.000	.152			
3	Condition	-.041	.141	.087	1	.768	.959	.728	1.264	0.177
	Severity	1.023	.295	12.013	1	.001	2.781	1.560	4.959	
	Certainty 1	-.473	.082	33.515	1	.000	.623	.531	.731	
	Constant	.776	.503	2.377	1	.123	2.172			
4	Condition	-.005	.143	.001	1	.974	.995	.751	1.319	0.185
	Severity	.960	.298	10.373	1	.001	2.611	1.456	4.681	
	Certainty 1	-.476	.084	31.899	1	.000	.621	.526	.733	
	Gender	-.123	.296	.173	1	.678	.884	.496	1.578	
	Experience	.016	.016	1.076	1	.300	1.016	.986	1.048	
	AOT score	-.034	.374	.008	1	.929	.967	.464	2.014	
	Constant	.857	1.596	.288	1	.591	2.356			
Dependent Variable: Change in diagnosis										

Table 4.16. Regression model of diagnostic change (Control versus Aided 2).

Logistic regression of diagnostic change (Aided 1 versus Aided 2)										
Step		B	S.E.	Wald	df	Sig.	Exp(B)	95% CI.for Exp(B)		Nagelkerke R^2
								Lower	Upper	
1	Condition	.809	.309	6.854	1	.009	2.247	1.226	4.119	0.034
	Constant	-3.072	.529	33.760	1	.000	.046			
2	Condition	.811	.310	6.859	1	.009	2.250	1.226	4.128	0.039
	Severity	.308	.299	1.060	1	.303	1.361	.757	2.447	
	Constant	-3.240	.557	33.801	1	.000	.039			
3	Condition	.753	.322	5.484	1	.019	2.124	1.131	3.991	0.158
	Severity	.547	.317	2.976	1	.084	1.728	.928	3.216	
	Certainty 1	-.447	.091	24.431	1	.000	.639	.535	.763	
	Constant	-.675	.738	.835	1	.361	.509			
4	Condition	.860	.327	6.921	1	.009	2.364	1.245	4.488	0.177
	Severity	.454	.322	1.980	1	.159	1.574	.837	2.960	
	Certainty 1	-.469	.096	23.763	1	.000	.625	.518	.755	
	Gender	.077	.336	.052	1	.819	1.080	.559	2.085	
	Experience	-.020	.019	1.069	1	.301	.981	.945	1.018	
	AOT score	-.332	.443	.562	1	.453	.718	.301	1.708	
	Constant	.851	1.926	.195	1	.659	2.341			
Dependent Variable: Change in diagnosis										

Table 4.17. Regression model of diagnostic change (Aided 1 versus Aided 2).

Table 4.18. presents the difference in frequency of changes in diagnosis across conditions and patient scenarios.

Patient scenario	Control	Aided 1	Aided 2	Overall
Chest pain (MSK steer)	11.10% (6/54)	7.80% (4/51)	21.20% (11/52)	13.40% (21/157)
Chest pain (cardiac steer)	9.70% (6/62)	3.40% (2/59)	3.40% (2/59)	5.60% (10/180)
Constipation (IBS steer)	8.90% (4/45)	7.00% (3/43)	12.20% (5/41)	9.30% (12/129)

Constipation (cancer steer)	51.30% (20/39)	23.70% (9/38)	47.40% (18/38)	40.90% (47/115)
Overall	18.00% (36/200)	9.40% (18/191)	18.90% (36/190)	15.50% (90/581)

Table 4.18. Frequency of changes in diagnosis across conditions and scenarios.

In line with Studies 1 and 2, change in diagnosis was negatively associated with change in certainty: GPs who increased their certainty in their initial diagnosis more were less likely to change their diagnosis (Control versus Aided 2: OR =,0.359, [0.28, 0.46] $p < 0.001$ and Aided 1 versus Aided 2: OR = 0.396, [0.31, 0.51], $p < 0.001$).

4.6.6. Exploratory analyses

4.6.6.1. Initial certainty (H5)

On average, GPs' certainty in their initial diagnosis was 6.31 (SD 1.72) on the 0-10 scale across the three conditions. GPs in Control condition had the highest certainty (i.e., 6.38, SD 1.77) compared to Aided 1 condition (6.35, SD 1.66) and Aided 2 condition (6.18, SD 1.74). None of these differences were significant (Control versus Aided 2: $t = 1.39$, $p = 0.673$ and Aided 1 versus Aided 2: $t = 1.278$, $p = 0.631$). The ratings of initial certainty across scenarios are presented in Table 4.19.

In Study 2, initial certainty was found to be the main driver of the process that affected all the downstream variables (i.e., perceived support, number of cues requested and diagnostic changes) in the expected direction. In the present study we expected the same results and hypothesised that GPs with higher certainty in their initial diagnosis would perceive the neutral clinical cues as more supportive, request fewer cues and change their initial diagnosis less frequently than GPs with lower certainty. Initial certainty was included in all the regression models and confirmed our expectations for all three dependent variables.

Patient scenario	Control Mean (SD)	Aided 1 Mean (SD)	Aided 2 Mean (SD)	Overall Mean (SD)
Chest pain (MSK steer)	5.67 (1.70)	6.29 (1.66)	6.02 (2.08)	5.99 (1.68)
Chest pain (cardiac steer)	7.23 (1.63)	7.17 (1.54)	7.08 (1.61)	7.16 (1.51)
Constipation (IBS steer)	6.27 (1.34)	5.74 (1.42)	5.71 (1.55)	5.91 (1.54)
Constipation (cancer steer)	6.13 (1.99)	5.84 (1.64)	5.24 (2.53)	5.74 (1.85)
Overall	6.38 (1.77)	6.35 (1.66)	6.13 (1.76)	6.29 (1.73)

Table 4.19. Initial certainty per scenario and condition.

4.6.6.2. Final certainty (H6)

Across conditions and patient scenarios, GPs' average certainty in their final diagnosis was 7.04 (SD 1.67). GPs who did not receive diagnostic suggestions had the highest final certainty (7.19, SD 1.50), compared to Aided 1 condition (6.97, SD 1.67) and Aided 2 condition (6.95, SD 1.82). None of these differences was significant (Control versus Aided 2: b = -0.116, [-0.28, 0.05], p = 0.169 and Aided 1 versus Aided 2: b = -0.021, [-0.37, 0.33], p = 0.906).

When Initial certainty was included in the models, it was found to be associated with final certainty in both models (Control versus Aided 2: b = 0.541, [0.46, 0.62], p < 0.001 and Aided 1 versus Aided 2: b = 0.564, [0.48, 0.65], p < 0.001), suggesting that GPs with higher initial certainty were also high in their final certainty. There were no interactions. Table 4.20. presents final certainty ratings across scenarios and conditions.

Patient scenario	Control Mean (SD)	Aided 1 Mean (SD)	Aided 2 Mean (SD)	Overall Mean (SD)
Chest pain (MSK steer)	7.31 (1.59)	6.98 (1.85)	7.46 (1.49)	7.25 (1.65)
Chest pain (cardiac steer)	7.18 (1.48)	7.61 (1.13)	7.32 (1.26)	7.37 (1.31)
Constipation (IBS steer)	7.69 (1.14)	7.21 (1.26)	7.12 (1.60)	7.35 (1.36)
Constipation (cancer steer)	6.44 (1.53)	5.71 (1.89)	5.50 (2.42)	5.89 (2.00)
Overall	7.19 (1.50)	6.97 (1.67)	6.95 (1.82)	7.04 (1.67)

Table 4.20. Final certainty per scenario and condition.

4.6.6.3. Number of differential diagnoses

The number of differential diagnoses corresponds to the number of alternative hypotheses that GPs recorded for each patient, after they provided their final diagnosis. On average, GPs recorded 1.27 differential diagnoses per scenario version (range 0 to 8 diagnoses, median 1). GPs in the Aided 2 condition gave a higher average number of differential diagnoses than GPs in the other conditions (Control: 1.26, Aided 1: 1.20, Aided 2: 1.34). The number of differential diagnoses differed by scenario version (Table 4.21.).

Patient scenario	Control Mean (SD)	Aided 1 Mean (SD)	Aided 2 Mean (SD)	Overall Mean (SD)
Chest pain (MSK steer)	1.07 (0.91)	0.96 (0.92)	1.15 (0.98)	1.06 (0.93)
Chest pain (cardiac steer)	0.92 (0.97)	1.00 (0.95)	1.19 (0.94)	1.03 (0.95)
Constipation (IBS steer)	1.56 (1.18)	1.21 (1.12)	1.39 (1.05)	1.39 (1.12)
Constipation (cancer steer)	1.72 (1.43)	1.82 (1.18)	1.76 (1.48)	1.77 (1.36)

Overall	1.26 (1.15)	1.20 (1.07)	1.34 (1.11)	1.27 (1.11)

Table 4.21. Number of differential diagnoses that GPs recorded by scenario and scenario version.

The regression models revealed no association between condition and number of differentials diagnoses that GPs provided after their final diagnosis (Control versus Aided 2: IRR = 0.943, [0.79, 1.12], p = 0.513 and Aided 1 versus Aided 2: IRR = 0.897, [0.75, 1.07], p = 0.232). Severity of initial diagnosis was added to the models and not found to be associated with the number of differentials diagnoses (Control versus Aided 2: IRR = 0.967, [0.81, 1.15], p = 0.709 and Aided 1 versus Aided 2 IRR = .854, [0.71, 1.02], p = 0.084). There was no interaction between severity and condition in any of the regression models.

We also controlled for GPs' initial certainty and found that it was negatively associated with the number of differential diagnoses in both models. This suggests that GPs who indicated higher certainty in their initial diagnosis recorded fewer differential diagnoses compared to their less certain counterparts (Control versus Aided 2: IRR = 0.917, [0.87, 0.96], p < 0.001 and Aided 1 versus Aided 2: IRR = 0.903, [0.86, 0.95], p < 0.001). There were no interactions.

4.6.6.4. Appropriateness of management

On average, GPs management was deemed appropriate in 13% of the cases (75/581). There was a small difference between conditions (Control: 10% versus Aided 1: 16% versus Aided 2: 14%) which was not significant. Separate logistic regression models were used to test associations with predictors and covariates. No association was found.

4.6.6.5. Gender

As in Study 2, gender was found to be associated with initial certainty: male GPs were more certain in their initial diagnosis than their female counterparts (6.60, SD 1.65 versus 6.04, SD 1.75). This result was significant in both regression models

(Control versus Aided 2: b = 0.465, [0.11, 0.82], p = 0.0011 and Aided 1 versus Aided 2: b = 0.667, [0.32, 1.02], p < 0.001).

4.6.6.6. Experience

As with gender, all regression models tested for associations between the variables of interest and GPs' professional experience. No association was found in any of the regression models.

4.6.6.7. Actively Open-minded Thinking

GPs' mean AOT score was 3.94 (SD 0.39) on the 1-5 scale. AOT score was included as a covariate in all regression models. We found the following:

Similar to Study 2, there was a positive association between AOT scores and the number of cues that GPs requested, in one of the regression models (Control versus Aided 2: IRR = 1.217, [1.07, 1.39], p = 0.003) suggesting that GPs with higher scores on the AOT scale requested more information about the patient before reaching a final diagnosis. In terms of perceived cue support, there was a positive association with AOT scores in one of the models (Aided 1 versus Aided 2: b = 0.519, [0.015, 1.02], p = 0.043): in contrast to what we expected, GPs with higher scores in AOT rated the cues as more supportive of their initial diagnosis.

We also found a negative association between AOT scores and initial certainty in one of the regression models (Control versus Aided 2: b = -0.512, [-0.97, -0.06], p = 0.028), suggesting that GPs high in AOT had lower certainty in their initial diagnosis. The opposite was true for final certainty, in one of the regression models (Control versus Aided 2: b = 0.427, [0.07, 0.78], p = 0.019): GPs high in AOT had higher certainty in their final diagnosis. Finally, and in line with Study 2, we found that GPs with high AOT scores recorded a greater number of differential diagnoses compared to GPs with lower AOT scores in one of the regression models (Control versus Aided 2: IRR = 1.429, [1.12, 1.82], p = 0.004).

4.6.7. Attitudes towards the decision aid

As mentioned above, after completing the diagnostic task and AOT scale, GPs were presented with a short questionnaire about the decision aid. Specifically, GPs were asked whether they found the diagnostic suggestions helpful, whether they would trust the suggestions from such a decision aid, and whether they would like to have a similar aid in their practice. In the first question about helpfulness of the suggestions, 147 of the 248 GPs who took part in the study (59.30%) responded that they found the suggestions helpful ("definitely" or "probably"). Likewise, 148 of the 248 GPs (59.70%) said that they would trust the suggestions from such a decision aid. Finally, 140 of the 248 GPs (56.50%) said that they would like to have a similar decision aid in their practice.[19]

4.7. Discussion

The present study was a follow-up to Study 2, in which GPs assessed two patient scenarios with and without diagnostic suggestions in a within-participant design. The suggestions included the most likely diagnosis for each scenario. No significant differences were detected between experimental conditions. The present study aimed to explore one possible reason for the lack of impact of the diagnostic suggestions on GP behaviour; namely that GPs used the diagnostic suggestions to validate their own, initial diagnosis, rather than to question in.

Study 3 differed from Study 2 in two respects: 1) GPs saw only one patient scenario rather than two, and always with diagnostic suggestions; 2) the list of diagnostic suggestions did not include the GPs' initial diagnoses. There was no way of knowing in advance which diagnosis GPs would initially give, thus, we determined the most likely diagnosis for each scenario version based on the information initially provided and removed it from the list. We then excluded from the analyses those GPs who did not provide this diagnosis.

[19] For a recent, detailed, qualitative analysis of physicians attitudes towards computerised decision aids see McParland, Cooper and Johnston (2019).

In this study, we made two separate comparisons: a) a comparison between GPs' responses given without seeing diagnostic suggestions (Control condition) and responses given after seeing the full list of diagnoses (including the most likely diagnosis – Aided 1 condition) and b) a comparison between GPs' responses given after seeing the full list (Aided 1 condition) and responses given after seeing the reduced list that did not include the most likely diagnosis (Aided 2 condition). Overall, the study replicated the results of Study 2; only trends were detected and no significant differences between conditions. Only one significant difference was detected, which was in the expected direction. Specifically, GPs who saw the reduced list changed their mind more frequently than those who saw the full list.

Similar to Study 2, the main driver that affected the variables of interest, and always in the expected direction, was initial certainty. It was found to be positively related to perceived support and negatively related to the amount of clinical information that GPs requested, the frequency of diagnostic changes and the number of alternative diagnoses that GPs recorded in addition to their final diagnosis. These relationships were observed in both pairs of comparisons, suggesting that initial certainty affects clinical reasoning irrespective of diagnostic suggestions.

One of the limitations of the present study was GPs' moderate level of initial certainty. Specifically, the initial information in each patient scenario was designed to steer GPs toward a specific diagnosis with relatively high certainty. However, despite the fact that this information was diagnostic, and that most GPs accepted the steers and provided the expected initial diagnoses, average initial certainty was rather moderate. This could indicate that GPs were already sceptical about their initial diagnosis, given that the information provided was not sufficient for a definite diagnosis. As explained in Study 2 (see Chapter 3: Discussion), this could have undermined the underlying hypothesis that first impressions are formed intuitively and are often accompanied by high certainty. In turn, this could have weakened any (de-biasing) effects of the diagnostic suggestions on GPs' thinking.

Another limitation was that approximately 25% of GPs responses were excluded from the analysis, due to not having provided the expected diagnosis in each scenario version, reducing, thus, the power of the study. The expected diagnoses

had been removed from the lists of diagnostic suggestions (reduced lists) before data collection; it was therefore important that GPs were included in the analyses only when they gave the expected diagnoses at the start of the scenarios.

The present study had no control condition: GPs saw only one patient scenario which was always presented with the reduced list. This reduced the duration of the online task but did not allow us to make direct comparisons between the three conditions in one single model (e.g., ANOVA). This was because of some of the data were dependent, since they came from the same participants (i.e., responses in the Control and Aided 1 conditions in Study 2), while other data were independent (i.e., responses from the Aided 2 condition in Study 3). As a result, we carried out a separate regression model for each pair of conditions.

We observed high variability in GPs' perceived support across patient scenarios, which could have prevented between-conditions differences to emerge. The baseline ratings of GPs' perceived support of cues in the Control condition varied widely both between and within scenarios (i.e., between versions of the same scenario). In the chest pain scenario, GPs evaluated the cues as more supportive of MSK than a cardiac diagnosis (6.32 versus 4.94 on a 0-10 scale). Likewise, in the constipation scenario, GPs found more support for IBS rather than for cancer (7.29 versus 3.16). These findings suggest that, despite the fact that the clinical cues were designed to be neutral, the diagnosticity of the information that the cues provided differed across the steers. It is possible that the severity, as well as the familiarity of diagnosis (e.g., availability heuristic), had an impact on GPs' evaluation of information, who found more support for common and less serious diagnoses (i.e., MSK and IBS) than for uncommon and more serious diagnosis (i.e., angina, cancer). There was also high variability in perceived support within the same scenario version/steer. For instance, in the cardiac steer of the chest pain scenario, the self-treatment cue ("David says that he has not taken any pain relief so far. He does not like taking tablets.") had a rating of 2.94, whereas the pain intensity cue ("David says that it is not really intense, but it does make him want to stop what he is doing") had a rating of 6.05. Although they were designed to be neutral, the findings suggest that different GPs perceived high diagnostic value in different cues and weighed them accordingly.

In line with Studies 1 and 2, the present study did not yield any different results, failing to show any effects of diagnostic suggestions on physicians thinking. However, it provided some evidence that a full list may enable GPs to validate their initial diagnosis if they see it on the list, thereby increasing their certainty even more than controls. We could also say that including the most likely diagnosis in the list of suggestions may lead to close-mindedness, as the full list (Aided 2 condition) appeared to undermine, rather than to help, reasoning, given that it led to a) more biased evaluations, b) greater increase in initial certainty, c) fewer diagnostic changes and d) almost the same (fewer though) differential diagnoses, compared to controls. Removing this diagnosis from the list appears to remove this effect but the reduced list does not provide additional improvement beyond the baseline (controls).

There are several avenues for expanding this work. For instance, future research could employ different, more realistic patient scenarios with richer information that would probably induce higher initial diagnostic certainty. This would be particularly important, as all three studies of this thesis showed that initial certainty was the main driver that affected physicians' reasoning. Initial diagnostic certainty could also be increased with priming (e.g., by asking physicians to think of accurate diagnoses or successful treatments they made in the past). Future studies could also use different information to be provided as neutral clinical cues, assessing its diagnosticity more systematically (e.g., through a separate control group that provides baseline ratings without being introduced to a specific patient scenario, or without having to provide a diagnosis – also see Nurek, Kostopoulou and Hagmayer, 2014).

Another way to expand this work is to test the effects of diagnostic suggestions on a different population (e.g., medical specialists or consultants), to test whether expertise plays a role in the process. This could also be done using a more interactive decision aid that allows users' input and provides additional information, such as risk scores, prevalence of diseases, and/or likelihood of diagnostic error (as opposed to only a list of suggestions). In a study design similar to the studies of this thesis, this information could be separately used as decision aids and tested in different experimental conditions to identify how it affects reasoning (e.g., one providing risk scores, another providing likelihood estimates and so on). It would be particularly interesting to test which type of information affects the aspects we

investigated in this study (e.g., diagnostic certainty, information search, information interpretation and diagnostic change). Practice sessions prior to the actual experiment could also help physicians familiarise themselves with such decision aids, minimising, possibly, any scepticism about their credibility.

Treatment recommendations, as opposed to diagnostic suggestions could also be an avenue for future research. A decision aid that provides management suggestions instead of diagnostic ones, could show different results. In fact, this could be tested in a study similar to the ones described in this thesis, to see whether physicians would be more likely to change their treatment decisions when presented with alternative management options. Justifications for and against the initial hypothesis, as well as reasons to support alternative hypotheses could also be included as a corrective strategy to potentially biased clinical judgements (e.g., Mussweiler and colleagues, 2000 and Arkes and colleagues, 1988).

The majority of medical studies measure diagnostic certainty on numerical scales (e.g., 0-100%). However, some physicians' feedback from the present study indicates that assigning certainty scores to a diagnosis often feels unnatural. The same applies to probability estimations. Taken together, future research could consider either a) using certainty scales with only textual anchors instead of numerical values, or b) measuring certainty indirectly, by rephrasing the question about it (e.g., likelihood of seeking advice as a proxy for certainty – see Friedman *et al.*, 2001, 2005) or even c) introducing the decision aid as advice from colleagues, rather than from a computerised source (see Medow, Arkes and Shaffer, 2010). Other methods such as what-if analysis and decision trees (e.g., Melone *et al.*, 1993; Gail, 2000; Morgan *et al.*, 2000), as well as the notable "pre-mortem" (i.e., prospective hindsight, see Klein, 2007 and Kahneman, 2011)[20] could be implemented in decision aids and used as debiasing techniques.

[20] The pre-mortem is a debiasing technique that could help organisations or individuals to minimise the possibility of failure: in the process of arriving to a decision or planning a project, the pre-mortem calls the decision makers to assume that a given decision was made and led to a bad outcome (e.g., launching a product that did not perform well in the market or following a medical treatment that caused harm). The decision makers are then asked to find reasons why this happened, in an attempt to prevent failure. This technique can help organisations to improve their decision making "pre-mortem"- i.e., pre-decisionally, by considering factors that could undermine the success of a plan (Klein, 2007, Kahneman, 2011).

Finally, future research could use different methodology. This could include interviews with physicians to explore their insights about decision aids, as well as think aloud protocols while interacting with a decision aid. Towards this direction, eye-tracking methodology could also contribute to the better understanding of how physicians interact with decision aids. Exploring physicians' visual behaviour could provide insights about information processing during the diagnostic process. This could include the number of fixations on certain information (e.g., do physicians focus on confirmatory information/their own diagnosis when presented with a list of suggestions?), the amount of time spent on reading the information provided, and/or fixation sequence (e.g., do they read all the information, and if so, which piece of information they start with?). In fact, the use of eye-tracking methodology was the initial plan for the present study. However, due to the COVID-19 pandemic, we were not able to conduct a lab-based study and, as a result, we opted for an online one.

Summary

- 248 GPs assessed one online patient scenario, always with suggestions
- Diagnostic suggestions (versus control) led to more changes in diagnosis
- Physicians with higher initial certainty requested fewer cues, had more biased evaluations, changed diagnosis less frequently, and recorded fewer differential diagnoses
- Physicians who provided a less serious (versus serious) initial diagnosis were more biased in their evaluations and changed diagnosis less frequently
- Physicians with higher AOT score were more biased in their evaluations, requested more cues and recorded more differential diagnoses

CHAPTER 5: Initial confidence in a judgement and changing of mind: a systematic review

5.1. Introduction

This systematic review was an attempt to capture the existing empirical work on the relationship between decision confidence and resistance to changing one's decision or judgement. It was also an attempt to expand the consistent finding from Studies 1, 2 and 3, that higher confidence in an initial hypothesis predicts the likelihood of change of the final judgement, beyond the field of medical decision making.

5.1.1. Theoretical and empirical background

It is almost self-evident that beliefs held with greater confidence are more resistant to change. The first studies that inspired later studies to investigate this association were conducted by James Stoner and Michael Wallach and colleagues in the early 1960s. Stoner's study was an unpublished master thesis in social psychology. The study was mainly focused on the differences between collective and individual decision making, and the effects of group interaction on individuals' risky shifts of opinion (Stoner, 1961). The risky shift phenomenon refers to people's tendency to change their individual opinions to riskier ones when they are acting as members of a group (Stoner, 1961; Bell and Jamieson, 1970; Burnstein and Katz, 1971). In Stoner's study (1961), participants were provided with 12 binary-choice questions. One choice was always riskier than the other (e.g., risky, high-return investments versus secure, low-return investments). Participants were asked to indicate their choice and their confidence in it, first individually and then as members of a group. The study found that groups tended to make riskier decisions compared to individuals (i.e., the risky shift phenomenon). However, confidence was not related to the riskiness of the choice (Stoner, 1961).

The risky shift phenomenon was subsequently explored by Wallach and colleagues during the 1960s. In a series of studies, Wallach and colleagues tried to explain the risky shifts by attributing them to the diffusion of responsibility (Wallach, Kogan and

Bem, 1963). Specifically, they suggested that individuals feel less responsible when they make decisions as members of a group, since the decisions are made collectively. As a result, they are more willing to change their initial decisions toward a riskier option. A similar explanation about the impact of accountability on reducing overconfidence and resistance to change was also proposed by Jermias (2006). Back in Wallach and colleagues series of studies, only one study tested the association with decision confidence and found no relationship between confidence and risky shifts (Wallach, Kogan and Bem, 1963), while the rest of the studies did not account for participants' confidence in their choice (Wallach, Kogan and Bem, 1962; Wallach, Kogan and Burt, 1965). The results of these studies are, therefore, beyond the scope of this review.

In the early 1970s, another two studies by Bell and Jamieson (1970) and Burnstein and Katz (1971) investigated the risky shift phenomenon and its links to people's confidence in their choices. The former study was similar to Stoner's 1961 study with slightly different results: despite the fact that risky shifts were not related to participants' confidence, a sub-group analysis (i.e., a comparison between participants with more than one standard deviation above or below the mean confidence) showed that lower confidence led to more risky shifts (Bell and Jamieson, 1970). The latter study used a different methodology. Instead of a binary-choice questionnaire, Burnstein and Katz (1971) used a reaction time task where participants had to toggle a switch as fast as they could. They were also asked to predict a reaction time interval for their performance in case they were able to participate again, and to rate their confidence in their prediction. Finally, they indicated their intention to change their choice if they had the opportunity to do so. The results showed that participants who provided a shorter interval (i.e., a riskier choice) were more confident in their choice and less likely to change it if they could (Burnstein and Katz, 1971). Group time reaction intervals were shorter, and therefore riskier, than the average time interval provided by individuals.

Moving from group to individual decision making, and from changing to adjusting opinion, Kahneman and Tversky (1974) introduced the anchoring and adjustment heuristic, suggesting that people's judgements, or current beliefs, are influenced by initial values (i.e., anchors) or prior beliefs (Kahneman and Tversky, 1974, also see

Chapter 1: Debiasing strategies). The subsequent adjustment of the current belief is mostly determined by the initial anchor, as they suggested. Einhorn and Hogarth (1992) investigated anchoring and adjustment further and proposed a model of "belief revision". They distinguished their model from Bayesian hypothesis testing by highlighting people's tendency to evaluate information relative to a single hypothesis (the hypothesis under consideration), as it is cognitively easier, rather than to a multiple one (i.e., the hypothesis under consideration versus its alternative), which is cognitively more demanding. They further suggested that the magnitude of belief adjustment is largely influenced, not only by the size of the initial anchor, but also by the order in which the subsequent pieces of information are presented. Specifically, they proposed that if the pieces of information are evaluated individually (i.e., one by one), the last pieces in the sequence will weigh more in the adjustment of the belief (recency effect). This effect is more prominent when the information is not consistent (e.g., confirmatory versus contradicting information). In contrast, if the information is evaluated as a whole, then the first pieces of information will have greater impact on belief (primacy effect). However, this proposed order effect seems to disappear when the information is consistent (Hogarth and Einhorn, 1992).

More recently, Daniel Kahneman famously distinguished System's 1 fast and intuitive thinking from System's 2 deliberate and slow thinking (Kahneman, 2011). Kahneman suggested that System 1 seeks cognitive ease and is, therefore, rushed to jump to conclusions based only on little information. Given that System 1 is averse to cognitively demanding tasks, evaluating the quality of information is not its primary goal. It is rather the coherence of the story that System 1 seeks, in order to reach a conclusion, as Kahneman suggested. Any conflicting information that could undermine the coherence of the story and the associative confidence in the initial judgement could be easily disregarded by System 1, in an attempt to maintain the status quo with minimum effort (Kahneman, 2011). It is also the coherence of the story that creates to people an illusion that their judgements are valid, and makes them more confident in their judgements, and therefore, less likely to revise them (Kahneman and Tversky, 1974).

The effects of coherence-seeking and the illusion of validity on judgements formation and revision were further investigated by other researchers. In a series of

experiments in legal decision making, Simon and colleagues (2001) tested whether coherence-seeking can affect the evaluation of conflicting arguments and the associative inferences (Simon *et al.*, 2001). Specifically, in their experiments they asked participants to assess a hypothetical legal case (i.e., a lawsuit) between two parties and indicate their initial leaning, as well as their final judgement (i.e., the verdict). In brief, participants were asked to evaluate a set of arguments that contained conflicting information about the two parties (e.g., party A claimed that the motive of the act was X while party B claimed that the motive of the act was Y), first independently (i.e., out of context), without being introduced to the legal case (baseline), and then contextually, after being introduced to the case. The results showed that participants changed their baseline evaluations in the direction that fitted with their verdict, suggesting that the coherence of a story (and the associative confidence in a judgement) can be assured, even in situations that require evaluation of conflicting information (Simon *et al.*, 2001).

Apart from the coherence of the story that intuitive thinking constructs, it is also the sense of correctness that boosts people's confidence in their judgements. This sense of correctness was explained as a metacognitive process by Valerie Thompson (Thompson, 2009). In her metacognitive framework of reasoning, Thompson (2009) suggested that intuitive judgements are often followed by a feeling of rightness (an analogy to subjective confidence) which will determine whether subsequent processing is required. Higher feeling of rightness (FOR) allows little space for System's 2 engagement, as it creates what Kahneman and Tversky called an *illusion of validity* (Kahneman and Tversky, 1974). The link between FOR and System's 2 analytical engagement was extensively investigated in various studies by Thompson and colleagues (2011, 2013, 2014, 2019). These were studies mainly on conditional or syllogistic reasoning (e.g., if p then q), where participants were asked first to make a quick, intuitive judgement about the validity of conclusions and then a more deliberative one, without time constraints. In all studies, Thompson and colleagues found a consistent pattern: lower FOR (i.e., lower confidence in an initial judgement) was associated with longer rethinking times and more frequent revisions of initial judgements, which indicates some degree of System 2 analytical processing (Thompson, Prowse Turner and Pennycook, 2011; Thompson, Ackerman, *et al.*, 2013; Thompson, Evans and Campbell, 2013; Wang and Thompson, 2019).

5.1.2. Objectives and links to Studies 1, 2 and 3

The empirical studies conducted for the present research work were designed to measure confidence in an initial diagnostic judgement based on limited information. This was an attempt to capture the very first, intuitive thoughts that come to physicians' mind, as soon as they are being presented with a patient case. The studies also measured whether physicians change their initial diagnostic judgement after being presented with a list of diagnostic suggestions as a decision aid. Overall, the three studies did not detect any significant effect of the diagnostic suggestions on physicians' reasoning. However, it is important to note that all three studies showed a consistent result; that is, physicians' confidence in their initial diagnostic judgement was negatively associated with the likelihood of changing this judgement later, with more information.

The consistent result found in the three studies of this research work is in line with the work conducted by Thompson and colleagues (2011, 2013, 2014, 2019), as well as with a number of studies presented in this review. Furthermore, this finding adds to previous research and highlights the importance of the first (intuitive) judgements for the final decision (Kahneman and Tversky, 1974; Hogarth and Einhorn, 1992; Stone, 1994; Ames *et al.*, 2010; Kahneman, 2011; Kostopoulou, Sirota, *et al.*, 2017). Assuming that first impressions are generated almost automatically as a result of intuitive thinking, rather than of analytical processing, they are potentially fallible (Kahneman and Tversky, 1974; Einhorn and Hogarth, 1978; Kahneman, 2011). And given that first impressions may be accompanied with high confidence, it would be important to know whether there is, indeed, an association between confidence and resistance to decision/judgement change. Consequently, the present review aimed to systematically assess the evidence for the relationship between confidence in an initial judgement and the likelihood of changing/updating this judgement. This will allow for conclusions about the robustness of this effect (if any) and could inform future interventions that aim to induce decision change.

5.2. Methods

A study protocol outlining the elements of the review was developed. The elements of the protocol are described below. The articles captured by the search were reviewed in three phases: title and abstract screening of all the articles, full-text review of potentially eligible articles and, finally, data extraction of eligible articles. Quality assessment of the eligible articles was also conducted. A sample of 10% of randomly selected articles was assessed by a second reviewer (Dr Olga Kostopoulou).

5.2.1. Search strategy

The search was conducted across five electronic databases: Web of Science, Scopus, Embase, Medline, and PsycINFO. In addition to the database searches, the first 20 pages of Google scholar web engine, as well as reference lists of relevant articles were reviewed. The search strategy was developed relative to the two components of the research question: “Decision confidence” and “Change of judgement”. Appropriate search terms and their synonyms were used to capture articles related to the research question. Table 5.1. presents the search terms used for each component in one of the five databases (Web of Science).

Decision confidence	Change of judgement	
Confiden* Certainty Overconfiden* “Feeling of rightness” “Illusion of validity”	belief* NEAR/2 updat* belief* NEAR/2 revis* belief* NEAR/2 chang* belief* NEAR/2 adjust* belief* NEAR/2 persever* belief* NEAR/2 persist*	opinion* NEAR/2 updat* opinion* NEAR/2 revis* opinion* NEAR/2 chang* opinion* NEAR/2 adjust* opinion* NEAR/2 persever* opinion* NEAR/2 persist*

Overestimat* “Degree of belief” “Strength of belief”	belief* NEAR/2 shift* answer* NEAR/2 updat* answer* NEAR/2 revis* answer* NEAR/2 chang* answer* NEAR/2 adjust* answer* NEAR/2 persever* answer* NEAR/2 persist* answer* NEAR/2 shift* judg$ment* NEAR/2 updat* judg$ment* NEAR/2 revis* judg$ment* NEAR/2 chang* judg$ment* NEAR/2 adjust* judg$ment* NEAR/2 persever* judg$ment* NEAR/2 persist* judg$ment* NEAR/2 shift* mind NEAR/2 chang* resist* NEAR/2 chang* reluctan* NEAR/2 chang* reconsider* rethink*	opinion* NEAR/2 shift* decision* NEAR/2 updat* decision* NEAR/2 revis* decision* NEAR/2 chang* decision* NEAR/2 adjust* decision* NEAR/2 persever* decision* NEAR/2 persist* decision* NEAR/2 shift* diagnos$s NEAR/2 updat* diagnos$s NEAR/2 revis* diagnos$s NEAR/2 chang* diagnos$s NEAR/2 adjust* diagnos$s NEAR/2 persever* diagnos$s NEAR/2 persist* diagnos$s NEAR/2 shift* "ego bias" "belief bias" "primacy effect" "perseverance effect" "first impression*"

Table 5.1. Main components of the search and search terms used in Web of Science database.

Wildcards and operators were adjusted according to the syntax of each database. Restrictions and limits were applied to meet the type of source and the language requirements. The results of the search were last updated on the 20th of June 2020.

5.2.2. Inclusion and exclusion criteria

Only articles in English were included. There was no limit for the year of publication. Study participants had to be healthy adults. Only primary research was included (i.e., no reviews, editorials, or opinion pieces). We only included studies that used quantitative methods. Initial confidence in a decision or judgement, as well as change in initial decision or judgement, had to be among the outcome measures. Studies that measured confidence only in a final judgement were excluded. Furthermore, studies that measured confidence in skills, knowledge, attitudes, or future performance were excluded. Studies that measured willingness to change judgement rather than actual change were also excluded.

A decision-making task was defined as any task where participants had to make a decision or judgement, including forecasts, that required at least two responses (e.g., an initial and a final decision). General knowledge (e.g., multiple choice quizzes) and memory or recognition tasks (e.g., identifying the direction of moving dots) were excluded.

5.2.3. Reference management

The Mendeley platform (version 1.17.12) was used to manage the references and remove duplicates. The records were organised in Microsoft Excel spreadsheets for screening, data extraction and quality assessment of the eligible studies.

5.2.4. Selection and data collection process

The initial screening for eligibility (i.e., titles and abstracts) was conducted by one researcher (Ploutarchos Kourtidis). Potentially eligible articles were included for second screening (i.e., full-text review). Throughout the second screening, the second reviewer (Dr Olga Kostopoulou) assessed a sample of 10% of the eligible articles. Any disagreements were resolved through discussion. A data extraction form was developed in line with the eligibility criteria, to extract data from the eligible articles. Only the most relevant information from each article was extracted, including journal, participants, sample size, task, type of intervention, measurement scales and results.

5.2.5. Quality assessment

The methodological quality of the eligible articles was assessed through the QATSDD tool for studies with diverse designs (Sirriyeh *et al.*, 2012). The original tool consists of 16 criteria to assess different aspects of quantitative and qualitative studies including theoretical framework, statement of objectives and discussion of limitations. However, the tool was adapted to the needs of the present review. Specifically, only five of the 16 criteria were used to assess the methodological quality of each article, as well as any sample size considerations: 1. Clear description of research setting, 2. Evidence of sample size considered in terms of analysis, 3. Representative sample of target group of a reasonable size, 4. Description of procedure for data collection and 5. Fit between research question and method of data collection. For each criterion, the articles were awarded a quality score that ranged from 0 to 3, which was averaged across all the articles to provide an overall score of the evidence. The quality score was expressed as a percentage of the maximum possible score.

5.3. Results

The search undertaken across the five electronic databases yielded the following results:

- Web of Science (2324 articles)
- Scopus (3,656 articles)
- Embase (646 articles)
- Medline (732 articles)
- PsycINFO (282 articles).

The total number of potentially eligible articles was 7,640. After removing duplicates, this number was reduced to 4,322. A further seven articles were included for titles and abstracts screening, after assessing reference lists from relevant articles. From the 4,322 articles that were initially screened, 4,130 were excluded for various reasons (see Figure 5.1.). The main reason for excluding the majority of the articles (i.e., 3,799) was that they were reporting “confidence intervals” in their abstract, without measuring confidence in a decision or judgement. For the remaining 199 articles, full texts were obtained.

After full-text screening, 41 eligible articles were included for data extraction. 38 articles came from the databases search and three from reference lists of relevant articles. The reasons that those three articles were not captured by the search was that the second component of the search strategy (i.e., change of judgement) was either not mentioned in their abstracts, or was described in different terms (e.g., openness to persuasion). From those 41 eligible articles, 21 studies were in the medical domain. 27 studies (i.e., 65.80%) provided evidence for an association between confidence in an initial decision or judgement and change of this judgement (i.e., higher confidence was associated with greater resistance to change), while five studies (i.e., 12.20%) supported the above hypothesis only partially. Overall, 78% of the studies found at least some evidence for the relationship between initial confidence and resistance to change. Finally, the remaining nine studies (i.e., 21.95%) did not find any evidence for this relationship. Below is the PRISMA flowchart of the review:

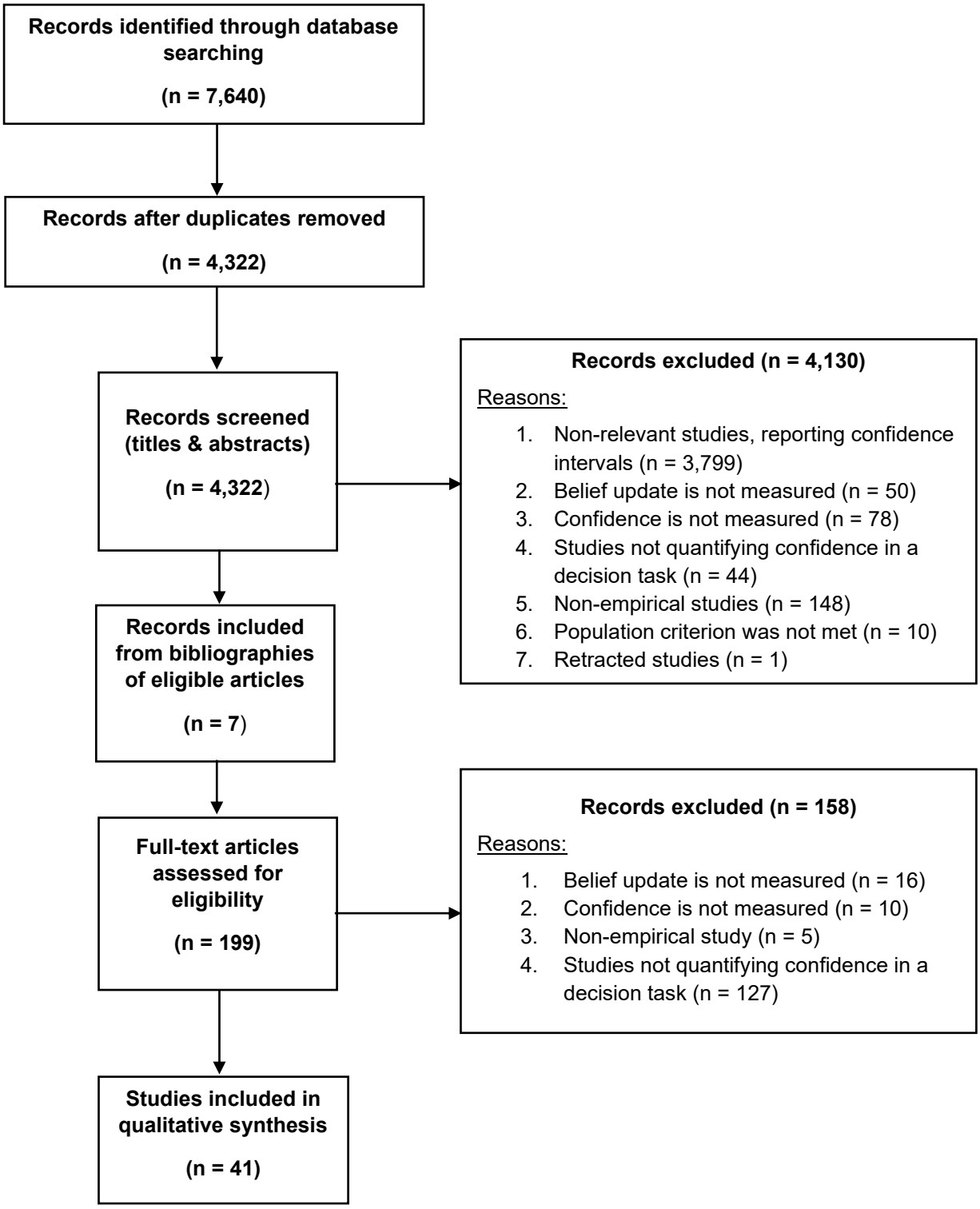


Figure 5.1. PRISMA flow diagram.

5.3.1. Methodological quality

The quality assessment of each eligible study was based on five criteria, adopted from the QATSDD tool (Sirriyeh *et al.*, 2012): description of settings, sample size considerations, representativeness of sample, description of procedure and appropriateness of methodology. The quality score of the 41 eligible articles ranged from 26.67% to 93.33% (mean 60.98%, median 60%). 11 studies were of quality below 50% and 30 above 50%. Most of the studies appeared not to consider the size of the sample in the study design, neither did they provide a power analysis for minimum sample size requirements. As a result, this criterion yielded the lowest quality score (i.e., 18.70%). The quality score about representativeness of the sample was also low (i.e., 31.71%), while the rest three criteria about research settings, procedure and methodology had an average score of 84.82%. The 27 studies that supported the association between confidence in an initial decision and the likelihood of changing this decision had a quality score of 59.01%. The five studies that showed partial support for this association had a quality score of 66.67%, while the nine studies that did not find any evidence for the association between initial confidence and decision change had a score of 63.70%. Medical and non-medical studies had an average quality score of 61.59% and 60.33% respectively (Tables 5.2. - 5.4.).

The difference in quality scores between studies that found support for the relationship between initial confidence and change of judgement and those that did not find one could be due to a number of different reasons. One reason could be that half of the supportive studies used either students or lay people as participants, whereas the studies that did not find support for the relationship between confidence and change of judgement mostly used (medical) professionals. Also, approximately 20% of the supportive studies had a considerably small sample size (n > 10 participants), while another 15% of the supportive studies did not report the sample size. Finally, more than 20% of the supportive studies did not report whether the results were significant.

5.3.2. Studies that found evidence for a relationship between initial confidence and change of judgement

From the 27 studies (65.85%) that found evidence for the relationship between initial confidence and change of judgement, 15 were medical studies, with study population either healthcare professionals or medical students. The rest of the studies mainly used syllogistic reasoning tasks, with either lay participants or students (Table 5.2.). The sample size of the studies ranged from one to 319 participants (mean 79, median 64 participants). In four studies (Lewis *et al.*, 2007; Boccardi *et al.*, 2016; Ceccaldi *et al.*, 2018; Spallazzi *et al.*, 2019), the number of participants was not reported. These were all studies in the medical domain, where physicians assessed a number of patient cases with cognitive impairments or abdominal pain (i.e., 205, 104, 228 and 126 patients respectively). Furthermore, one of the medical studies reported that decisions were made by consensus (Spallazzi *et al.*, 2019).

Study	Type of study	Population	Sample size	Quality score
Boccardi et al., 2016	Medical	Neurologists	Not reported	66.67%
Ceccaldi et al., 2018	Medical	Neurologists	Not reported	66.67%
Dreiseitl & Binder, 2005	Medical	Dermatologists	52	53.33%
Ferrarin et al., 2008	Medical	Physiatrists	2	53.33%
Jaimes et al., 2013	Medical	Dermatologists	22	60%
Krupat et al., 2017	Medical	Students	75	73.33%
Lewis et al., 2007	Medical	Emergency physicians	Not reported	53.33%
Mower et al., 2019	Medical	ED physicians	319	60%
Ossenkoppele et al., 2013	Medical	Neurologists	4	60%
Pandharipande et al., 2016	Medical	Trainee physicians	245	93.33%
Pfister et al., 1999	Medical	Nephrologists	18	60%

Spallazzi et al., 2019	Medical	Neurologists	Not reported	40%
Weston et al., 2016	Medical	Neurologists	1	46.67%
Yelamos et al., 2020	Medical	Dermatologists	7	46.67%
Zwan et al., 2017	Medical	Neurologist/Psychiatrist	2	40%
Bajsanski & Zauhar, 2019	Non-medical	Students	70	66.67%
Dujmovic et al., 2020	Non-medical	Students	53	60%
Folke et al., 2017	Non-medical	Lay people	24/28	53.33%
Kruglanski et al. 1993	Non-Medical	Students	60/74/41	66.67%
Phillips et al., 2014	Non-medical	Students	161	53.33%
Shynkaruk & Thompson, 2006	Non-medical	Students	128	73.33%
Thompson et al., 2011	Non-medical	Students	30/48/128/64	46.67%
Thompson, Turner et al., 2013	Non-medical	Students	108 and 72	73.33%
Thompson & Johnson, 2014	Non-medical	Students/Lay people	103	80%
Thompson, Evans et al., 2013	Non-medical	Students	104	73.33%
Tullis & Goldstone, 2020	Non-medical	Students	208	46.67%
Vega et al., 2020	Non-medical	Lay people	89	53.33%

Table 5.2. Supportive studies.

5.3.2.1. Medical studies

15 eligible studies conducted in healthcare were captured by the search. The studies were published between 1999 and 2020. Six studies were on patients with cognitive impairment, assessed by neurologists (Ossenkoppele *et al.*, 2013; Boccardi *et al.*, 2016; Weston *et al.*, 2016; Zwan *et al.*, 2017; Ceccaldi *et al.*, 2018; Spallazzi *et al.*, 2019). One study was on movement analysis and used physiatrists as participants

(Ferrarin *et al.*, 2008). Two studies were on dermoscopic diagnosis, and one on decision aids with dermatologists as participants (Dreiseitl and Binder, 2005; Jaimes *et al.*, 2013; Yélamos *et al.*, 2020). One study was on renal disease where nephrologists' diagnostic judgements were analysed (Pfister *et al.*, 1999). Another study on premature closure in clinical diagnosis used medical students as participants (Krupat *et al.*, 2017). Finally, three studies on the effects of computed tomography (CT) and utrasonography in emergency medicine used medical students, trainees, as well as fully qualified physicians as participants (Lewis *et al.*, 2007; Pandharipande, Reisner, *et al.*, 2016; Mower *et al.*, 2019).

The sample size of the medical studies that found support for the association between confidence and resistance to change ranged from one to 319 physicians (mean 75, median 53). In four studies, the number of physicians was not reported (Lewis *et al.*, 2007; Boccardi *et al.*, 2016; Ceccaldi *et al.*, 2018; Spallazzi *et al.*, 2019). In most of the studies, the physicians' task included the assessment of a real or hypothetical patient (e.g., a vignette) in two steps: before and after an intervention. In the first step, physicians provided a preliminary diagnosis and/or a treatment plan, while in the second step (i.e., after the intervention) they provided a final diagnosis and/or a treatment plan. The type of intervention (experimental manipulation) between the initial and final diagnoses ranged across studies, with most of the studies using Positron Emission Tomography (PET) results (Ossenkoppele *et al.*, 2013; Boccardi *et al.*, 2016; Weston *et al.*, 2016; Zwan *et al.*, 2017; Ceccaldi *et al.*, 2018; Spallazzi *et al.*, 2019). Other studies used CT scan results (Lewis *et al.*, 2007; Pandharipande, Reisner, *et al.*, 2016), Instrumented Gait Analysis (Ferrarin *et al.*, 2008), Reflectance Confocal Microscopy (Yélamos *et al.*, 2020), ultrasonography (Mower *et al.*, 2019), time constraints (Jaimes *et al.*, 2013), disconfirming information (Krupat *et al.*, 2017), management suggestions from a decision aid (Dreiseitl and Binder, 2005) or kidney biopsy results (Pfister *et al.*, 1999).

Confidence in an initial decision or judgement was measured on various scales: most of the studies used a 0-100% visual analogue scale (Pfister *et al.*, 1999; Ossenkoppele *et al.*, 2013; Boccardi *et al.*, 2016; Pandharipande, Reisner, *et al.*, 2016; Zwan *et al.*, 2017; Ceccaldi *et al.*, 2018), while other studies used a 1-100% VAS (Krupat *et al.*, 2017), a 0-10 Likert scale (Dreiseitl and Binder, 2005; Yélamos *et*

al., 2020), a 1-10 Likert scale (Ferrarin *et al.*, 2008; Jaimes *et al.*, 2013), a 3-point scale (Mower *et al.*, 2019) or a 50-85% VAS (Spallazzi *et al.*, 2019).

All studies found that confidence in an initial decision (either diagnosis or treatment) was negatively associated with decision change. Most of the studies reported significance with the relevant p values, without necessary reporting the size of the effect (Dreiseitl and Binder, 2005; Jaimes *et al.*, 2013; Ossenkoppele *et al.*, 2013; Boccardi *et al.*, 2016; Pandharipande, Reisner, *et al.*, 2016; Krupat *et al.*, 2017; Zwan *et al.*, 2017; Ceccaldi *et al.*, 2018; Spallazzi *et al.*, 2019), while other studies did not mention whether the results were significant (Pfister *et al.*, 1999; Lewis *et al.*, 2007; Ferrarin *et al.*, 2008; Weston *et al.*, 2016; Mower *et al.*, 2019; Yélamos *et al.*, 2020). These studies provided only frequencies but no tests of statistical significance. These results are therefore interpreted with caution.

In one study, where instrumented movement assessment was conducted on patients with focal dystonia, when physicians' confidence in their initial diagnosis was lower (i.e., prior to assessment), changes in diagnosis (i.e., after the assessment) were more frequent (Ferrarin *et al.*, 2008). However, the study does not report whether this difference was significant. In another study, where reflectance confocal microscopy (RCM) was used to assess skin lesions for diagnosis of skin cancer, dermatologists who were more confident in their initial diagnosis pre-RCM were less likely to change their diagnosis post-RCM (Yélamos *et al.*, 2020). However, the study does not report any numerical results or p values. In a third study where no statistical significance was reported, 18 nephrologists assessed patient cases with kidney diseases. For each patient, they were asked to provide their diagnosis twice, over a period of six months. The study found that the frequency of change in diagnosis was negatively associated with nephrologists' confidence in their initial diagnosis (Pfister *et al.*, 1999). However, the study did not provide effect sizes or p values.

Similar results were found in another three studies that did not report statistical significance. Specifically, in one study (Weston *et al.*, 2016), one neurologist assessed 20 patients with dementia syndromes before and after amyloid PET scan. The neurologist was asked to provide his/her initial diagnosis and the level of confidence (0-100% scale). After PET results, the neurologist provided his/her final

diagnosis and confidence level. The study reported that there was a change in diagnosis in 7 out of 20 patients (i.e., 35%) and that in those patients, the neurologist's initial confidence was lower, compared to patients whose initial diagnosis was not changed (55% versus 76%).

In another study (Mower *et al.*, 2019), 319 emergency medicine physicians assessed patient cases before and after ultrasonographic evaluations. They indicated how confident they were that an abscess was present on a 3-point scale, as well as their management decision. After the ultrasonographic evaluation, physicians were asked to update their management decision. The study reported that ultrasonography changed physicians' management in 3.1% of the cases, and that physicians who were confident about their initial management decision were 20 times less likely to change their initial judgement, compared to less confident physicians (1.20% versus 23.80%). However, there was a large difference between the number of confident and less confident management decisions (i.e., 1,111 versus 105 instances). Also, the study did not report whether the difference in frequency of changes in management was significant.

Finally, in another study in emergency medicine (Lewis *et al.*, 2007), physicians were asked to assess patient cases with abdominal pain before and after an optional CT scan. Initially, physicians provided their preliminary diagnosis, based on patient history and physical examination, as well as their level of confidence on a 0-100% scale. After the CT scan results, physicians provided their final diagnosis and confidence. The results showed that physicians who ordered a CT scan had lower confidence than physicians who did not order a CT scan ($z = 3.96$, $p < 0.001$) and that they changed their initial diagnosis more often (46% versus 29%). However, the study did not report whether the difference in change in diagnosis between physicians was significant. Furthermore, the number of physicians who took part in the study was not reported.

Amongst the medical studies that found support for the relationship between initial confidence and decision change, there was a study where the diagnostic decision was made by consensus (Spallazzi *et al.*, 2019). In this study, a group of neurologists assessed 104 patients with mild cognitive impairment or dementia. The

neurologists were asked to provide their diagnosis and their level of confidence on a scale from 50% to 85% twice: before and after PET results. The study showed that neurologists who changed their diagnosis after obtaining PET results had lower confidence in their initial diagnosis compared to those who did not change their diagnosis (63% versus 70.40%, $p < 0.01$). However, the diagnostic decisions and confidence ratings were not made individually but collectively. The study did not report how the neurologists solved potential disagreements about the final diagnosis of confidence ratings.

5.3.2.2. Non-medical studies

A further 12 studies conducted in non-medical settings found support for the relationship between initial confidence and decision change. The studies were published between 1993 and 2020. Most of the studies used syllogistic reasoning tasks (e.g., conditional statements, probability judgements, moral dilemmas) with students and/or lay people as participants (Shynkaruk and Thompson, 2006; Thompson, Prowse Turner and Pennycook, 2011; Thompson, Ackerman, *et al.*, 2013; Thompson, Evans and Campbell, 2013; Thompson and Johnson, 2014; Bajšanski and Žauhar, 2019; Dujmovic *et al.*, 2020; Vega *et al.*, 2020). Other studies were of lay participants choosing between food items of different values (Folke *et al.*, 2017), and students answering multiple choice questions (Phillips, Prybutok and Peak, 2014; Tullis and Goldstone, 2020) or assessing a legal scenario (Kruglanski, Webster and Klem, 1993). The number of participants in these 12 studies ranged from 24 to 208 (84 participants per study on average).

In order to measure change in judgement, the studies required participants to produce a dual response: an initial response before the intervention or manipulation and a final response. The type of manipulation that most studies used was constraining the time of the first response (Shynkaruk and Thompson, 2006; Thompson, Prowse Turner and Pennycook, 2011; Thompson, Ackerman, *et al.*, 2013; Thompson, Evans and Campbell, 2013; Thompson and Johnson, 2014; Bajšanski and Žauhar, 2019; Dujmovic *et al.*, 2020; Vega *et al.*, 2020): the typical task included the presentation of a reasoning problem or dilemma (e.g., the Linda

problem, see Kahneman, 2011, p.156 or the Wason selection task, see Wason and Shapiro, 1971, p. 64), where participants were initially asked to provide a quick, intuitive response (within a few seconds) and, subsequently, a final response without time constraints. Other studies used peer instruction, i.e., discussing one's own performance with a peer (Tullis and Goldstone, 2020); disconfirming information (Phillips, Prybutok and Peak, 2014); the jury method - similar to peer instruction but for consensual legal decisions - also described in the General Introduction (Kruglanski, Webster and Klem, 1993); and task repetition without any explicit manipulation, e.g., a set of choices that appear more than once during a selection task (Folke *et al.*, 2017).

To measure confidence in an initial diagnosis, six out of the 12 studies used a 7-point Likert scale, ranging from "guessing" to "certain I am right" (Shynkaruk and Thompson, 2006; Thompson, Prowse Turner and Pennycook, 2011; Thompson, Ackerman, *et al.*, 2013; Thompson, Evans and Campbell, 2013; Phillips, Prybutok and Peak, 2014; Thompson and Johnson, 2014). One study used a 4-point scale ranging from "completely uncertain" to "completely certain" (Vega *et al.*, 2020), two studies used a 50% to 100% visual analogue scale (Bajšanski and Žauhar, 2019; Dujmovic *et al.*, 2020), another study used a 10-point scale ranging from "not at all confident" to "highly confident" (Tullis and Goldstone, 2020), another study used a 9-point scale ranging from -4 ("extremely confident about A") to 4 ("extremely confident about B") (Kruglanski, Webster and Klem, 1993), and finally, there was one study where the confidence scale was not reported (Folke *et al.*, 2017).

The relative consistency of the non-medical studies in terms of confidence measurement scale (e.g., 1-7 scale), nature of the task (e.g., conditional statements) and type of intervention (e.g., time constraints), is explained by the fact that most of them were conducted, or inspired by, the work of Professor Valerie Thompson and her colleagues. In their studies, Thompson and colleagues have been typically using two-response reasoning tasks. The tasks could include judging the validity of conclusions based on conditional statements, e.g., if p then q (Shynkaruk and Thompson, 2006; Thompson, Prowse Turner and Pennycook, 2011); estimating the probability of an event, e.g., base rate neglect tasks (Thompson, Prowse Turner and Pennycook, 2011; Thompson, Ackerman, *et al.*, 2013; Thompson and Johnson,

2014) and denominator neglect tasks (Thompson and Johnson, 2014); and assessing the relationship between events, e.g., causal reasoning tasks (Thompson and Johnson, 2014). In these studies, Thompson and colleagues asked participants to provide two responses to any given problem: a quick, initial response that comes to mind within seconds, and a more deliberative, final response without time constraints. For both initial and final responses, participants provided ratings of confidence on a scale from 1 ("guessing") to 7 ("certain I am right"). Thompson and colleagues termed the confidence in participants' initial judgements as "feeling of rightness" (Thompson, Prowse Turner and Pennycook, 2011).

All studies showed that there is a negative association between initial confidence in a judgement and the likelihood that this judgement will be revised. Statistical significance with the related p values and effect sizes were reported in all of the studies. In one study on conjunction fallacy (e.g., the Linda problem, where people falsely assume that the co-occurrence of two specific events or conditions is more probable than the occurrence of each event separately, see Kahneman, 2011, p. 156), participants were presented with a description of a person (i.e., John is a lawyer at a prestigious firm who spends time on golf courses). They were then presented with two statements about this person: a single and a conjunctive (e.g., [a] John owns an expensive car and [b] John owns an expensive car and is a workaholic) and asked to judge which statement is more probable. There were four versions of statements, varied in representativeness and order in which they were presented. The results showed that in three out of four versions of the problem, participants who changed their judgements were significantly less confident compared to those who did not change, after being allowed to think without time constraints ($t = 3.00$, $t = 3.12$, $t = 3.18$, $p < 0.01$). The study reported that, despite the fact that there was evidence for the relationship between initial confidence and change of judgement, in two versions of the problem (where there was a significant difference in initial confidence), change of judgement was similar, i.e., 13.96% versus 12.45% (Dujmovic *et al.*, 2020).

In another study on confirmation bias (Phillips, Prybutok and Peak, 2014), participants were presented with a hypothetical scenario about the unauthorised use of laptops in a classroom, and asked to indicate their decision for or against it, as

well as their confidence in their decision. Subsequently, participants were presented with information that contradicted their initial decision and were asked whether they wish to obtain more information and then make a final decision. The results showed that participants with high initial confidence were less willing to seek further information before reaching a conclusion and, subsequently, less likely to change their initial decision (Phillips, Prybutok and Peak, 2014). Despite the fact that the study reports that there was an association between initial confidence and change of judgement (possibly via the intention to seek further information), the authors only report the significant effect of information seeking intention on change of judgement ($p < 0.05$).

Finally, there is an early study by Shynkaruk and Thompson, (2006) on deductive reasoning, where they found evidence for the relationship between initial confidence and change of judgement. The study included two experiments, similar to the ones described above. However, in the second experiment the relevant results (about change of judgement) were not reported.

5.3.3. Studies that did not find evidence for a relationship between initial confidence and change of judgement

Nine studies (21.95%) did not find evidence to support the relationship between initial confidence and change of judgement. Five of those studies were conducted in the medical domain with healthcare professionals and trainees as participants. The remaining four studies explored various topics (e.g., forecasting, conditional reasoning), using students, lay people or professionals as participants (Table 5.3.). The sample size of these studies ranged from two to 256 participants (mean 116, median 91). One medical study did not report the number of physicians who took part (Ferrarin *et al.*, 2015).

Study	Type of study	Population	Sample size	Quality score
Bergeron et al., 2015	Medical	Neurologists	2	60%
Eva et al., 2010	Medical	Primary care physicians	256	93.33%
Ferrarin et al., 2015	Medical	Physicians	Not reported	53.33%
Medow et al., 2010	Medical	Intern. Medicine physicians	108	73.33%
Wood et al., 2017	Medical	Physicians	23	73.33%
Ahlawat, 1999	Non-medical	Managers/Seniors	91	60%
Rowe & Right, 1996	Non-medical	Students	60	66.67%
Rowe et al., 2005	Non-medical	Students	90	66.67%
Trouche et al., 2014	Non-medical	Lay people	215/204	26.67%

Table 5.3. Non-supportive studies.

5.3.3.1. Medical studies

The five medical studies were published between 2010 and 2017. One study on decision aid advice used internal medicine residents as participants (Medow, Arkes and Shaffer, 2010). Another study on dementia and the clinical impact of a repeated PET used neurologists as participants (Bergeron *et al.*, 2016). In another study on diagnostic flexibility, participants were primary care physicians (Eva *et al.*, 2010). One study on Objective Structured Clinical Examination (OSCE) used physicians with various specialties, from internal medicine to surgery and anaesthesia (Wood *et al.*, 2017). Finally, in one study on gait analysis and its effects on clinical decision making, physicians' specialty, as well as the number of physicians who took part, were not reported (Ferrarin *et al.*, 2015). The sample size of the medical studies ranged from two to 256 physicians (mean 97, median 108 participants).

In most of the studies, the physicians' task was to assess a real or hypothetical patient scenario in two steps, before and after an intervention (e.g., CT scan or

PET). Specifically, in one study, internal medicine residents were presented with a community-acquired pneumonia (CAP) case and asked to decide about admission to intensive care (Medow, Arkes and Shaffer, 2010). They also provided confidence ratings (50-100% scale) about their decision. They were then presented with advice that contradicted their initial decision, either from a computerised decision aid, or an anonymous pulmonologist, and then provided their final decision. The results showed that physicians who received advice from the decision aid were more likely to change their management decision and reduce their confidence in it. However, regardless of the source of advice, physicians' confidence in their initial decision was not associated with their changing of judgement ($p = 0.50$).

In another study (Bergeron *et al.*, 2016), two neurologists were presented with atypical dementia patient cases and asked to provide an initial diagnosis and their level of confidence in it (0-100% scale). The neurologists were then presented with PET results and were asked to provide their final diagnosis and confidence. The results showed that lower confidence in initial diagnosis led to more frequent diagnostic changes (29% versus 15%). However, the study reported that this difference was not significant (the p value was not reported).

In a third study (Ferrarin *et al.*, 2015), physicians assessed patients who had suffered a cerebrovascular accident before and after gait analysis and provided their treatment recommendations and confidence (1-10 scale). Treatment recommendations included surgery, physiotherapy, orthotic management, and neuromuscular blockade. The study found that gait analysis led to more frequent changes of management recommendations. However, physicians' initial confidence was not related to changes in recommendations (the p value was not reported).

In another study in primary care (Eva *et al.*, 2010), 256 physicians were presented with a patient case with symptoms that suggested coronary heart disease (CHD). Half of the physicians were primed to consider CHD (e.g., "a colleague physician said that a CHD is possible"). Physicians were asked to provide a preliminary diagnosis based on a brief patient description, and a final diagnosis, after further patient information was revealed. Physicians also indicated their confidence in their diagnosis on a scale from 0% to 100%. The results showed that 36.80% of

physicians changed their initial diagnosis (Eva *et al.*, 2010). However, confidence in initial diagnosis was not associated with diagnostic change (p = 0.95).

Finally, there was a study on OSCE (Wood *et al.*, 2017), where 23 physicians from different specialties were presented with a clinical examination video and asked to rate the examinee's (i.e., physician's) performance on a scale from 1 ("inferior") to 6 ("excellent"). Performance scoring criteria included history taking, diagnosis, organisation, communication skills and data interpretation. Examiner physicians provided two ratings for the examinee physicians' performance: an initial rating after 60 seconds of the video, and a final rating once the video was completed. Examiners also indicated their confidence in their performance ratings on a 5-point scale. Despite the fact that there was variability between confidence ratings in examiners' initial decision (i.e., examinees' performance ratings), initial confidence was not related to the likelihood of changing their judgement (p value was not reported).

5.3.3.2. Non-medical studies

Four non-medical studies, published between 1996 and 2014, found no evidence for the relationship between initial confidence and changing one's judgement. Two studies by Rowe and colleagues (1996, 2005) recruited 60 and 90 undergraduate students respectively (Rowe and Wright, 1996; Rowe, Wright and McColl, 2005). One study on deductive reasoning used 204 and 215 lay participants in two experiments (Trouche, Sander and Mercier, 2014), and the last study on individual versus group decision making used 91 professionals from a Certified Public Accountant (CPA) firm (Ahlawat, 1999). The average sample size for these four non-medical studies was 132 participants.

In the first two studies by Rowe and colleagues, participants' task was to provide their forecast for a number of (mainly economic and political) events. In the 1996 study, participants were given multiple choice questions to answer and indicated their confidence on a 7-point scale (Rowe and Wright, 1996), whereas in the 2005 study participants were given binary choice questions and indicated their confidence on a 50-100% scale (Rowe, Wright and McColl, 2005). After different types of

feedback (e.g., statistical feedback: what other participants have answered, or explanatory feedback: why other participants have answered this way), participants were asked to complete the task again. Both studies failed to find any association between participants' confidence in their initial answers and the likelihood of changing these answers when the task was repeated after feedback ($p > 0.05$).

In the study on deductive reasoning (Trouche, Sander and Mercier, 2014), in two similar experiments, Trouche and colleagues (2014) presented participants with a reasoning task similar to the Cognitive Reflection Task (Frederick, 2005) – e.g., "Paul is looking at Linda and Linda is looking at Patrick. Paul is married but Patrick is not. Is a person who is married looking at a person who is not married? Yes/No/Cannot be determined". In brief, they were asked to provide their answers and rate their confidence on a 9-point scale. They were also asked to justify their answers. Subsequently, participants were given an argument against their initial answer and asked to provide a final answer. The results showed that, in both experiments, participants with lower confidence in their initial answers were not more likely to change their answers compared to participants with higher confidence ($p = 0.42$ and $p = 0.40$ respectively).

Finally, in the study on individual versus group decision making (Ahlawat, 1999), participants were presented with a hypothetical auditing scenario. Their task was to evaluate the probability of a company's continued existence (0-100% scale), and to produce an audit report with three possible outcomes (i.e., unqualified, modified or disclaimer). Participants indicated their decision (i.e., the outcome of the audit) and their confidence in their report on a scale from 0% to 100%. Subsequently, participants were provided with six pieces of information: half of them were in favour of the company's existence and the other half provided information against the company's future existence. After each piece of information, participants were able to re-estimate the likelihood of the company's continued existence. At the end of the task, participants made a final decision about the audit report and indicated their confidence in their report. The judgement task was taken either individually (i.e., by one participant alone), or as a group (i.e., by three participants who made decisions by consensus). The results showed that confidence in the initial audit report was not associated with change in the report's outcome ($F = 3.26$, $p = 0.078$). The study did

not report whether this was the case in both group and individual decisions. However, it reported that there was a statistically significant difference in initial confidence between groups and individuals (i.e., 75.57% versus 63.57% respectively, $t = 2.11$, $p < 0.05$).

5.3.4. Studies with mixed results

From the eligible articles captured by the search, five studies (12.20%) found only partial support for the relationship between initial confidence and change of judgement. Two were conducted in the medical domain with fully qualified physicians as participants. The remaining three studies explored various topics (e.g., choosing between risky alternatives, forecasting football game results, assessing the validity of conclusions), with lay people and students as participants (Table 5.4.). The number of participants ranged from three to 192 (mean 91, median 99).

Study	Type of study	Population	Sample size	Quality score
Krudop et al., 2016	Medical	Neurologists	3	66.67%
Pandharipande et al., 2016	Medical	Primary care physicians	91	86.67%
Bago & De Neys, 2017	Non-medical	Lay people	101/99/115	60%
Bell & Jamieson, 1970	Non-medical	Students	192	60%
Bolger et al., 2011	Non-medical	Students/Lay people	39	60%

Table 5.4. Studies providing partial support.

5.3.4.1. Medical studies

The two eligible medical studies were both published in 2016. The average number of participants was 47 physicians. In one study (N = 91) on the effects of CT scan on clinical decision making, participants were primary care physicians (Pandharipande, Alabre, *et al.*, 2016). In the other study (N = 3) on dementia syndromes, participants

were neurologists (Krudop *et al.*, 2016). As in the medical studies described in previous sections, physicians' main task was to assess a number of either hypothetical or real patient cases in two steps, before and after an intervention. The type of intervention was either a CT scan (Pandharipande, Alabre, *et al.*, 2016) or Cerebrospinal fluid analysis (Krudop *et al.*, 2016).

In the first study (Pandharipande, Alabre, *et al.*, 2016), 91 primary care physicians were presented with a number of clinical cases (from 1 to 34 cases per physician) based on real patients, who presented with either weight loss, haematuria or abdominal pain. Initially, they were asked to provide their leading diagnosis and indicate their confidence in it on a scale from 0% to 100%. Physicians were subsequently presented with CT scan results and asked to provide a final diagnosis and their level of confidence. The frequency of changes in diagnosis after the CT scan was relatively similar across patient cases (i.e., weight loss: 57%, haematuria: 49%, abdominal pain: 53%). There was a negative association between physicians' confidence in their initial diagnosis and the likelihood of changing this diagnosis after CT scan results. However, this was significant only in the cases of weight loss ($p = 0.03$) and abdominal pain ($p < 0.0001$), and not in cases with haematuria ($p = 0.14$).

In the second study on dementia syndromes (Krudop *et al.*, 2016), three neurologists were presented with 137 real patient cases and were asked to provide their initial diagnosis and confidence level on a 0-100% scale. They were subsequently presented with neuroimaging and Cerebrospinal Fluid Biomarker results and asked to update their diagnosis and confidence. Diagnoses were classified into three groups: behavioural variant frontotemporal dementia (bvFTD), psychiatry, and other neurological disorders. The results showed that when there was a diagnostic change from bvFTD towards another diagnosis, neurologists' confidence in their initial diagnosis was lower (OR = 1.57, [1.04-2.38], p value was not reported). The same was true for diagnostic changes from other diagnoses towards bvFTD (OR = 1.08, [1.01–1.16], p value was not reported). However, the study did not include results for diagnostic changes from psychiatry to other neurological disorders and vice versa (i.e., approximately 33% of total number of diagnostic changes was not included).

5.3.4.2. Non-medical studies

The three non-medical studies that provided only partial support for the relationship between initial confidence and changing of judgement were published in 1970 (N = 192), 2011 (N = 39) and 2017 (N = 101, 99 and 115 in three experiments). The studies are briefly described above.

The first study (Bell and Jamieson, 1970) investigated groups' tendency to make riskier choices than individuals (i.e., the risky shift phenomenon) and the role of publicity in this tendency. In brief, 192 students were presented with 12 binary-choice dilemmas (e.g., "A person with severe heart problems must either change strong life habits or undertake a potentially fatal surgery. How would you advise this person?") and asked to provide their initial choice and confidence on a 0-100% scale. After a number of different manipulations (e.g., discussion of choices made with a group of people, feedback about other participants' responses), the students were able to reconsider their responses and make a final choice. The results showed that less confident participants changed their initial choices more frequently. However, this difference was not significant ($t = 0.96$). An additional comparison of changes in initial choices, for participants with more than one standard deviation above or below the mean initial confidence, was made. The results showed that lower confidence in the initial choice led to more changes ($t = 2.30$, $p < 0.05$). However, both analyses took into account only the risky shifts (i.e., changes from less risky to risky choices). Information about no-risky shifts was not provided.

In the second study (Bolger *et al.*, 2011), 39 participants (students and lay people) were asked to predict the winning team of various football games. They initially provided their forecasts and their confidence on a 50-100% scale. They were then provided with either numeric feedback (other participants' forecasts and confidence ratings) or reason feedback (justifications about the forecasts) and provided a final forecast. In the numeric feedback condition, participants with lower confidence in their forecasts were more likely to change their judgement compared to their more confident counterparts ($p = 0.01$). However, this was not true in the reason feedback condition ($p = 0.36$).

The final study (Bago and De Neys, 2017) consisted of a series of experiments on syllogistic reasoning, similar to Thompson and colleagues' studies described above (Thompson, Prowse Turner and Pennycook, 2011; Thompson, Ackerman, *et al.*, 2013; Thompson and Johnson, 2014). In three experiments, 101, 99 and 115 participants, recruited via an online platform, were presented with reasoning tasks, and asked to provide first a quick response and then a final response without time constraints. Participants also provided confidence ratings in their initial response on a 0-100% scale. The study found that, in two out of three experiments (i.e., experiments 1 and 4), there was a significant, negative association between participants' confidence in their initial response and the likelihood of changing their judgement ($p = 0.01$ and $p < 0.0001$ respectively). In the third experiment though, this effect was not significant ($p = 0.16$).

5.4. Discussion

5.4.1. Summary

Overall, this systematic review presented a comprehensive body of evidence for the association between confidence in an initial decision and the likelihood of changing one's judgement. The majority of the studies identified by this review (i.e., 78%) found that confidence in an initial decision is, at least partially, negatively associated with the likelihood of changing this decision. Approximately half of the studies were on medical decision making. The overall methodological quality of the studies that supported the link between initial confidence and changing of judgement was 60.21% (~moderate). This is mostly due to their sample size and representativeness, as discussed above.

In the body of literature that was assessed through this systematic review, the degree of belief in an initial judgement (i.e., confidence) has been found to be a good predictor of the subsequent revision of this judgement. As the evidence of the present review suggests, judgements made with higher confidence are, indeed, less likely to change. This resistance to change has been found not only in individual but in group decision making and ranged from simple consumer choices and syllogistic

reasoning problems to more complex medical diagnoses and moral dilemmas. This phenomenon is therefore not limited to lay participants or students, but it is also applicable to experienced professionals.

As mentioned above, different approaches from the field of judgement and decision making tried to explain this phenomenon by attributing it to various cognitive biases such as: illusion of validity or overconfidence (Kahneman and Tversky, 1974; Kahneman, 2011; Thompson, Prowse Turner and Pennycook, 2011; Thompson, Prowse Turner, *et al.*, 2013), anchoring and adjustment (Kahneman and Tversky, 1974; Hogarth and Einhorn, 1992), information distortion (Simon *et al.*, 2001; Kostopoulou *et al.*, 2012), confirmation bias (Kahneman, 2011; Mendel *et al.*, 2011; Rollwage *et al.*, 2020) or premature closure (Webster and Kruglanski, 1994; Kruglanski and Webster, 1996; Krupat *et al.*, 2017). With regards to closure, Kruglanski and Webster (1996) highlighted that people often form judgements on the basis of inconclusive evidence, as they seek to reach a conclusion (i.e., closure). This *need* for (cognitive) closure is characterised by two tendencies, what Kruglanski and Webster call “seizing” and “freezing” (Kruglanski and Webster, 1996). Seizing refers to peoples’ tendency to reach a (relatively confident) conclusion as soon as possible (i.e., urgency), whereas freezing refers to the tendency to commit to this conclusion for as long as possible (i.e., permanence). By extension, people who seek for closure may reach a conclusion prematurely, due to urgency, and be reluctant in changing their judgement, as found in Kruglanski and colleagues (1993) – also see Chapter 1: General introduction.

Taken together, one could argue that the above annotations can create a coherent narrative about how resistance to change occurs. That is, intuitive thinking is fast in forming an initial judgement. This is because intuitive thinking seeks cognitive ease. The initial judgement or impression is usually accompanied by a feeling of rightness. People usually anchor themselves in this initial impression and, any adjustment away from it is often insufficient. In addition, being effort-averse, intuitive thinking is reluctant to evaluate information that contradicts its initial judgement, as this would require additional processing. Instead, it seeks only information that confirms its judgement, as this would allow for reaching a conclusion quickly. If it is not possible to avoid contradicting information, it can always alter the meaning of the information

in a way that fits with the initial judgement. At this point the feeling of rightness would be even higher, firstly because there is “more” information that “confirms” the initial judgement and secondly, because the urgency for a conclusion did not allow for alternative hypotheses. Hence, it would feel better to stick with the first impression rather than to change it.

In summary, revising a judgement or a belief is cognitively and timely expensive. As the results of this review suggest, initial confidence can determine whether, and to what extent, a judgement would be revised or not. For well-calibrated decision makers (i.e., those who are appropriately confident in the accuracy of their judgements) it would be safe to maintain their initial judgements. However, for those who are overconfident about the accuracy of their judgements, resistance to change judgement could lead to erroneous decisions. This is potentially problematic even for underconfident decision makers, as they could be more open to other opinions and potentially change their judgement toward a less accurate decision. In any case, assuming that initial confidence correlates with resistance to change, one could argue that by lowering confidence in a belief or judgement, we can indirectly induce change. This would be particularly important for judgements made with unwarranted confidence on the basis of insufficient information.

5.4.2. Limitations

One of the main limitations of this review is that it did not capture all the available evidence for the relationship between initial confidence and changing one’s judgement. This is due to the way the review was designed. Specifically, the inclusion criteria allowed for studies that measure judgements and confidence in them in decision making tasks. Studies that measure confidence in attitudes or personal opinions about various topics were excluded (Mello *et al.*, 2020). Studies that measure confidence in memory or recognition tasks or general knowledge questions were also excluded (Saito, 1998; Lau and Coiera, 2008; Morgan *et al.*, 2012; Relling and Giuliodori, 2015; Couchman *et al.*, 2016; Cross *et al.*, 2017; Griffiths and Higham, 2018; Rollwage *et al.*, 2020). Finally, studies that measure willingness to change judgement and not actual change of judgement were not

included either (Burnstein and Katz, 1971; Jermias, 2006; Couchman *et al.*, 2016; Wang and Thompson, 2019). As a result, a considerable number of studies (i.e., at least a further 12 articles identified by the search) that found that higher initial confidence is negatively associated with change of judgement were not included. Likewise, two studies that did not confirm the association between initial confidence and change of judgement were not included either, as they measured confidence in recognition or general knowledge tasks (Saito, 2003; Simon and Kim, 2017).

Another limitation is the representativeness of the sample: in approximately half of the eligible studies participants were undergraduate students. This is also reflected on the relevant quality assessment score of the studies, which was found to be low (i.e., 30%). Furthermore, in five studies (four supportive of and one contradicting the positive relationship between confidence and resistance to change) the number of participants was not reported (Lewis *et al.*, 2007; Ferrarin *et al.*, 2015; Boccardi *et al.*, 2016; Ceccaldi *et al.*, 2018; Spallazzi *et al.*, 2019). In two studies, one supportive (Spallazzi *et al.*, 2019) and one non-supportive (that found a negative relationship between initial confidence and resistance to change though, Ahlawat, 1999), judgements and confidence in them were measured in group decision making, without reporting how consensus was reached. There are also several studies that used a very small sample size. These are mainly medical studies where the target population was the patients and not the physicians. For the purpose of this review though, the population of interest was actually the physicians (i.e., the decision makers) and not the patients.

Furthermore, there is a study on deductive reasoning by Trouche and colleagues (2014) that found no support for the relationship between initial confidence and resistance to change judgement. The study was briefly described above (Trouche, Sander and Mercier, 2014). However, while in the rest of the studies participants provided their judgements in two steps (e.g., Step 1: an initial judgement before the intervention and Step 2: a final judgement after the intervention), in this study participants had an additional, intermediate step to reflect on their initial judgement and confidence (Step 3). In the rest of the studies, initial confidence was confidence in Step 1, and revision of answer was computed as the difference in answers between Step 1 and Step 3. In this study though, initial confidence was confidence in

Step 2 and revision of answer was computed by the difference in answers between Step 2 and Step 3. In other words, this procedure allowed participants to revise their initial answer and confidence twice, while counting the second measurements as if they were the initial ones. This is potentially problematic in the context of this review, as it suggests that participants' intuitive thinking could have already been disrupted intermediately by the justification of their initial answer, before an intervention was introduced. If that were the case, then initial confidence in that study did not reflect the degree of belief in an initial, possibly intuitive judgement, but the degree of belief in an already revised, and possibly more deliberate judgement.

5.4.3. Implications for future research

This work could be expanded in various ways. Firstly, the review assessed the evidence for the relationship between confidence in an initial judgement and resistance to changing one's decision or judgement. It did not account for the accuracy of the initial judgement, or the direction of the change. In particular, the review did not test whether initial judgements made with higher confidence were correct or not, and how this related to resistance to change. Likewise, the review did not test whether changes of initial judgements led to more accurate decisions. Also, the present review did not check whether certain interventions/manipulations led to more frequent changes and what the role of initial confidence in this was.

Furthermore, several articles were excluded due to the way the review was designed. Specifically, and as mentioned above, certain studies that found a relationship between initial confidence and resistance to change were not included, as they did not use a decision-making task but measured confidence in other types of tasks. Expanding the review to include this kind of studies would possibly allow for more definite conclusions about the effect of confidence in resistance to change. In addition, this review was limited to five bibliographic databases, two of them related to the medical and pharmacological literature. As a result, half of the eligible studies came from the medical domain. A future systematic review could include different databases that would capture studies that possibly have not been identified by this search (e.g., legal and financial decision making). Finally, it is important to note that

the results of the present review could provide behavioural insights and inform future policies and interventions that focus on triggering deliberate reasoning, as an attempt to change people's judgements toward optimal, and more accurate decisions.

Finally, there are some studies, both supportive and non-supportive, that did not test the significance of the relationship between initial confidence and change of judgement. These studies provided only descriptive information, such us means and frequencies, that indicate trends. However, since there is no statistical significance reported, any interpretation based on such findings should be made with caution. Future reviews could control for such studies and use stricter criteria for their inclusion in order to find more conclusive evidence.

Summary

- Studies 1, 2, and 3 found that confident initial judgements are less likely to change
- Systematic evidence from 32 (out of 41 eligible) studies confirmed the finding above: initial decisions and judgements made with high confidence are more resistant to change
- This finding applies to domains outside medicine (e.g., in moral dilemmas and syllogistic reasoning tasks)
- This finding is not limited to lay people but applies to professionals.

CHAPTER 6: General discussion

6.1. Summary

This research consisted of three empirical studies on the effects of early diagnostic suggestions on physicians' reasoning, and a systematic review on the role of confidence in changing one's mind. The empirical studies investigated different aspects of the diagnostic process including confidence in an initial diagnosis, information search, information evaluation and diagnostic changes, when physicians encountering new information. Using online clinical scenarios, we tested whether providing diagnostic suggestions early on can reduce the effects of cognitive biases that potentially operate during the diagnostic process, namely overconfidence, premature closure, information distortion and resistance to change.

Study 1 tested if the provision of diagnostic suggestions can reduce GPs' confidence in their initial diagnosis and whether this can lead to a change in diagnosis when equivocal information is later encountered. GPs assessed two patient scenarios, one with and the other without suggestions, allocated at random. Results were in the expected direction: GPs reduced their initial confidence and changed their diagnosis more frequently when provided with diagnostic suggestions. However, these differences were not significant. Only in one scenario there was a significant effect of suggestions on GPs' confidence. Across conditions and scenarios, GPs with higher initial confidence ordered fewer investigations, changed their diagnosis less frequently, and had lower scores in the Actively Open-minded Thinking scale (AOT). Female GPs were more likely to change their diagnosis than males but were not less confident, neither did they have higher AOT scores.

Study 2 tested two additional aspects of the diagnostic process: information search and information evaluation. Specifically, we tested whether providing diagnostic suggestions can encourage GPs to seek for more clinical information before settling on a final diagnosis. We also investigated whether diagnostic suggestions could make GPs evaluate neutral patient information in a more balanced way and change their diagnosis more frequently. GPs assessed two patient scenarios in a random

order, one with and the other without suggestions. Contrary to our predictions, diagnostic suggestions had no effect on GPs' initial confidence, information search, information evaluation or diagnostic change. In fact, some of the trends were counterintuitive: when GPs received diagnostic suggestions, neutral patient information tended to be rated as more supportive of the initial diagnosis rather than less; they changed diagnosis less frequently and gave fewer differential diagnoses at the end of the scenarios; they did however seek more information before providing a final diagnosis, as would be expected.

Rather than diagnostic suggestions, it was initial confidence that appeared to be the main driver of behaviour, and it significantly affected the variables of interest: GPs with higher initial confidence sought less patient information, evaluated this information in a more biased way and changed their diagnosis less frequently when receiving equivocal information. The only effect of diagnostic suggestions was on GPs' final confidence: when GPs were presented with the suggestions, they were less confident about their final diagnosis, regardless of whether they had changed diagnosis or not. Severity of GPs' initial diagnosis had a significant effect on the variables of interest. Specifically, GPs who initially provided a common, less serious diagnosis (MSK and IBS, depending on the scenario) requested more information and rated neutral information as more supportive of their initial diagnosis. They also increased their confidence in their initial diagnosis more after receiving additional patient information and were less likely to change it. GPs with lower AOT scores sought less patient information and generated fewer differential diagnoses. This time, female GPs were found to be less confident in their initial diagnosis compared to their male counterparts and sought more patient information before they provided their final diagnosis. No differences in AOT scores were detected between genders.

Given the counterintuitive trends seen in Study 2, we conducted a further study, which was in essence an additional experimental condition but with different participants, given that Study 2 was completed. The list of diagnostic suggestions did not include the diagnosis that most GPs were expected to give based on the initial patient description. This was done to test the hypothesis that GPs in Study 2 were simply confirming rather than questioning their initial diagnosis. GPs in Study 3 assessed one patient scenario only. GPs who did not provide the expected diagnosis

were excluded from the analyses. GP responses in Study 3 were compared with responses in Study 2, separately for the control and aided conditions. We found that there were significantly more diagnostic changes in Study 3 than the aided condition of Study 2 but there was no difference from Study 2 control. All other expected differences were not significant, but they were in line with our hypothesis. Specifically, providing suggestions that did not include the GPs' initial diagnosis resulted in more information being requested compared to both the control and aided conditions of Study 2. Confidence about the initial diagnosis increased but less so in Study 3 than the aided condition of Study 2; no differences with the control condition were detected. Furthermore, in Study 3, the neutral patient information was perceived as less supportive of the initial diagnosis than in the aided condition of Study 2; there was no difference from Study 2 controls.

In line with both previous studies, initial confidence was the main driver of behaviour also in Study 3. GPs with higher confidence in their initial diagnosis sought less patient information before concluding, evaluated this information as more supportive of their initial diagnosis and changed diagnosis less frequently. In line with Study 2, severity of the GPs' initial diagnosis was significantly associated with perceived cue support, change in initial confidence, and change of diagnosis: GPs who provided a less serious initial diagnoses perceived the neutral cues as more supportive of their initial diagnosis, had a greater increase in their initial confidence after receiving additional patient information and changed diagnosis less frequently. The provision of diagnostic suggestions in both Studies 2 and 3 tended to depress final confidence compared to controls but not significantly so. In line with Study 2, GPs with higher AOT engaged in more extensive search of information to reach a conclusion and recorded a greater number of differential diagnoses in Study 3. Interestingly so, higher AOT scores were positively associated with confidence in final diagnosis in Study 3 only but, in line with Study 1, they were negatively associated with confidence in initial diagnosis. In line with Study 2, male GPs were more confident in their initial diagnosis compared to their female counterparts. As discussed above, all these associations were significant.

The negative relationship between initial confidence and diagnostic change, a consistent finding in all three studies, was further explored through a systematic

literature review. The review was not limited to medical judgements and decisions. It collated and assessed the evidence for a relationship between confidence in an initial judgement and likelihood of changing one's judgement (or resistance to changing one's mind). The review was conducted across five electronic databases and identified 41 eligible articles. The large majority of the articles captured by the search, provided evidence for a negative association between initial confidence and changing one's mind, confirming both the pattern found in our studies, as well as the received wisdom that beliefs held with greater confidence are more resistant to change.

6.2. General conclusions

Three studies explored the potential psychological mechanisms underlying the impact of a diagnostic decision aid in the form of diagnostic suggestions, provided early in the diagnostic process. Each study was an attempt to increase the likelihood to detect any effect of the diagnostic suggestions, addressing the limitations of the previous study/ies. Overall, the effects of diagnostic suggestions were found to be subtle but in the expected direction. The consistent finding of the negative relationship between initial confidence and resistance to changing one's judgement was further explored through a systematic review, which produced evidence of this effect beyond medical decision making.

Does the provision of diagnostic suggestions reduce physicians' confidence?

Based on the three studies conducted in this research work, diagnostic suggestions appear to have only a weak effect on physicians' confidence in their initial diagnostic hypothesis. In Study 1, they reduced confidence only in one of the two patient scenarios. Study 2 did not detect this effect; instead, diagnostic suggestions were associated with reduced confidence in the final diagnosis. Study 3 showed that diagnostic suggestions that do not include the physician's initial diagnostic hypothesis led to a smaller increase in initial confidence and lower final confidence, but these effects were not statistically significant. Thus, the findings regarding

diagnostic confidence are only partially in line with studies which found that providing suggestions and considering alternative hypotheses can lower decision confidence (e.g., Koriat, Lichtenstein and Fischhoff, 1980; Hirt and Markman, 1995; McKenzie, 1997; Haran, Moore and Morewedge, 2010).

There may be several reasons for the weak effect of diagnostic suggestions on physicians' confidence. Firstly, the patient scenarios used in the studies intended to induce relatively high confidence in an initial diagnosis. However, in all three studies, initial confidence was rather moderate. Secondly, in Studies 1 and 2, the list of diagnostic suggestions included the expected diagnosis, and this could have been perceived as a reassuring indication that their initial diagnosis was probably correct. Thirdly, as participants' feedback suggested, assigning explicit confidence ratings in diagnostic decisions might have felt unnatural. It is therefore possible that the numeric scale did not sufficiently captured physicians' diagnostic confidence. Finally, it is also possible that the list of suggestions provided by the aid was discounted, given that it was relatively lengthy, and included diagnoses that were perhaps considered too unlikely. Furthermore, no information about the aid was provided to physicians to engender trust and motivate them to take it into account (e.g., Yaniv, 2004; Wang and Du, 2018).

Does the provision of diagnostic suggestions reduce premature closure?

Studies 2 and 3 that explored physicians' information search found that the provision of diagnostic suggestions led to more extensive search of information before reaching a final diagnosis. This was irrespective of whether the suggestions included physicians' own diagnosis or not. However, this effect was not significant, therefore, it cannot be claimed that diagnostic suggestions can reduce premature closure. This appears to be somewhat in contrast to Kruglanski and Webster's (1996) view that higher need for closure leads to fewer alternative hypotheses generated ("lack of alternatives", see Chapter 1: General Introduction). This is because, as Studies 2 and 3 showed, the provision of diagnostic suggestions (i.e., alternative hypotheses) did not reduce physicians' premature closure, taking into account that not all

alternatives would be actively considered, due to working memory constraints (Thomas and colleagues, 2008, 2014, also see Chapter 1: General Introduction).

It is possible that our ability to detect an effect of diagnostic suggestions on information search was undermined by the fact that physicians could select from a limited number of certain items of information. If the patient scenarios contained more information available to the physicians, they could reveal a stronger effect. Real clinical consultations allow physicians to actively search and collect as many pieces of information as they need in order to reach a conclusion about the patient.

Does the provision of diagnostic suggestions reduce information distortion?

Studies 2 and 3 measured the evaluation of neutral, open to interpretation, information under the influence of a leading diagnostic hypothesis. We had expected that the provision of diagnostic suggestions would make GPs more balanced in their evaluation. In Study 2, physicians who were presented with diagnostic suggestions did not perceive the information as less supportive of their initial diagnosis. However, as Study 3 showed, when diagnostic suggestions did not include physicians' own diagnosis, the perceived support was smaller than when diagnostic suggestions included the physicians' own diagnosis. This suggests that "contradictory advice" (as opposed to advice that contains both confirming and contradicting information) can make physicians evaluate neutral information in a more balanced way. Nevertheless, since the differences were not significant, we cannot argue that the provision of diagnostic suggestions reduced information distortion.

A potential explanation of the failure to detect an effect on information evaluation could be the variability in diagnostic value between the different pieces of patient information. This information was designed to be neutral, with minimal diagnostic value that would not help physicians differentiate between possible diagnoses. However, as Studies 2 and 3 showed, some pieces of information were evaluated as more supportive of the physicians' initial diagnosis, while other pieces were evaluated as less supportive or neutral.

Does the provision of diagnostic suggestions reduce resistance to changing one's mind?

Diagnostic confidence and diagnostic changes were measured in all three studies. Study 1 showed that when diagnostic suggestions were provided, physicians changed their initial diagnosis more frequently after encountering neutral information. However, this was not statistically significant in any of the patient scenarios. In Study 2, in contrast to our hypothesis, the provision of diagnostic suggestions did not lead to more frequent changes of the initial diagnosis. Study 3 found that diagnostic suggestions which excluded the physicians' initial diagnosis led to more frequent diagnostic changes compared to suggestions that included that diagnosis. This difference was significant, though the difference from the control condition (no suggestions) was not. Overall and in line with findings from previous studies (e.g., Mussweiler, Strack and Pfeif, 2000; Dreiseitl and Binder, 2005; Huang, Hsu and Ku, 2012), the three studies provide some evidence that diagnostic suggestions can reduce resistance to changing diagnosis, when subsequent information is not confirmatory of an initial hypothesis. Nevertheless, this evidence is weak and should be interpreted with caution.

What else is worth noting?

As previously discussed, the most consistent finding was related to physicians' confidence in their initial diagnosis. Study 1 found that physicians with higher initial confidence were less likely to change their mind. Studies 2 and 3 replicated this finding, expanding the effect to physicians' information search and information evaluation. Higher initial confidence was associated with less extensive search for information and more biased evaluation of information. It was also associated with fewer differential diagnoses. Overall, the pattern of the positive association between initial confidence and resistance to change was further investigated by reviewing studies outside medicine. The systematic review conducted to assess the existing evidence on this association revealed that, indeed, initial judgements made with high confidence are less likely to change.

Limitations

The limitations of this research work were discussed throughout this thesis. However, it is important to highlight again the scepticism about the external validity of the findings. In all the studies, we used hypothetical patient scenarios that were assessed online. Even if we take into account the sharp increase of online clinical consultations due to the COVID-19 pandemic (Car *et al.*, 2020; Webster, 2020), the design of our studies differs from a virtual clinical consultation on several counts, notably, the pre-fixed patient history, the limited pool of available information, the requirement of only one diagnostic hypothesis and the response scales aiming to make explicit (usually implicit) cognitive processes. Our requirement for a diagnostic "conclusion" based on inconclusive evidence was possibly impetuous. Taken together, one could argue that the diagnostic process was narrowed down in a non-realistic way. Such sacrifices of realism though were inevitable in order to standardise the process and make sense of the data.

Multiple choice diagnoses (as opposed to free-text responses) would be another sacrifice of realism for the sake of a neater analysis. By providing physicians with a set of initial diagnoses to choose from (not necessarily mutually exclusive), we would be able to measure their confidence in differential diagnoses, as well as their perceived support of subsequent patient information relative to more than one diagnoses. This would potentially undermine the disruptive purpose of the early diagnostic suggestions, given that it would plant the consideration of alternative hypotheses pre-decisionally. However, it would possibly allow us to capture the different diagnostic hypotheses that physicians may consider at the start of the consultation.

Another limitation derives from our own hypotheses. Since we hypothesised that the decision aid could disrupt physicians' intuition by reducing their diagnostic confidence, the effect of diagnostic suggestions would be ideally demonstrated in a situation where physicians are relatively biased. That is, overconfident in their own diagnosis. We were not able to determine whether physicians were well-calibrated, given that we did not measure their diagnostic accuracy. However, their relatively

moderate confidence in their initial diagnosis, which was found to be the main driver throughout the three studies, did not possibly help us to detect a stronger effect.

Furthermore, the decision aid could have been introduced to physicians in different ways. For instance, it could have been more interactive, similar to the prototype developed by Kostopoulou and colleagues (2015, 2017). Physicians would have been able to enter patient symptoms and risk factors and receive suggestions that would vary depending on the information entered. Also, the suggestions could have been presented in a likelihood/prevalence order, rather than in an alphabetical one, so they would have been more informative. The suggestions could have also been limited to only the most likely ones, reducing their length, and possibly the distance from physicians' own views to half (e.g., six suggestions). Finally, there could have been more information about the decision aid, and possibly a few practice sessions, so physicians would have been able to build trust in the aid and become more open to its suggestions. Introducing the aid as suggestions provided by other colleagues could also be an option, to reduce any reluctance to computerised, as opposed to human, advice (also see McParland, Cooper and Johnston, 2019).

Finally, throughout this research work we tested several hypotheses and ran a large number of analyses. We were looking for subtle differences and powered the empirical studies to detect small effects, given the exploratory nature of our research. We did not control for Type 1 error rates, and we therefore did not penalise the significant level (p values) for hypotheses testing. For this reason, any significant results found could be due to Type 1 error.

Implications for future research

Several implications have been discussed throughout this thesis. This could include different patient scenarios, richer patient information, more active information search, more neutral patient information to be evaluated, different framing of the decision aid such as advice from colleagues, more information about the aid, different outputs such as risk calculations or counterarguments, different scales to measure confidence and perceived support. There is also space for qualitative methodology

including think aloud protocols, in-depth interviews with physicians, as well as eye-tracking methodologies. Eye-tracking would be particularly insightful, given that it would allow us to explore how physicians interact with such decision aids, without substantially interrupting the diagnostic process.

The ultimate aim of this research work was to use the findings of the empirical studies to inform how the decision aid discussed throughout this thesis could be optimised for adoption in clinical practice. Adoption and implementation of such interventions that improve patient safety may be beyond the scope of this research work. However, it is worth noting that studies on the potential of decision aids to improve medical decision making, along with the different models of diffusion of innovations (e.g., Rogers, Singhal and Quinlan, 1983 and Greenhalgh *et al.*, 2004) can inform policy makers for the successful implementation of such innovations within healthcare systems.

CHAPTER 7: Appendices

7.1. Appendix 1

Participant information sheet (Study 1)

Study title: Influences of a diagnostic decision aid on GPs' assessment of patients

This research study aims to investigate the influence of a diagnostic aid on General Practitioners' assessment of patients, and to suggest improvements in the way diagnostic support for primary care is designed. The researchers are Ploutarchos Kourtidis, Olga Kostopoulou, and Brendan Delaney. They are all based at Imperial College London. The research is funded by the NIHR Imperial PSTRC (Patient Safety Translational Research Centre).

We are inviting currently practising GPs and GP trainees to participate. Before you decide to take part, it is important to understand what participation involves.

What to expect

Participation is **anonymous** and takes about 10 minutes.

1. First, you will be asked to indicate your gender and whether you are a GP or GP trainee.
2. Then, you will see two hypothetical patient cases.
    - In one case, you will receive suggestions from a diagnostic aid, in the other you will not.
    - Information in each case will be presented sequentially in two steps.
    - At each step, you will be asked to give a working diagnosis, and indicate your certainty on a 0-10 scale.
    - You will also be asked what investigations you would order and what other information you would like.
3. At the end, you will be asked to complete a short, standardised questionnaire about thinking styles.
4. You will then be re-directed to another site where you can claim a £10 Amazon voucher. You will need to enter your name and e-mail address to which the voucher will be sent. Since this is done on a different site, **it is not possible to link your personal data to your research data.**

Please note: Case information is not sufficient for a definite diagnosis.

Try to complete the study in one sitting.

If you decide not to take part, you can simply close the browser and any responses you have entered will be deleted.

What will happen to the data and results of the study

The results of this study will be submitted for publication in academic journals and presented at academic conferences. Upon completion of the study, the research data will be deposited in a public repository (Open Science Framework: https://osf.io), where they will be freely available to other researchers.

This study has been approved by the Head of Department of Surgery and Cancer and the Joint Research Compliance Office.

If you have any questions about the study, please contact **Ploutarchos Kourtidis (p.kourtidis17@imperial.ac.uk) or Olga Kostopoulou (o.kostopoulou@imperial.ac.uk)**

GDPR-related information

Imperial College London is the sponsor for this study based in the United Kingdom. We will be using information from you in order to undertake this study and will act as the data controller for this study. This means that we are responsible for looking after your information and using it properly. Imperial College London will keep identifiable information about you.

As a university we use personally-identifiable information to conduct research to improve health, care and services. As a publicly-funded organisation, we have to ensure that it is in the public interest when we use personally-identifiable information from people who have agreed to take part in research.

For auditing purposes, Imperial College London will collect information about you (name and e-mail address) from the researchers. This information is regarded as a special category of information. The University's Finance Office will use this information only for auditing purposes.

You can download a copy of this information sheet here.

On the next page, you will be asked to give your informed consent.

7.2. Appendix 2

Consent Form (Study 1)

Study title: Influences of a diagnostic decision aid on GP's assessment of patients
Name of Researchers: Olga Kostopoulou and Ploutarchos Kourtidis

1. I confirm that I have read and understood the information for the above study.
2. I understand that my participation is voluntary, and that I can withdraw at any point without giving a reason, by closing the browser. In this case, my data will be deleted.
3. I understand that the data are anonymous; there is no way that individuals can be identified from the data.
4. I agree that upon completion of the study, the data will be deposited in a public repository (Open Science Framework: https://osf.io), where they will be freely available to other researchers.
5. I understand that it is not possible to withdraw my responses after completing the survey.

You can download a copy of this consent form here.

7.3. Appendix 3

Invitation e-mail (Study 1)

Study title: Influences of a diagnostic decision aid on GPs' assessment of patients

Dear Doctor

We are contacting you because you either took part in one of our previous studies or have indicated your interest in participating in future studies on Medical Decision Making.

We are inviting currently practising GPs and GP trainees in England to participate in an online research study into the effects of a diagnostic decision aid on GPs' assessment of patients.

Participation is anonymous, takes approximately 10 minutes, and is entirely online. Upon clicking the link below:

1. First, you will be asked to read a study information sheet and provide consent online.
2. Then, you will be asked to indicate your gender and whether you are a GP or GP trainee.
3. Next, you will read two hypothetical patient descriptions and be asked to provide a diagnosis and your certainty. In one case, you will receive the advice of a diagnostic decision aid, in the other you will not.
4. You will then be given a short questionnaire about thinking styles.
5. At the end, you will be able to claim a £10 Amazon voucher.

We are interested in whether the aid has any impact on the assessment of patients, and if this depends on the GPs' usual thinking style.

To claim the £10 Amazon voucher, you will be re-directed to a website where you can enter your name and e-mail address. In this way, it is not possible to link your personal details to the research data.

Please forward this invitation to any other GP colleagues who might be interested.

Thank you in advance for your help!

Dr Olga Kostopoulou (o.kostopoulou@imperial.ac.uk)
Professor Brendan Delaney (brendan.delaney@imperial.ac.uk)

The Medical Decision Making Research Group
Imperial College London

To participate, please click on the link below. **Try to complete the study in one sitting. The link will remain active for one week from the moment that you will access it.**

If you have any problems or questions, please e-mail Ploutarchos Kourtidis (p.kourtidis17@imperial.ac.uk)

Take me to the survey

7.4. Appendix 4

Participant information sheet (Studies 2 and 3 with minor changes)

Title: Influences of a diagnostic decision aid on clinical reasoning

We are inviting currently practising GPs and GP trainees to participate. Before you decide to take part, it is important to understand what participation involves.

What is the purpose of the study?

This research study aims to investigate the influence of a diagnostic aid on General Practitioners' assessment of patients, and to suggest improvements in the way diagnostic support for primary care is designed. The researchers are Ploutarchos Kourtidis, Olga Kostopoulou, and Brendan Delaney. They are all based at Imperial College London. The study is being undertaken as part of a PhD study.

Why have I been chosen?

Through the National Institute of Health Research (NIHR) Clinical Research Network (CRN) North West London, we are inviting currently practising General Practitioners and GP trainees in England to participate in an online study on the effects a diagnostic decision aid on clinical reasoning. We aim to recruit 250 GPs.

Do I have to take part?

It is up to you to decide whether or not to take part. If you do decide to take part you will be able to download this information sheet and provide your implied consent. If you decide to take part you are still free to withdraw at any time and without giving a reason by closing the browser. In this case, your data will be deleted.

What to expect

Participation is **anonymous** and takes about 15 minutes.

1. First, you will be asked to indicate your gender and whether you are a GP or GP trainee.
2. Then, you will see two hypothetical patient scenarios and will receive suggestions from a diagnostic aid in one of them.

- Scenario information will be presented sequentially in two steps. At the first step, you will see a short patient description and be asked to give your initial diagnosis and level of confidence. At the second step, you will have the opportunity to request specific items of information about the patient.
- For each item you request, you will be asked to indicate how much this item supports your initial diagnosis.
- Finally, you will be asked to update your diagnosis and confidence.
- You will also be asked how you would manage the patient.
- **Scenario information is not sufficient for a definite diagnosis.**

3. At the end, you will be asked to complete a short-standardised questionnaire about your thinking style.
4. You will then be re-directed to another site where you can claim a £10 Amazon voucher. You will need to enter your name and e-mail address to which the voucher will be sent. Since this is done on a different site, **it is not possible to link your personal details (name and e-mail address) to your study responses.**

We ask you to complete the study in one sitting.

What are the possible benefits of taking part?

Participating GPs will be sent the study results, if they wish, and find out if and in what way our decision aid impacts clinical reasoning.

What if something goes wrong?

If you are harmed by taking part in this research project, there are no special compensation arrangements. If you are harmed due to someone's negligence, then you may have grounds for a legal action. Regardless of this, if you wish to complain, or have any concerns about any aspect of the way you have been treated during the course of this study then you should immediately inform the Investigator (Olga Kostopoulou, o.kostopoulou@imperial.ac.uk). The normal National Health Service complaints mechanisms are also available to you. If you are still not satisfied with the response, you may contact the Imperial AHSC Joint Research Compliance Office.

Will my taking part in this study be kept confidential?

All information collected about you will be kept confidential.

Imperial College London is the sponsor for this study based in the United Kingdom. We will be using information from you in order to undertake this study and will act as the data controller for this study. This means that we are responsible for looking after your information and using it properly. Imperial College London will keep identifiable information about you:

- 10 years after the study has finished in relation to data subject consent forms.
- 10 years after the study has completed in relation to primary research data.

Personal data i.e., name and e-mail address, as well as your study responses will be collected via Qualtrics (www.qualtrics.com) in two separate websites. This will ensure that your personal data cannot be linked to your study responses. Qualtrics is a General Data Protection Regulation (GDPR) compliant, third-party online platform. Further information about Qualtrics privacy policy can be found at https://www.qualtrics.com/privacy-statement/.

Further information on Imperial College London's retention periods may be found at https://www.imperial.ac.uk/media/imperial-college/administration-and-support-services/records-and-archives/public/RetentionSchedule.pdf.

Your rights to access, change or move your information are limited, as we need to manage your information in specific ways in order for the research to be reliable and accurate. If you withdraw from the study, we will keep the information about you that we have already obtained. To safeguard your rights, we will use the minimum personally identifiable information possible.

You can find out more about how we use your information by contacting the investigator (o.kostopoulou@imperial.ac.uk).

LEGAL BASIS

As a university, we use personally-identifiable information to conduct research to improve health, care and services. As a publicly-funded organisation, we have to ensure that it is in the public interest when we use personally-identifiable information from people who have agreed to take part in research. This means that when you agree to take part in a research study, we will use your data in the ways needed to conduct and analyse the research study.

Health and care research should serve the public interest, which means that research sponsors have to demonstrate that their research serves the interests of society as a whole. They do this by following the UK Policy Framework for Health and Social Care Research.

INTERNATIONAL TRANSFERS

There may be a requirement to transfer information to countries outside the European Economic Area (for example, to a research partner). Where this information contains your personal data, Imperial College London will ensure that it is transferred in accordance with data protection legislation. If the data is transferred to a country, which is not subject to a European Commission (**EC**) adequacy decision in respect of its data protection standards, Imperial College London will enter into a data sharing agreement with the recipient organisation that incorporates EC approved standard contractual clauses that safeguard how your personal data is processed.

CONTACT US

If you wish to raise a complaint on how we have handled your personal data or if you want to find out more about how we use your information, please contact Imperial College London's Data Protection Officer via email at dpo@imperial.ac.uk, via telephone on 020 7594 3502 and via post at Imperial College London, Data Protection Officer, Faculty Building Level 4, London SW7 2AZ.

If you are not satisfied with our response or believe we are processing your personal data in a way that is not lawful you can complain to the Information Commissioner's

Office (ICO). The ICO does recommend that you seek to resolve matters with the data controller (us) first before involving the regulator.

Imperial College London will collect information from you for this research study in accordance with our instructions.

Imperial College London will use your name and contact details, i.e., e-mail address to contact you about the research study and make sure that relevant information about the study is recorded for your care, and to oversee the quality of the study. Individuals from Imperial College London and regulatory organisations may look at your research records to check the accuracy of the research study. The only people at Imperial College London who will have access to information that identifies you will be people who need to contact you to disseminate the results of the study or to audit the data collection process. The people who analyse the information will not be able to identify you and will not be able to find out your name or contact details.

Imperial College London will keep identifiable information about you from this study for 10 years after the study has finished and archived in accordance to Imperial's archiving policy.

When you agree to take part in a research study, the information about your health and care may be provided to researchers running other research studies in this organisation and in other organisations. These organisations may be universities, NHS organisations or companies involved in health and care research in this country or abroad. Your information will only be used by organisations and researchers to conduct research in accordance with the UK Policy Framework for Health and Social Care Research.

Your information could be used for research in any aspect of health or care and could be combined with information about you from other sources held by researchers, the NHS or government.

Where this information could identify you, the information will be held securely with strict arrangements about who can access the information. The information will only be used for the purpose of health and care research, or to contact you about future

opportunities to participate in research. It will not be used to make decisions about future services available to you, such as insurance.
Where there is a risk that you can be identified your data will only be used in research that has been independently reviewed by an ethics committee.

Any information collected about you may be used to support other research in the future. In this case, this information will be coded and not identifiable.

What will happen to the data and results of the study?

The results of this study will be submitted for publication in academic journals and presented at academic conferences. Upon completion of the study, the research data will be deposited in a public repository (Open Science Framework: https://osf.io), where they will be freely available to other researchers. Only anonymous data will be deposited in the repository.

You may opt in to receive a copy of the results of this study or be contacted about participation in future studies. In this case, you will be asked to provide your consent for us to retain your personal data, according to the General Data Protection Regulation (GDPR).

Who is organising and funding the research?

Imperial College London is the study sponsor. The research is funded by the NIHR Imperial PSTRC (Patient Safety Translational Research Centre).

Who has reviewed the study?

This study was approved by the Health Research Authority (HRA).

Contact for Further Information

If you have any questions about the study, please contact **Ploutarchos Kourtidis (p.kourtidis17@imperial.ac.uk) or Olga Kostopoulou (o.kostopoulou@imperial.ac.uk).**

You can download a copy of this information sheet here.
On the next page, you will be asked to give your implied consent.
Thank you for taking part in the study!

Version 9, 16/12/2019

7.5. Appendix 5

Consent Form (Studies 2 and 3)

Study title: Influences of a diagnostic decision aid on clinical reasoning

Name of researchers: Olga Kostopoulou and Ploutarchos Kourtidis

1. I confirm that I have read and understood the subject information sheet dated 16/12/2019 version 9 for the above study and have had the opportunity to ask questions which have been fully answered.

2. I understand that my participation is voluntary, and that I can withdraw at any point without giving a reason, by closing the browser. In this case, my data will be deleted.

3. I understand that the data is anonymous; there is no way that individuals can be identified from the data.

4. I agree that upon completion of the study, the anonymous data will be deposited in a public repository (Open Science Framework: https://osf.io), where they will be freely available to other researchers.

5. I understand that it is not possible to withdraw my responses after fully completing the survey.

6. I give consent for information collected about me to be used to support other research in the future. This information will be coded and not identifiable.

7. I agree to participate in this study.

You can download a copy of this consent form here.

To proceed to the survey, please click on the arrow below

Optional

Please indicate by ticking the applicable boxes whether you wish to be contacted about the findings of this study and possible participation in future studies.

- ☐ **I agree to be contacted by the research team to potentially taking part in other research studies.**
- ☐ **I agree to be contacted by the research team about the results of this study and give permission for my name and email address to be retained until the findings are available.**
- ☐ **I give consent for identifiable information collected about me to be exported outside of the EEA.**

7.6. Appendix 6

Invitation e-mail (Studies 2 and 3 with minor changes)

Study title: Influences of a diagnostic decision aid on clinical reasoning

Dear Doctor

We are inviting currently practising GPs and GP trainees in England to participate in an online survey study into the effects of a diagnostic decision aid on GPs' assessment of patients.

Participation is **anonymous**, takes approximately 15 minutes, and is entirely online.

Upon clicking the link below:

6. First, you will be asked to read a study information sheet and provide implied consent online.
7. Then, you will be asked to indicate your gender and whether you are a GP or GP trainee.
8. Next, you will read two hypothetical patient scenarios and be asked to provide a diagnosis and level of certainty. In one patient scenario, you will receive diagnostic suggestions from a decision aid, in the other you will not.
9. You will then be given a short questionnaire about your preferred thinking style.
10. At the end, you will be able to claim a £10 Amazon voucher.

To claim the £10 Amazon voucher, you will be re-directed to a website where you can enter your name and e-mail address. In this way, it is not possible to link your personal details to the research data.

Please forward this invitation to any other GP colleagues who might be interested.

To participate, please click on the link below.

Try to complete the study in one sitting.

The link will remain active for one week from the moment that you will access it.

If you have any problems or questions, please e-mail Ploutarchos Kourtidis (p.kourtidis17@imperial.ac.uk)

Thank you in advance for your help!

Dr Olga Kostopoulou (o.kostopoulou@imperial.ac.uk)
Professor Brendan Delaney (brendan.delaney@imperial.ac.uk)

The Medical Decision Making Research Group
Imperial College London

Take me to the survey

7.7. Appendix 7

Study flow in Studies 2 and 3 (as seen by the participating GPs)

Title: Influences of a diagnostic decision aid on clinical reasoning

Participant information sheet

We are inviting currently practising GPs and GP trainees to participate. Before you decide to take part, it is important to understand what participation involves.

What is the purpose of the study?
This research study aims to investigate the influence of a diagnostic aid on General Practitioners' assessment of patients, and to suggest improvements in the way diagnostic support for primary care is designed. The researchers are Ploutarchos Kourtidis, Olga Kostopoulou, and Brendan Delaney. They are all based at Imperial College London. The study is being undertaken as part of a PhD study.

Why have I been chosen?
Through the National Institute of Health Research (NIHR) Clinical Research Network (CRN) North West London, we are inviting currently practising General Practitioners and GP trainees in England to participate in an online study on the effects a diagnostic decision aid on clinical reasoning. We aim to recruit 250 GPs.

Do I have to take part?
It is up to you to decide whether or not to take part. If you do decide to take part you will be able to download this information sheet and provide your implied consent. If you decide to take part you are still free to withdraw at any time and without giving a reason by closing the browser. In this case, your data will be deleted.

What to expect
Participation is **anonymous** and takes about 15 minutes.

1. First, you will be asked to indicate your gender and whether you are a GP or GP trainee.
2. Then, you will see two hypothetical patient scenarios and will receive suggestions from a diagnostic aid in one of them.
    - Scenario information will be presented sequentially in two steps. At the first step, you will see a short patient description and be asked to give your initial diagnosis and level of confidence. At the second step, you will have the opportunity to request specific items of information about the patient.
    - For each item you request, you will be asked to indicate how much this item supports your initial diagnosis.
    - Finally, you will be asked to update your diagnosis and confidence.
    - You will also be asked how you would manage the patient.
    - **Scenario information is not sufficient for a definite diagnosis.**
3. At the end, you will be asked to complete a short-standardised questionnaire about your thinking style.
4. You will then be re-directed to another site where you can claim a £10 Amazon voucher. You will need to enter your name and e-mail address to which the voucher will be sent. Since this is done on a different site, **it is not possible to link your personal details (name and e-mail address) to your study responses.**

We ask you to complete the study in one sitting.

What are the possible benefits of taking part?
Participating GPs will be sent the study results, if they wish, and find out if and in what way our decision aid impacts clinical reasoning.

What if something goes wrong?

If you are harmed by taking part in this research project, there are no special compensation arrangements. If you are harmed due to someone's negligence, then you may have grounds for a legal action. Regardless of this, if you wish to complain, or have any concerns about any aspect of the way you have been treated during the course of this study then you should immediately inform the Investigator (Olga Kostopoulou, o.kostopoulou@imperial.ac.uk). The normal National Health Service complaints mechanisms are also available to you. If you are still not satisfied with the response, you may contact the Imperial AHSC Joint Research Compliance Office.

Will my taking part in this study be kept confidential?

All information collected about you will be kept confidential.

Imperial College London is the sponsor for this study based in the United Kingdom. We will be using information from you in order to undertake this study and will act as the data controller for this study. This means that we are responsible for looking after your information and using it properly. Imperial College London will keep identifiable information about you:

- 10 years after the study has finished in relation to data subject consent forms.
- 10 years after the study has completed in relation to primary research data.

Personal data i.e. name and e-mail address, as well as your study responses will be collected via Qualtrics (www.qualtrics.com) in two separate websites. This will ensure that your personal data cannot be linked to your study responses. Qualtrics is a General Data Protection Regulation (GDPR) compliant, third-party online platform. Further information about Qualtrics privacy policy can be found at https://www.qualtrics.com/privacy-statement/.

Will my taking part in this study be kept confidential?

All information collected about you will be kept confidential.

Imperial College London is the sponsor for this study based in the United Kingdom. We will be using information from you in order to undertake this study and will act as the data controller for this study. This means that we are responsible for looking after your information and using it properly. Imperial College London will keep identifiable information about you:

- 10 years after the study has finished in relation to data subject consent forms.
- 10 years after the study has completed in relation to primary research data.

Personal data i.e. name and e-mail address, as well as your study responses will be collected via Qualtrics (www.qualtrics.com) in two separate websites. This will ensure that your personal data cannot be linked to your study responses. Qualtrics is a General Data Protection Regulation (GDPR) compliant, third-party online platform. Further information about Qualtrics privacy policy can be found at https://www.qualtrics.com/privacy-statement/.

Further information on Imperial College London's retention periods may be found at https://www.imperial.ac.uk/media/imperial-college/administration-and-support-services/records-and-archives/public/RetentionSchedule.pdf.

Your rights to access change or move your information are limited, as we need to manage your information in specific ways in order for the research to be reliable and accurate. If you withdraw from the study, we will keep the information about you that we have already obtained. To safeguard your rights, we will use the minimum personally identifiable information possible. You can find out more about how we use your information by contacting the investigator (o.kostopoulou@imperial.ac.uk).

LEGAL BASIS

As a university, we use personally-identifiable information to conduct research to improve health, care and services. As a publicly-funded organisation, we have to ensure that it is in the public interest when we use personally-identifiable information from people who have agreed to take part in research. This means that when you agree to take part in a research study, we will use your data in the ways needed to conduct and analyse the research study.

Health and care research should serve the public interest, which means that research sponsors have to demonstrate that their research serves the interests of society as a whole. They do this by following the UK Policy Framework for Health and Social Care Research.

INTERNATIONAL TRANSFERS

There may be a requirement to transfer information to countries outside the European Economic Area (for example, to a research partner). Where this information contains your personal data, Imperial College London will ensure that it is transferred in accordance with data protection legislation. If the data is transferred to a country, which is not subject to a European Commission (**EC**) adequacy decision in respect of its data protection standards, Imperial College London will enter into a data sharing agreement with the recipient organisation that incorporates EC approved standard contractual clauses that safeguard how your personal data is processed.

CONTACT US

If you wish to raise a complaint on how we have handled your personal data or if you want to find out more about how we use your information, please contact Imperial College London's Data Protection Officer via email at dpo@imperial.ac.uk, via telephone on 020 7594 3502 and via post at Imperial College London, Data Protection Officer, Faculty Building Level 4, London SW7 2AZ.

If you are not satisfied with our response or believe we are processing your personal data in a way that is not lawful you can complain to the Information Commissioner's Office (ICO). The ICO does recommend that you seek to resolve matters with the data controller (us) first before involving the regulator.

Imperial College London will collect information from you for this research study in accordance with our instructions.

Imperial College London will use your name and contact details, i.e. e-mail address to contact you about the research study and make sure that relevant information about the study is recorded for your care, and to oversee the quality of the study. Individuals from Imperial College London and regulatory organisations may look at your research records to check the accuracy of the research study. The only people at Imperial College London who will have access to information that identifies you will be people who need to contact you to disseminate the results of the study or to audit the data collection process. The people who analyse the information will not be able to identify you and will not be able to find out your name, or contact details.

Imperial College London will keep identifiable information about you from this study for 10 years after the study has finished and archived in accordance to Imperial's archiving policy.

When you agree to take part in a research study, the information about your health and care may be provided to researchers running other research studies in this organisation and in other organisations. These organisations may be universities, NHS organisations or companies involved in health and care research in this country or abroad. Your information will only be used by organisations and researchers to conduct research in accordance with the UK Policy Framework for Health and Social Care Research.

Your information could be used for research in any aspect of health or care and could be combined with information about you from other sources held by researchers, the NHS or government.

Where this information could identify you, the information will be held securely with strict arrangements about who can access the information. The information will only be used for the purpose of health and care research, or to contact you about future opportunities to participate in research. It will not be used to make decisions about future services available to you, such as insurance.

Where there is a risk that you can be identified your data will only be used in research that has been independently reviewed by an ethics committee.

Any information collected about you may be used to support other research in the future. In this case, this information will be coded and not identifiable.

What will happen to the data and results of the study?

The results of this study will be submitted for publication in academic journals and presented at academic conferences. Upon completion of the study, the research data will be deposited in a public repository (Open Science Framework: https://osf.io), where they will be freely available to other researchers. Only anonymous data will be deposited in the repository.

You may opt in to receive a copy of the results of this study or be contacted about participation in future studies. In this case, you will be asked to provide your consent for us to retain your personal data, according to the General Data Protection Regulation (GDPR).

Who is organising and funding the research?

Imperial College London is the study sponsor. The research is funded by the NIHR Imperial PSTRC (Patient Safety Translational Research Centre).

Who has reviewed the study?

This study was approved by the Health Research Authority (HRA).

Contact for Further Information

If you have any questions about the study, please contact **Ploutarchos Kourtidis (p.kourtidis17@imperial.ac.uk) or Olga Kostopoulou (o.kostopoulou@imperial.ac.uk)**

You can download a copy of this information sheet here.

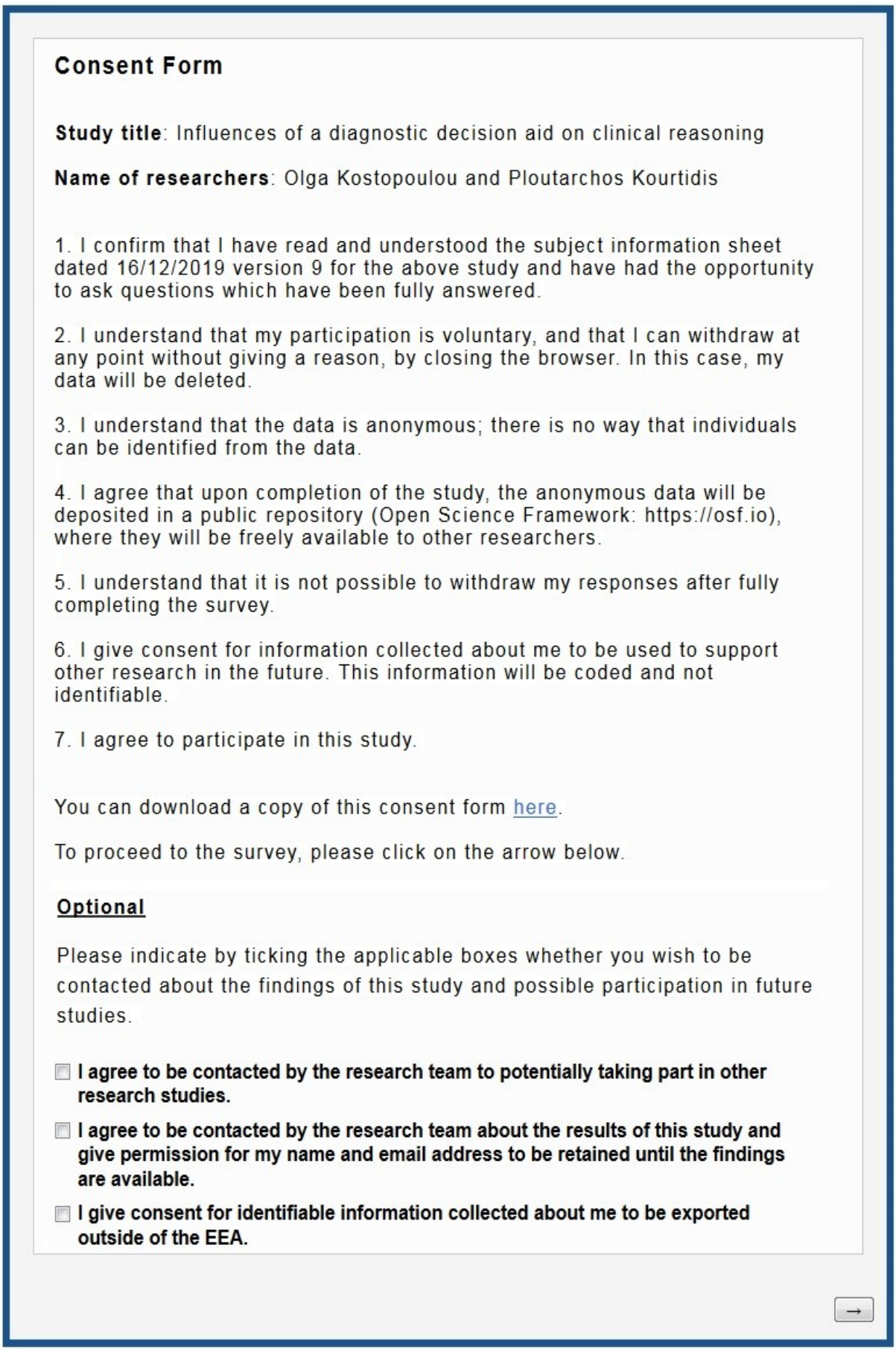
On the next page, you will be asked to give your implied consent.

Thank you for taking part in the study!

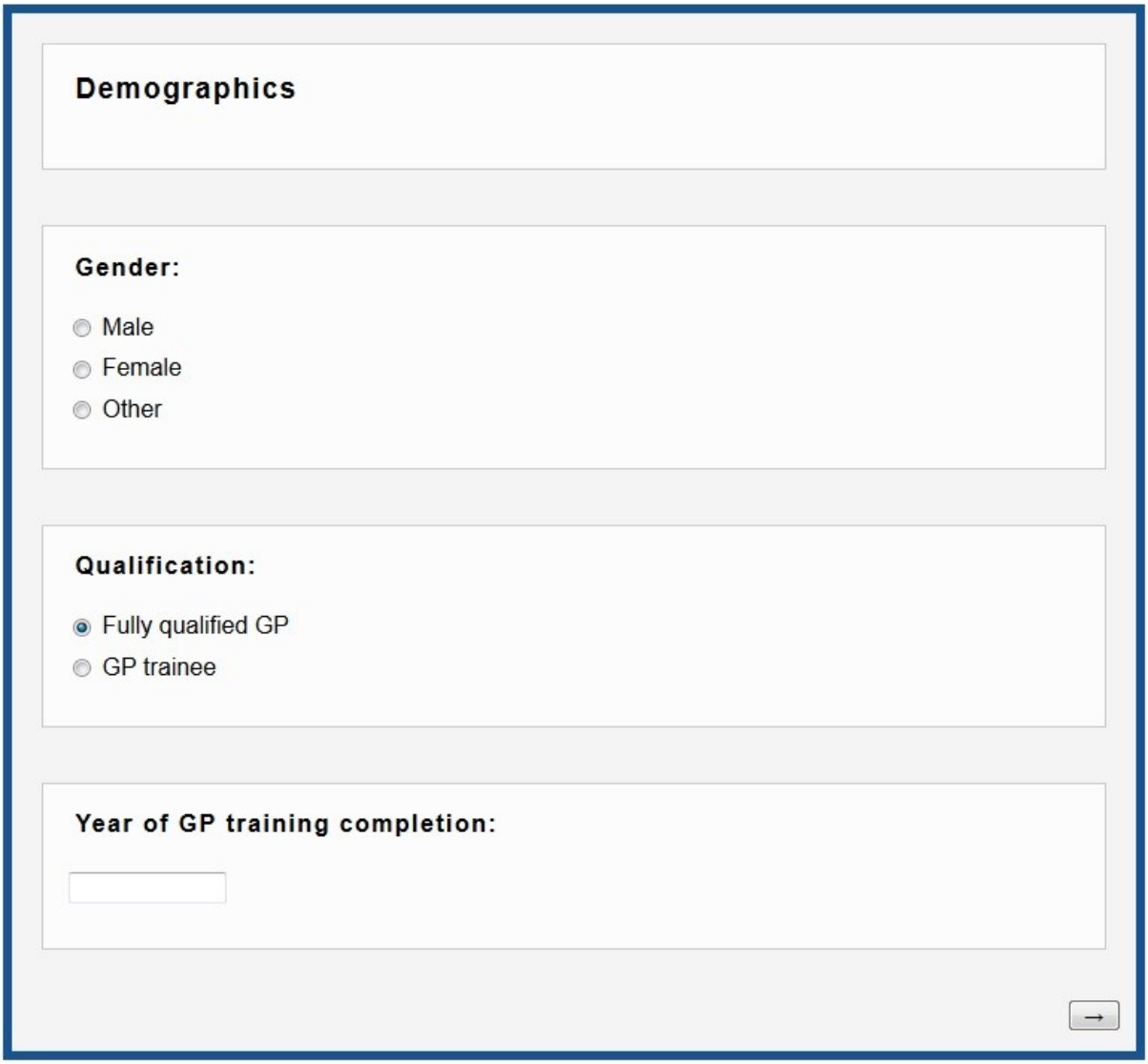
Version 9, 16/12/2019

→

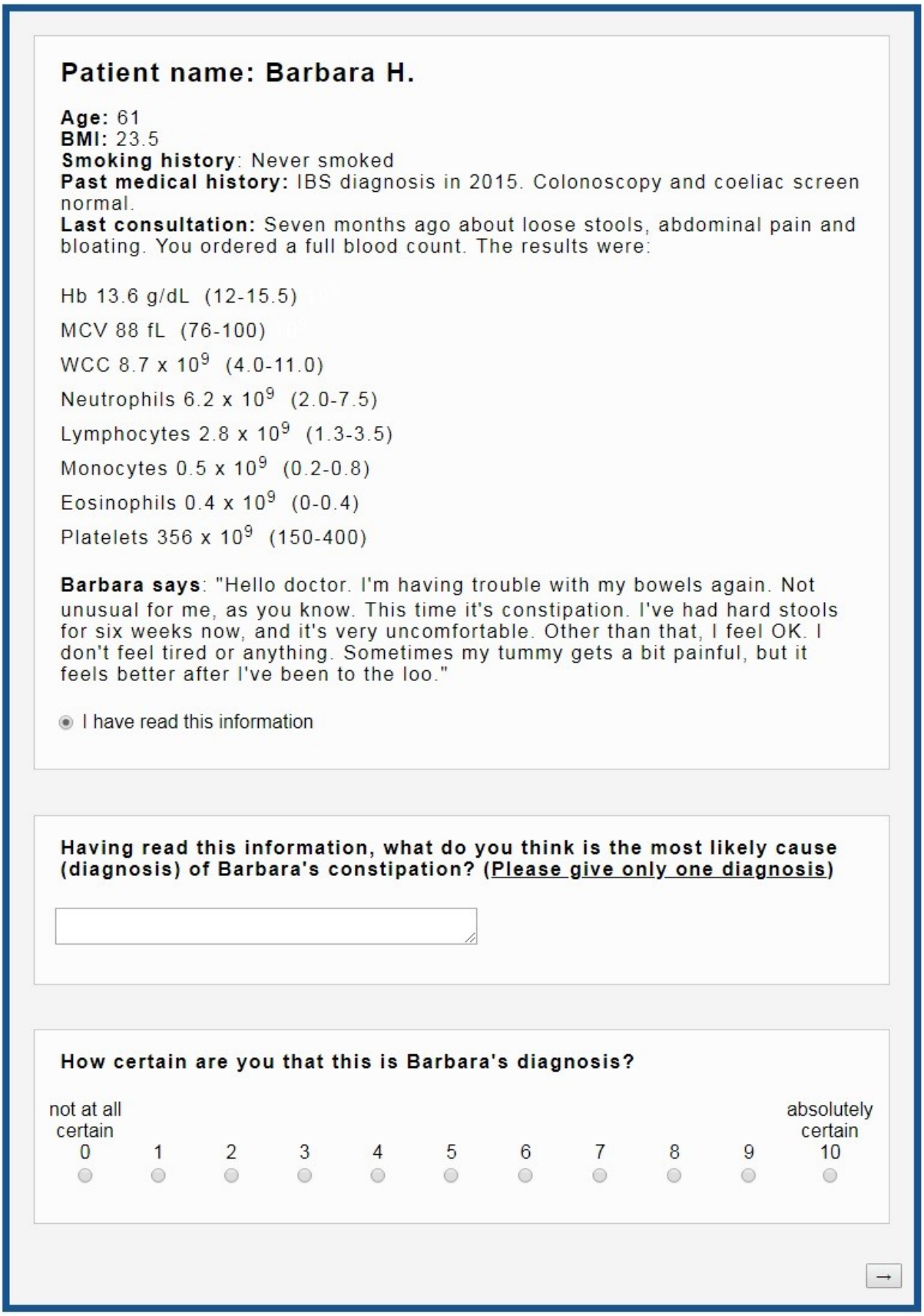
Screen 1 (Information sheet)



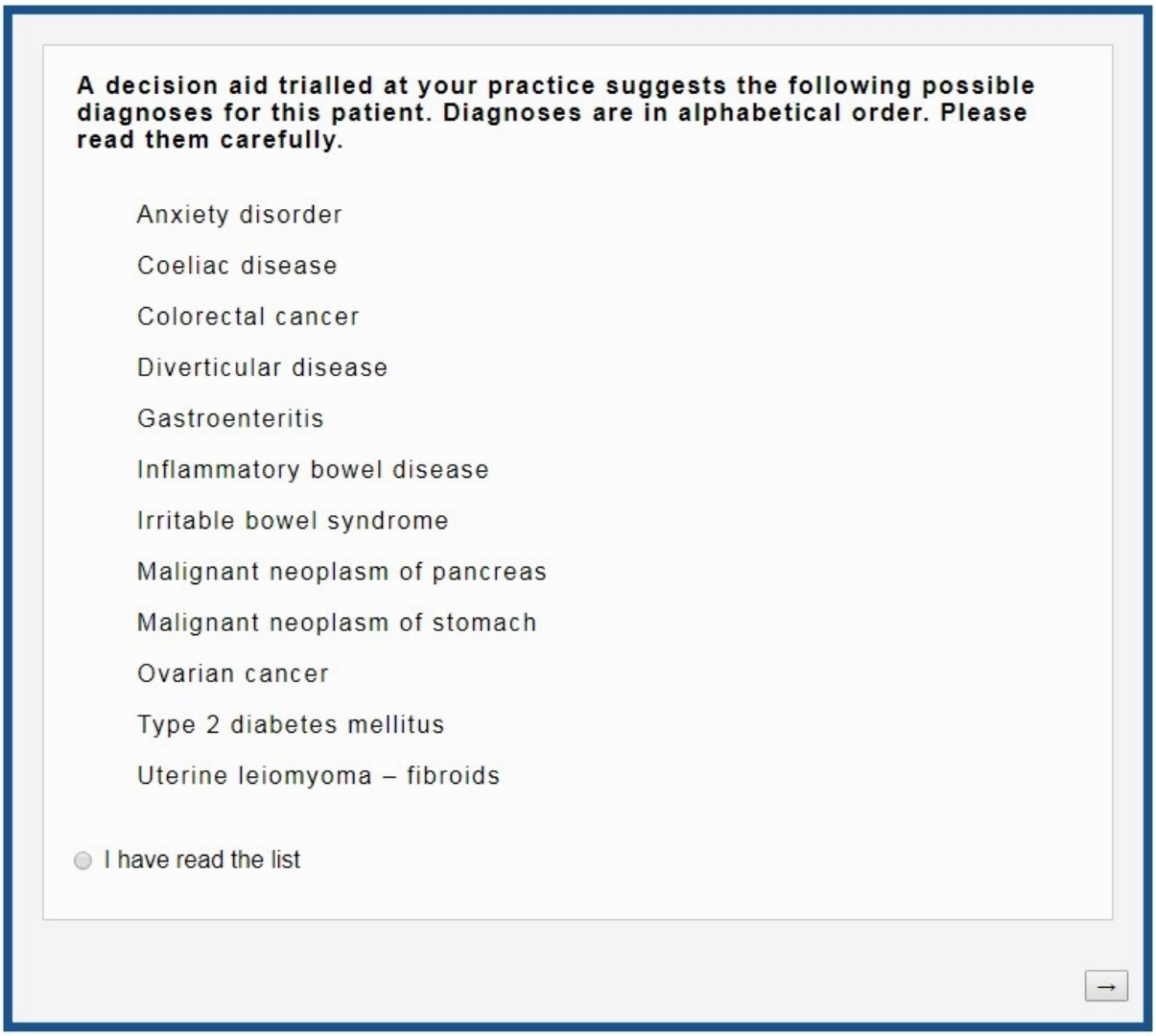
Screen 2 (Consent Form)



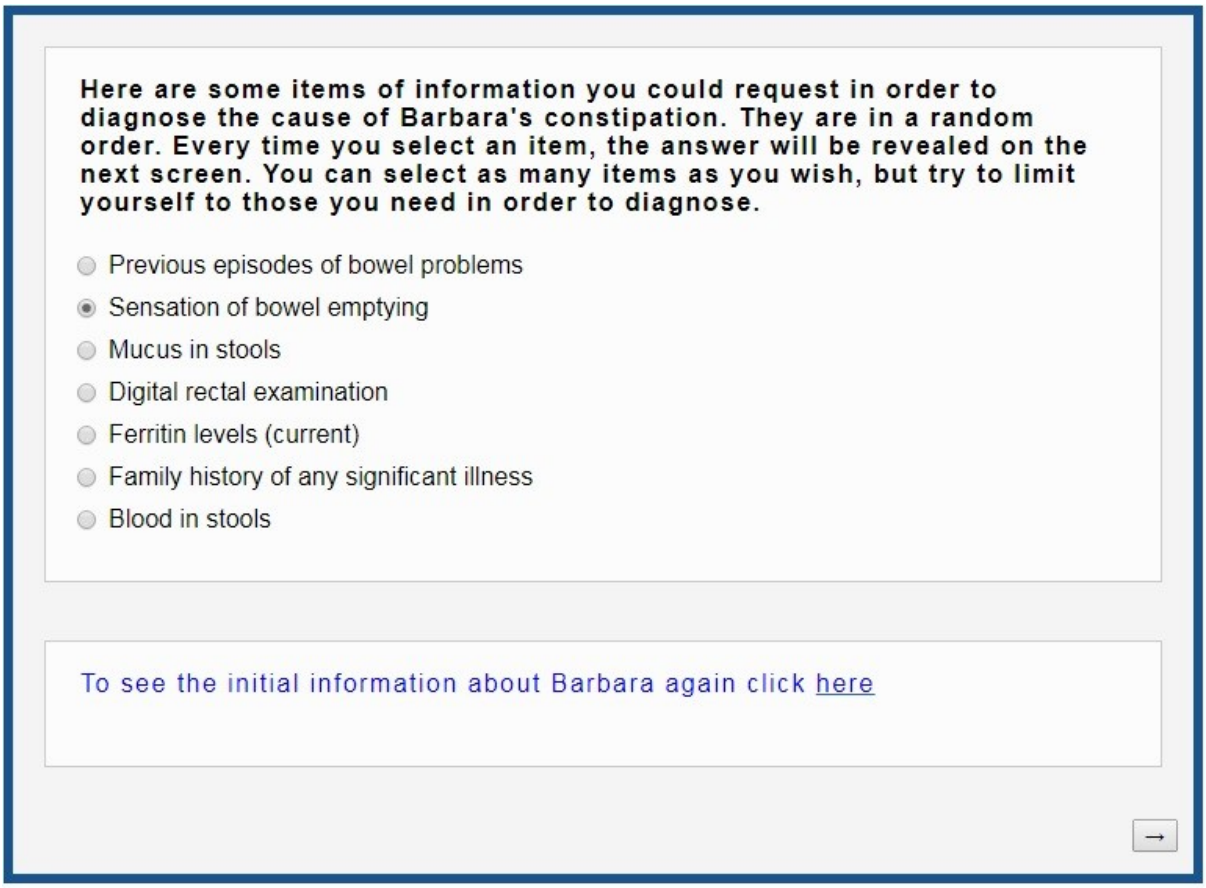
Screen 3 (Demographics)



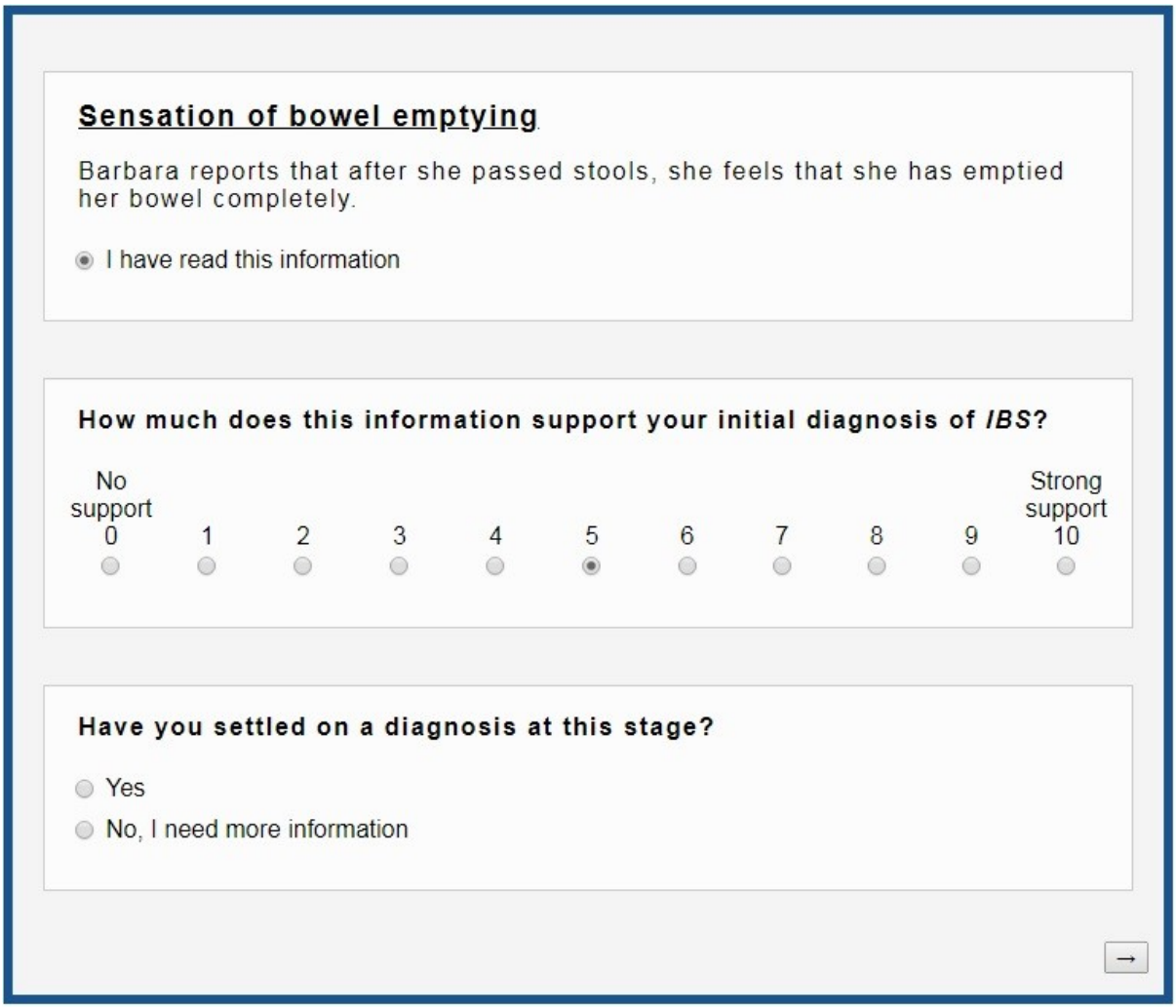
Screen 4 (initial patient description, initial diagnosis and level of confidence).



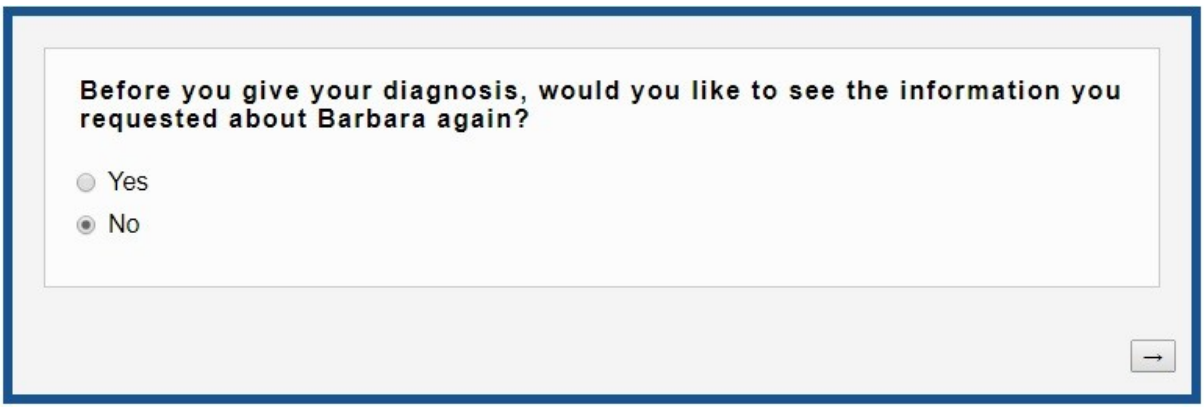
Screen 5 (The decision aid)



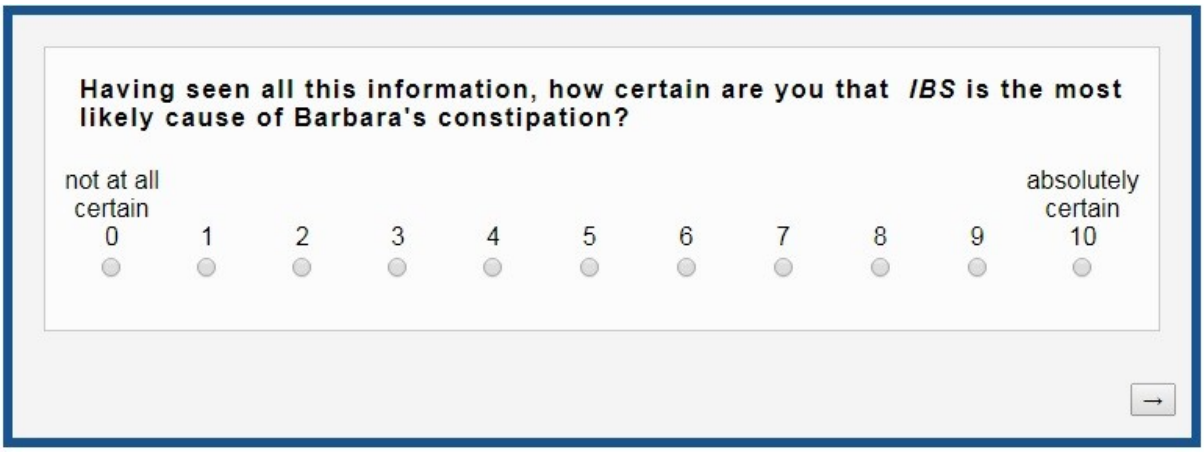
Screen 6 (List of clinical cues)



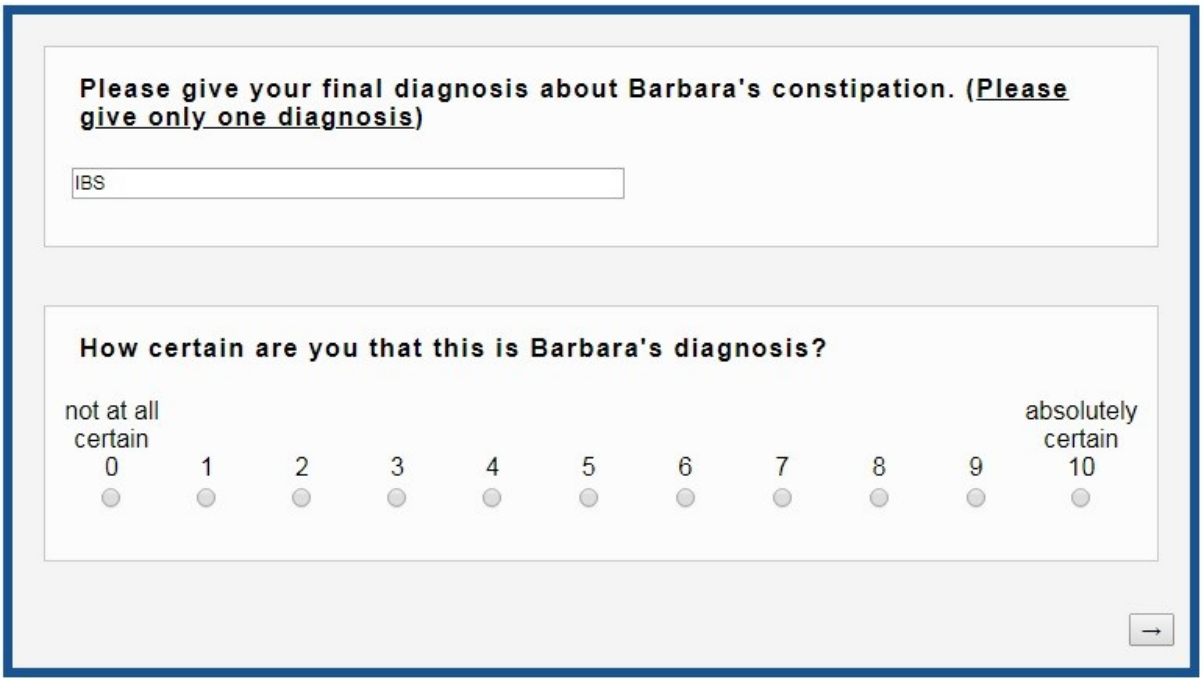
Screen 7 (Cue information and evaluation)



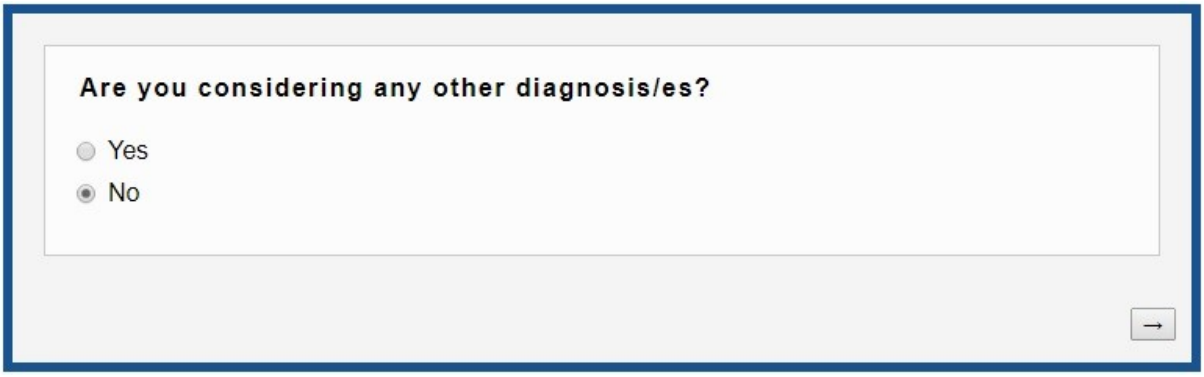
Screen 8 (Access to previous information before final diagnosis was given)



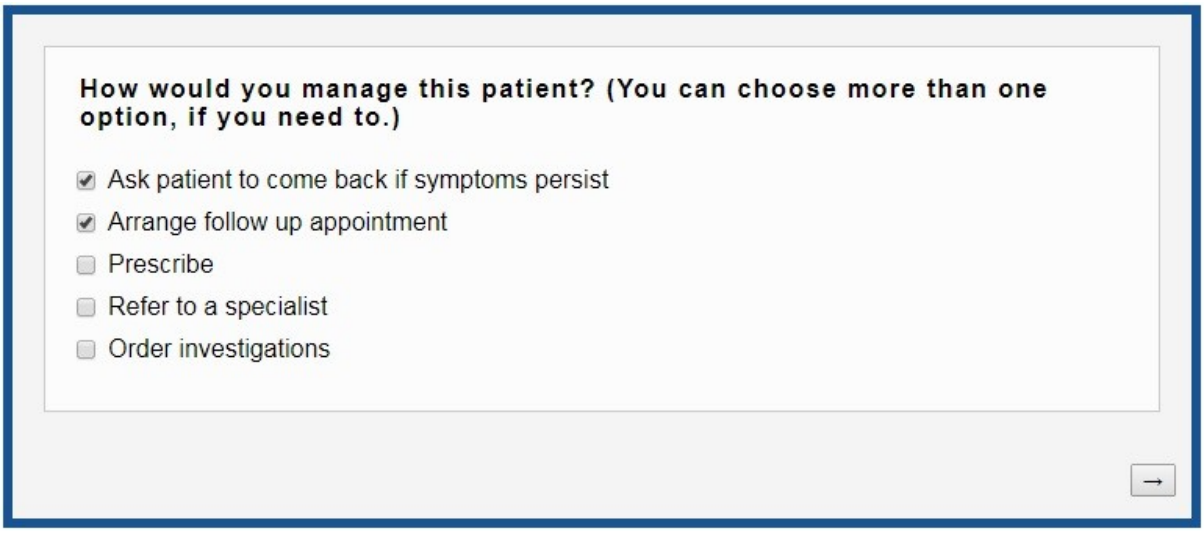
Screen 9 (Confidence in initial diagnosis – second measurement)



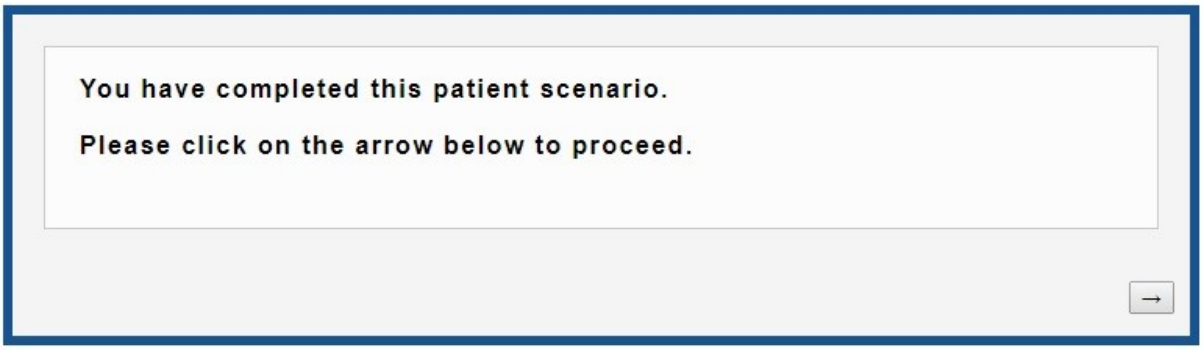
Screen 10 (Final diagnosis and confidence)



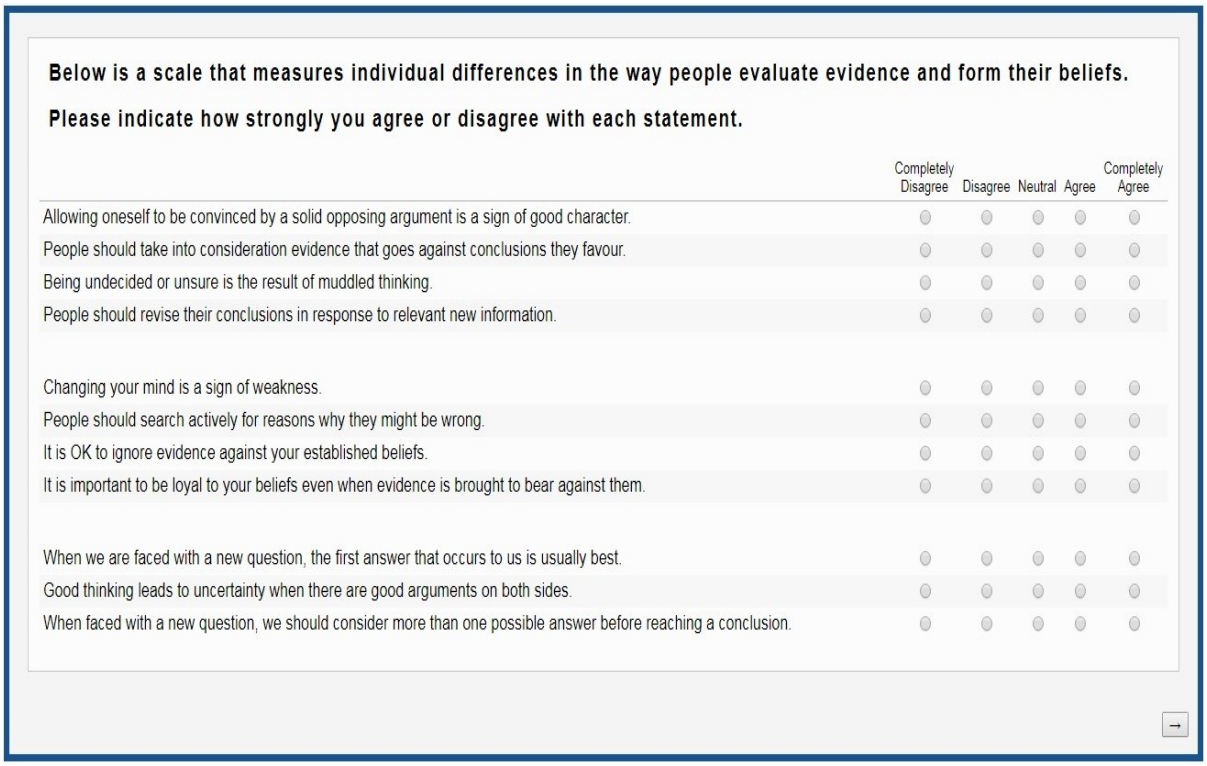
Screen 11 (Differential diagnoses)



Screen 12 (Patient's management).



Screen 13 (End of patient scenario)



Screen 14 (Actively Open-Minded Thinking Scale)

Debriefing

Study title: Influences of a diagnostic decision aid on clinical reasoning

The study was designed to explore how GPs evaluate clinical information during diagnosis and whether this evaluation is influenced by the provision of diagnostic suggestions.

Previous research has found that GPs tend to view neutral patient information (symptoms and signs) as supportive of their initial diagnostic hypothesis (e.g., https://journals.sagepub.com/doi/pdf/10.1177/0272989X12447241).

GPs' diagnostic accuracy improved when they received diagnostic suggestions early in the consultation (e.g., http://bjgp.org/content/bjgp/early/2017/01/30/bjgp16X688417.full.pdf).

In this study, we are trying to understand why this improvement occurred, by measuring how GPs evaluate information after the first diagnosis that comes to mind. We hypothesise that, when they receive suggestions about alternative diagnoses, GPs may evaluate neutral information as less supportive of their initial diagnosis than when they do not receive any suggestions. All participants saw the same two scenarios, one with and the other without suggestions. We will average responses across all our participants to compare responses on the scenarios with vs. on the scenarios without suggestions. We cannot draw any conclusions about you as an individual GP based on answers you provided on only two hypothetical scenarios, constrained in their information.

Due to the design of the study, you may feel that the diagnostic process has been narrowed down in an unrealistic way. For example, you may have felt reluctant to give a single diagnosis so early in the process and with so little information. Nevertheless, we needed your initial thoughts in order to measure how subsequent neutral patient information is evaluated, and we could only do this by asking you for a single diagnosis. Furthermore, you could only choose from seven pre-set items. In real life, you would take a more complete patient history before reaching a diagnostic conclusion, by asking open questions and/or different questions. Unfortunately, such constraints are necessary in order to reduce variability between respondents and ensure that your responses are analysable and meaningful for this study. Averaging responses across a large number of participants will allow us to compare scenarios with vs. without suggestions, despite the variability between individuals.

Thank you again for your time! If you have any further questions or comments, please email Dr Olga Kostopoulou (o.kostopoulou@imperial.ac.uk).

→

Screen 15 (Debriefing information)



Screen 16 (End of survey message).



Screen 17 (Participant's contact details to claim the Amazon voucher).

7.8. Appendix 8

7.8.1. Piloting for Study 1

Before conducting the actual study, we piloted the materials and procedure initially with two GPs known to the supervisors. GPs were invited to a one-to-one session. They were presented with three patient scenarios on a computer screen and were asked to assess them, thinking out loud. During the pilot, three different VAS were used to measure GPs' certainty (0-9, 0-10, 0-11). After the piloting, one of the patient scenarios was excluded, as it did not work as expected (i.e., the initial and additional patient information was too inconsistent, and this would possibly lead to a change in diagnosis irrespective of the provision of diagnostic suggestions). With regards to the certainty scales, all of them were found to be appropriate. We chose the 0-10 scale to keep it consistent with the Kostopoulou, Porat and colleagues (2017) study. Both sessions were audio-recorded with GPs' consent. The two GPs voluntarily spared their time to help us piloting the materials, without being paid for this.

After the first run of piloting, we invited another 11 GPs to pilot the updated materials online. GPs were sent an e-mail with information about the study and a link to access the study website. They were presented with two patient scenarios to diagnose and indicate their certainty on a 0-10 VAS. The results showed that there was variability between GPs' responses in both initial and final diagnosis. After the second run of piloting, we adjusted the two patient scenarios according to the clinical judgement of the second supervisor who is a practising GP (Professor Brendan Delaney). Specifically, in James' case (main symptom: chest pain), the symptom breathlessness was removed from the initial patient description, as it suggested a number of different diagnoses beyond MSK and Pleuritic chest pain (e.g., pneumothorax, infection, pulmonary embolism). In Derek's case (main symptom: breathlessness), the symptom dizziness was removed from the additional patient information as it suggested a heart disease. The patient scenarios described in Materials (see 2.4.4.) were the final versions assessed by the GPs in the actual study.

Finally, GPs who took part in the piloting were asked to comment on the list of diagnostic suggestions. Most of GPs found the list of suggestions “appropriate” and “reasonable”, noting that some diagnoses are very unlikely, or “out of context” based on the patient history (e.g., Tuberculosis, Herpes Zoster). Other GPs mentioned that “common and uncommon diseases are mixed on the same list which is confusing”, “common things come first in primary care” and that a lengthy list of possible diagnoses can “blur”, “distract” or “confuse” doctors’ judgement.

7.8.2. Piloting for Study 2

In Study 1, the first scenario (Derek) generally worked as expected. GPs who provided the expected initial diagnosis (COPD) changed it to a cardiac or a cancer diagnosis, when they received additional patient information. However, most GPs did not change their initial diagnosis. The other scenario (James) led to more diagnostic changes. However, most GPs provided “pleurisy” as an initial diagnosis, which is a symptom rather than a diagnosis.

As our intention was to have a clear initial diagnosis, we tried to adjust the scenarios to create a stronger leaning. Specifically, in Derek’s scenario (breathlessness) we did not change the initial patient description. From the additional information, we removed the cue about sleep (“Yesterday, I sat down on the sofa to watch TV and dozed off for a few minutes”) as it suggested sleep apnoea. We also added “light-headedness” to make the case more suggestive of a heart disease. From the list of suggestions, we removed the most unlikely diagnoses (e.g., allergic reaction, tonsilitis, obstructive sleep apnoea) and added heart failure, aortic or mitral valve disease and acute exacerbation of COPD. We also removed overlapping diagnoses that were covered by other ones (e.g., cor pulmonale and congestive cardiac failure were covered by heart failure).

In the “James” scenario (chest pain), we tried to steer GPs towards infection as an initial diagnosis and added a temperature of 37.5 °C in the patient description. From the additional patient information, we changed oxygen saturation from 98% to 95%, and respiratory rate from 20 to 30 in order to make it more suggestive of PE and

infection. We also removed the temperature feature from the additional patient information. From the list of suggestions, we removed the least likely diagnosis including angina, aortic valve regurgitation and stenosis, as well as cardiac arrythmias and Type 2 diabetes mellitus. We also removed rib fracture and lobar pneumonia as they were covered by MSK and acute LRTI respectively. We also used a third scenario (patient name “Barbara”, with the main symptom of constipation). Finally, we changed the study design from within-participant to between-participants to avoid potential spillover effects. We tested this with eight GPs and found no different results compared to the results from Study 1. As a result, we did not proceed with these materials and design.

We subsequently designed a new study to measure confirmation bias in physicians reasoning. We adapted two different scenarios, previously used in Kostopoulou *et al.*, 2012. Both scenarios were designed to initially steer physicians towards two possible diagnoses. One scenario was about chest pain (David) with possible diagnoses either MSK or angina, and the other scenario was about fatigue and weight loss (Patricia) with possible diagnoses either diabetes or depression. Physicians (GPs) had to indicate their preferred option between the two diagnoses (binary choice). They were then asked to choose between 10 questions to ask the patient in order to obtain further information. Half of the questions were about one diagnosis and the other half about the other diagnosis (e.g., question related to diabetes: “Have you been drinking more fluids than usual lately?”, question related to depression: “Have you been feeling sad lately?”). GPs were able to choose up to five questions to ask.

Our intention was to measure the number of questions chosen and whether they were related to GPs’ initial diagnostic hypothesis. This was to classify what their search strategy was (e.g., hypothesis-confirming or hypothesis-testing). We used a between participants design where GPs saw diagnostic suggestions only in one scenario. We piloted the materials with 15 GPs in two runs: the first run was focused on study flow, the scenarios, and the subsequent questions about the patient. The second run of piloting was focused on how GPs felt about their initial diagnosis after reading the suggestions and why they chose specific questions. In both pilots, we did

not find any effect of the diagnostic suggestions on GPs information searching strategy and, as a result, we did not proceed with this design.

After the second pilot, we designed a similar study to measure physicians' information search and information interpretation. We used the scenarios from the previous pilot and adapted them for the purposes of the new study. Specifically, we created two versions (the steers) of each scenario to allow for greater generalisability. Each version for each scenario was designed to include strong signals of a certain diagnosis so physicians could create an initial leaning. For instance, in the chest pain scenario (David), we manipulated the patient description so that one version suggested MSK, while the other version of the scenario suggested angina. Likewise, in the constipation scenario (Barbara), one version suggested IBS, while the other suggested colorectal cancer. Finally, in the fatigue/weight loss scenario (Patricia), one version of it suggested depression and the other suggested diabetes. We removed the binary choice for initial diagnosis and allowed a free-text response. We also reduced the number of diagnostic suggestions to 12 for each scenario, removing the least likely diagnoses.

The additional patient information was neutral and open to interpretation and was presented on a list of seven clinical cues. Physicians were able to request more details about each cue and rate how much this information supports their initial diagnosis. As before, we used a between-participants design. We tested the materials with 10 GPs over the phone while completing the online task. The scenarios worked as expected and the procedure was clear.

We removed one of the scenarios (Patricia), because our intention was to include "severity of initial diagnosis" as a second predictor. For the other two scenarios (chest pain and constipation) it was clear that there was a serious and a less serious diagnosis (i.e., MSK versus angina and IBS versus cancer). The fatigue/weight loss scenario implied either a physical or a psychological cause (i.e., diabetes versus depression) and it was not clear which one was more serious. As a result, we decided to exclude this scenario. We also changed the design from between-participants to within-participant to reduce the required sample size by 50%. Since there was a possibility of spill over effects, as found in Study 1, we decided to

provide the diagnostic suggestions always in the scenario seen last. We tested the materials and design online with another 10 GPs. We then finalised the materials and procedure, as described in 3.4.4. and 3.4.6.

7.8.3. Piloting for Study 3

Since Studies 1 and 2 did not show the expected results, we decided to use eye tracking methodology to explore in more depth why the decision aid did not work. The initial design would include the scenarios used in Study 2 to be assessed by Imperial's Clinical Research Fellows. It would be a lab-experiment where participants would be taking the decision task while an eye tracker is recording their eye movement. We intended to measure the time participants spend on reading the list of diagnostic suggestions as a proxy for deliberate engagement and whether participants focus on their own diagnosis, as a proxy for confirmation bias. However, due to the COVID-19 pandemic and the relevant restrictions, this plan never materialised.

Patient and Public Involvement (PPI)

Study 1, as well as an introduction about this research work was presented at Imperial Festival (April 2018), where several research groups across Imperial College London collaborated to create a Patient and Public Involvement (PPI) stand. On this stand, members of the public were invited to comment on and contribute to ongoing research. We took part in this stand, where we had the opportunity to talk to the public about the research project. We explained how the decision aid operates and what its effects on physicians' thinking could be. We also asked for any thoughts, concerns, and feedback that they might have. We had also prepared a set of questions about people's attitudes toward the use of decision aids by physicians (e.g., "Who would you trust more, physicians' knowledge and intuition or a computerised decision aid?", "Do you think that physicians should use decision aids?", "Do you have any concerns about the use of decision aids during the consultation?").

We collected 29 written responses from patients and the public. Most of the people who answered the questions said that they would either trust physicians' knowledge and intuition over a computerised decision aid or a combination of physicians' input and advice from the aid. Regarding the use of decision aids by physicians, the large majority of the people said that physicians should use decision aids to help them diagnose. However, they reported that the decision aid should be used as a "safety net" and that the "final decision" should be made by physicians' who have to be held "accountable".

References

Ahlawat, S. S. (1999) ‘Order effects and memory for evidence in individual versus group decision making in auditing’, *Journal of Behavioral Decision Making*, 12(1), pp. 71–88. doi: 10.1002/(SICI)1099-0771(199903)12:1<71::AID-BDM308>3.0.CO;2-Q.

Ames, D. R. *et al.* (2010) ‘Not so fast: The (not-quite-complete) dissociation between accuracy and confidence in thin-slice impressions’, *Personality and Social Psychology Bulletin*, 36(2), pp. 264–277. doi: 10.1177/0146167209354519.

Arkes, H. R. (1981) ‘Impediments to accurate clinical judgment and possible ways to minimize their impact’, *Journal of Consulting and Clinical Psychology*, 49(3), pp. 323–330. doi: 10.1037/0022-006X.49.3.323.

Arkes, H. R. *et al.* (1988) ‘Eliminating the Hindsight Bias’, *Journal of Applied Psychology*, 73(2), pp. 305–307. doi: 10.1037/0021-9010.73.2.305.

Arkes, H. R. (2001) ‘Overconfidence in Judgmental Forecasting’, in Armstrong, J. S. (ed.) *Principles of forecasting*. Boston, MA: Springer. doi: 10.1007/978-0-306-47630-3_22.

Arkes, H. R. (2013) ‘The Consequences of the Hindsight Bias in Medical Decision Making’, *Current Directions in Psychological Science*, 22(5), pp. 356–360. doi: 10.1177/0963721413489988.

Bago, B. and De Neys, W. (2017) ‘Fast logic?: Examining the time course assumption of dual process theory’, *Cognition*, 158, pp. 90–109. doi: 10.1016/j.cognition.2016.10.014.

Bajšanski, I. and Žauhar, V. (2019) ‘The Relationship between Consistency and Consensuality in Syllogistic Reasoning’, *Psychological Topics*, 28(1), pp. 73–91. doi: 10.31820/pt.28.1.4.

Barber, B. M. and Odean, T. (2001) ‘Boys Will be Boys: Gender, Overconfidence, and Common Stock Investment’, *The Quarterly Journal of Economics*, 116(1), pp. 261–292. doi: 10.1162/003355301556400.

Baron, J. (2006) *Thinking and Deciding*. Cambridge, MA: Cambridge University Press.

Baron, J. (2018) *Actively open-minded thinking in politics*. Unpublished manuscript.

Baron, J. (2019) 'Actively open-minded thinking in politics', *Cognition*, 188, pp. 8–18. doi: 10.1016/j.cognition.2018.10.004.

Baumann, A. O., Deber, R. B. and Thompson, G. G. (1991) 'Overconfidence among physicians and nurses: The "micro-certainty, macro-uncertainty" phenomenon', *Social Science and Medicine*, 32(2), pp. 167–174. doi: 10.1016/0277-9536(91)90057-J.

Bean, R. and Bean, W. (1950) *Sir William Osler: Aphorisms from his bedside teachings and writings*. New York, NY: Henry Schuman.

Bell, P. and Jamieson, B. (1970) 'Publicity of Risky Initial Shift Decisions Phenomenon and the Risky shift Phenomenon', *Journal of Experimental Social Psychology*, 6(3), pp. 329–345. doi: 10.1016/0022-1031(70)90067-3.

Van Den Berge, K. and Mamede, S. (2013) 'Cognitive diagnostic error in internal medicine', *European Journal of Internal Medicine*. European Federation of Internal Medicine., 24(6), pp. 525–529. doi: 10.1016/j.ejim.2013.03.006.

Bergeron, D. *et al.* (2016) 'Clinical Impact of a Second FDG-PET in Atypical/Unclear Dementia Syndromes', *Journal of Alzheimer's Disease*, 49(3), pp. 695–705. doi: 10.3233/JAD-150302.

Berner, E. S. *et al.* (1999) 'Effects of a decision support system on physicians' diagnostic performance', *Journal of the American Medical Informatics Association*, 6(5), pp. 420–427. doi: 10.1136/jamia.1999.0060420.

Berner, E. S. and Graber, M. L. (2008) 'Overconfidence as a cause of diagnostic error in medicine', *American Journal of Medicine*, 121(5), pp. 2–23. doi: 10.1016/j.amjmed.2008.01.001.

Berner, E. S. and Maisiak, R. S. (1999) 'Influence of case and physician characteristics on perceptions of decision support systems', *Journal of the American Medical Informatics Association*, 6(5), pp. 428–434. doi: 10.1136/jamia.1999.0060428.

Bhandari, G., Hassanein, K. and Deaves, R. (2008) 'Debiasing investors with decision support systems: An experimental investigation', *Decision Support Systems*, 46(1), pp. 399–410. doi: 10.1016/j.dss.2008.07.010.

Bhasale, A. (1998) ‘The wrong diagnosis: identifying causes of potentially adverse events in general practice using incident monitoring’, *Family Practice*, 15(4), pp. 308–318.

Boccardi, M. *et al.* (2016) ‘Assessment of the incremental diagnostic value of florbetapir F 18 imaging in patients with cognitive impairment: The incremental diagnostic value of amyloid PET with [18F]-florbetapir (INDIA-FBP) study’, *JAMA Neurology*, 73(12), pp. 1417–1424. doi: 10.1001/jamaneurol.2016.3751.

Bolger, F. *et al.* (2011) ‘Does the Delphi process lead to increased accuracy in group-based judgmental forecasts or does it simply induce consensus amongst judgmental forecasters?’, *Technological Forecasting and Social Change*, 78(9), pp. 1671–1680. doi: 10.1016/j.techfore.2011.06.002.

Borum, R., Otto, R. and Golding, S. (1993) ‘Improving clinical judgment and decision making in forensic evaluation’, *Journal of Psychiatry and Law*, 21(1), pp. 35–76. doi: 10.1177/009318539302100104.

Brezis, M. *et al.* (2019) ‘Does Physician’s Training Induce Overconfidence That Hampers Disclosing Errors?’, *Journal of Patient Safety*, 15(4), pp. 296–298. doi: 10.1097/PTS.0000000000000255.

Brussel van, S. *et al.* (2020) ‘“Consider the Opposite” – Effects of elaborative feedback and correct answer feedback on reducing confirmation bias – A pre-registered study’, *Contemporary Educational Psychology*, 60, p. 101844. doi: 10.1016/j.cedpsych.2020.101844.

Burnstein, E. and Katz, S. (1971) ‘Individual commitment to risky and conservative choices as a determinant of shifts in group decisions’, *Journal of Personality*, 39(4), pp. 564–580. doi: 10.1111/j.1467-6494.1971.tb00063.x.

Car, J. *et al.* (2020) ‘Video consultations in primary and specialist care during the covid-19 pandemic and beyond’, *The BMJ*, 371, pp. 1–6. doi: 10.1136/bmj.m3945.

Cavalcanti, R. B. and Sibbald, M. (2014) ‘Am I right when I am sure? Data consistency influences the relationship between diagnostic accuracy and certainty’, *Academic Medicine*, 89(1), pp. 107–113. doi: http://dx.doi.org/10.1097/ACM.0000000000000074.

Ceccaldi, M. *et al.* (2018) 'Added value of F-18-florbetaben amyloid PET in the diagnostic workup of most complex patients with dementia in France: A naturalistic study', *Alzheimer's & Dementia*, 14(3), pp. 293–305. doi: 10.1016/j.jalz.2017.09.009.

Choi, J. . *et al.* (2008) 'Need for Cognitive Closure and Information Search Strategy', *Psychology & Marketing*, 25(11), pp. 1027–1042. doi: 10.1002/mar.

Cipriano, M. and Gruca, T. S. (2014) 'The power of priors: How confirmation bias impacts market prices', *The Journal of Prediction Markets*, 8(3), pp. 34–56. doi: 10.5750/jpm.v8i3.974.

Couchman, J. J. *et al.* (2016) 'The instinct fallacy: the metacognition of answering and revising during college exams', *Metacognition and Learning*, 11(2), pp. 171–185. doi: 10.1007/s11409-015-9140-8.

Croskerry, P. and Norman, G. (2008) 'Overconfidence in Clinical Decision Making', *American Journal of Medicine*, 121(5A), pp. 24–29. doi: 10.1016/j.amjmed.2008.02.001.

Cross, C. P. *et al.* (2017) 'Sex differences in confidence influence patterns of conformity', *British Journal of Psychology*, 108(4), pp. 655–667. doi: 10.1111/bjop.12232.

Dani, M., Bowen-Carpenter, S. and McGown, P. J. (2019) 'Not All Strokes Are Strokes - An Example of Diagnostic Confirmation Bias', *European Journal of Case Reports in Internal Medicine*, 6(1). doi: 10.12890/2019_001006.

Delaney, B. (2015) *TRANSForm*. Available at: https://www.i-hd.eu/index.cfm/resources/ec-projects-results/transform/.

Dreiseitl, S. and Binder, M. (2005) 'Do physicians value decision support? A look at the effect of decision support systems on physician opinion', *Artificial intelligence in medicine*, 33(1), pp. 25–30. doi: https://dx.doi.org/10.1016/j.artmed.2004.07.007.

Dujmovic, M. *et al.* (2020) 'The role of representativeness in reasoning and metacognitive processes: an in-depth analysis of the Linda problem', *Thinking & Reasoning*, pp. 1–26. doi: 10.1080/13546783.2020.1746692.

Edwards, W. and Fasolo, B. (2001) 'Decision Technology', *Annual Review of*

Psychology, 52, pp. 581–606.

Einhorn, H. J. and Hogarth, R. M. (1978) 'Confidence in judgment: Persistence of the illusion of validity', *Psychological Review*, 85(5), pp. 395–416. doi: 10.1037/0033-295X.85.5.395.

Epstein, A. S. *et al.* (2021) 'Development of the Oncolo-GIST ("Giving Information Strategically & Transparently") Intervention Manual for Oncologist Skills Training in Advanced Cancer Prognostic Information Communication', *Journal of Pain and Symptom Management*. Elsevier Inc, 62(1), p. 10–19.e4. doi: 10.1016/j.jpainsymman.2020.11.023.

Eva, K. W. *et al.* (2010) 'Swapping Horses Midstream: Factors Related to Physicians' Changing Their Minds About a Diagnosis', *Academic Medicine*, 85(7), pp. 1112–1117. doi: 10.1016/j.immuni.2010.12.017.Two-stage.

Feinstein, A. R. (1990) 'The inadequacy of binary models for the clinical reality of three-zone diagnostic decisions', *Journal of Clinical Epidemiology*, 43(1), pp. 109–113. doi: 10.1016/0895-4356(90)90064-V.

Ferrarin, M. *et al.* (2008) 'Does instrumented movement analysis alter, objectively confirm, or not affect clinical decision-making in musicians with focal dystonia?', *Medical Problems of Performing Artists*, 23(3), pp. 99–106.

Ferrarin, M. *et al.* (2015) 'Does gait analysis change clinical decision-making in poststroke patients? Results from a pragmatic prospective observational study', *European Journal of Physical and Rehabilitation Medicine*, 51, pp. 171–184.

Festinger, L. (1957) *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Fisseni, G., Pentzek, M. and Abholz, H. H. (2008) 'Responding to serious medical error in general practice - consequences for the GPs involved: analysis of 75 cases from Germany', *Family Practice*, 25(1), pp. 9–13.

Folke, T. *et al.* (2017) 'Explicit representation of confidence informs future value-based decisions', *Nature Human Behaviour*, 1(2), pp. 1–32. doi: 10.1038/s41562-016-0002.

Frederick, S. (2005) 'Cognitive reflection and decision making', *Journal of Economic Perspectives*, 19(4), pp. 25–42. doi: 10.1257/089533005775196732.

Friedman, C. P. *et al.* (2001) 'Are clinicians correct when they believe they are correct? Implications for medical decision support', *Studies in Health Technology and Informatics*, 84(1), pp. 454–458. doi: 10.3233/978-1-60750-928-8-454.

Friedman, C. P. *et al.* (2005) 'Do physicians know when their diagnoses are correct? Implications for decision support and error reduction', *Journal of General Internal Medicine*, 20(4), pp. 334–339. doi: 10.1111/j.1525-1497.2005.30145.x.

Gail, S. (2000) 'Decision support for what-if analysis and the confirmation bias', *Journal of Computer Information Systems*, 40(4), pp. 84–92.

Gigerenzer, G. (2015) *Risk Savvy: How to make good decisions*. London, UK: Penguin.

Gilovich, T. (1983) 'Biased evaluation and persistence in gambling', *Journal of Personality and Social Psychology*, 44(6), pp. 1110–1126. doi: 10.1037/0022-3514.44.6.1110.

Goodie, A. S. (2005) 'The role of perceived control and overconfidence in pathological gambling', *Journal of Gambling Studies*, 21(4), pp. 481–502. doi: 10.1007/s10899-005-5559-1.

Greenhalgh, T. *et al.* (2004) 'Diffusion of innovations in service organizations: Systematic review and recommendations', *Milbank Quarterly*, 82(4), pp. 581–629. doi: 10.1111/j.0887-378X.2004.00325.x.

Griffiths, L. and Higham, P. A. (2018) 'Beyond hypercorrection: remembering corrective feedback for low-confidence errors', *Memory*, 26(2), pp. 201–218. doi: 10.1080/09658211.2017.1344249.

Harada, Y. *et al.* (2021) 'Effects of a Differential Diagnosis List of Artificial Intelligence on Differential Diagnoses by Physicians: An Exploratory Analysis of Data from a Randomized Controlled Study', *International Journal of Environmental Research and Public Health*, 18(5562), pp. 1–8. doi: 10.3390/ijerph18115562.

Haran, U., Moore, D. A. and Morewedge, C. K. (2010) 'A simple remedy for overprecision in judgment', *Judgment and Decision Making*, 5(7), pp. 467–476.

Haran, U., Ritov, I. and Mellers, B. a. (2013) 'The role of actively open-minded thinking in information acquisition, accuracy, and calibration', *Judgment and*

Decision Making, 8(3), pp. 188–201. doi: 10.1017/CBO9781107415324.004.

Hirt, E. R. and Markman, K. D. (1995) 'Multiple Explanation: A Consider-an-Alternative Strategy for Debiasing Judgments', *Journal of Personality and Social Psychology*, 69(6), pp. 1069–1086. doi: 10.1037/0022-3514.69.6.1069.

Hogarth, R. M. and Einhorn, H. J. (1992) 'Order effects in belief updating: The belief-adjustment model', *Cognitive Psychology*, 24(1), pp. 1–55. doi: 10.1016/0010-0285(92)90002-J.

Holyoak, K. J. and Simon, D. (1999) 'Bidirectional reasoning in decision making by constraint satisfaction.', *Journal of Experimental Psychology: General*, 128(1), pp. 3–31. doi: 10.1037/0096-3445.128.1.3.

Horstmann, N., Ahlgrimm, A. and Glöckner, A. (2009) 'How distinct are intuition and deliberation? An eye-tracking analysis of instruction-induced decision modes', *Judgment and Decision Making*, 4(5), pp. 335–354. doi: 10.2139/ssrn.1393729.

Huang, H. H., Hsu, J. S. C. and Ku, C. Y. (2012) 'Understanding the role of computer-mediated counter-argument in countering confirmation bias', *Decision Support Systems*, 53(3), pp. 438–447. doi: 10.1016/j.dss.2012.03.009.

IoM (2015) *Improving Diagnosis in Health Care*. Washington, DC: The National Academies Press. doi: 10.17226/21794.

Jaimes, N. *et al.* (2013) 'Influence of time on dermoscopic diagnosis and management', *Australasian Journal of Dermatology*, 54(2), pp. 96–104. doi: 10.1111/ajd.12001.

Jermias, J. (2006) 'The influence of accountability on overconfidence and resistance to change: A research framework and experimental evidence', *Management Accounting Research*, 17(4), pp. 370–388. doi: 10.1016/j.mar.2006.03.003.

Kahneman, D. (2003) 'A Perspective on Judgment and Choice: Mapping Bounded Rationality', *American Psychologist*, 58(9), pp. 697–720. doi: 10.1037/0003-066X.58.9.697.

Kahneman, D. (2011) *Thinking, Fast and Slow*. London: Penguin.

Kahneman, D. and Tversky, A. (1974) 'Judgment under uncertainty: Heuristics and biases', *Science*, 185(4157), pp. 1124–1131. doi: 10.1017/CBO9780511809477.002.

Kassin, S. M., Dror, I. E. and Kukucka, J. (2013) 'The forensic confirmation bias: Problems, perspectives, and proposed solutions', *Journal of Applied Research in Memory and Cognition*, 2(1), pp. 42–52. doi: 10.1016/j.jarmac.2013.01.001.

Kassirer, J. (1989) 'Our stubborn quest for diagnostic certainty. A cause of excessive testing', *The New England journal of medicine*, 320(22), pp. 1489–1491.

Katz, J. (1984) 'Why Doctors Don't Disclose Uncertainty', *The Hastings Center Report*, 14(1), pp. 35–44.

Klein, G. (2007) 'Performing a Project Premortem', *Harvard Business Review*. Available at: https://hbr.org/2007/09/performing-a-project-premortem.

Koriat, A., Lichtenstein, S. and Fischhoff, B. (1980) 'Reasons for Confidence', *Journal of Experimental Psychology: Human Learning and Memory*, 6(2), pp. 107–118.

Kostopoulou, O. *et al.* (2012) 'Information Distortion in Physicians' Diagnostic Judgments', *Medical Decision Making*, 32(6), pp. 831–839. doi: 10.1177/0272989X12447241.

Kostopoulou, O., Lionis, C., *et al.* (2015) 'Early diagnostic suggestions improve accuracy of family physicians: A randomized controlled trial in Greece', *Family Practice*, 32(3), pp. 323–328. doi: 10.1093/fampra/cmv012.

Kostopoulou, O., Rosen, A., *et al.* (2015) 'Early diagnostic suggestions improve accuracy of GPs: a randomised controlled trial using computer-simulated patients', *British Journal of General Practice*, 65(630), pp. 49–54. doi: 10.3399/bjgp15X683161.

Kostopoulou, O., Porat, T., *et al.* (2017) 'Diagnostic accuracy of GPs when using an early-intervention decision support system: A high-fidelity simulation', *British Journal of General Practice*, 67(656), pp. 201–208. doi: 10.3399/bjgp16X688417.

Kostopoulou, O., Sirota, M., *et al.* (2017) 'The Role of Physicians' First Impressions in the Diagnosis of Possible Cancers without Alarm Symptoms', *Medical Decision Making*, 37(1), pp. 9–16. doi: 10.1177/0272989X16644563.

Kostopoulou, O., Delaney, B. and Munro, C. W. (2008) 'Diagnostic difficulty and error in primary care - A systematic review', *Family Practice*, 25(6), pp. 400–413. doi:

10.1093/fampra/cmn071.

Kostopoulou, O., Mousoulis, C. and Delaney, B. (2009) 'Information search and information distortion in the diagnosis of an ambiguous presentation', *Judgment and Decision Making*, 4(5), pp. 408–418.

Krudop, W. A. *et al.* (2016) 'Impact of Imaging and Cerebrospinal Fluid Biomarkers on Behavioral Variant Frontotemporal Dementia Diagnosis within a Late-Onset Frontal Lobe Syndrome Cohort', *Dementia and Geriatric Cognitive Disorders*, 41(1–2), pp. 16–26. doi: 10.1159/000441023.

Kruglanski, A. W. and Webster, D. M. (1996) 'Motivated closing of the mind: "seizing" and "freezing"', *Psychological review*, 103(2), pp. 263–283. doi: 10.4324/9781315175867-3.

Kruglanski, A. W., Webster, D. M. and Klem, A. (1993) 'Motivated Resistance and Openness to Persuasion in the Presence or Absence of Prior Information', *Journal of Personality and Social Psychology*, 65(5), pp. 861–876. doi: 10.1037//0022-3514.65.5.861.

Krupat, E. *et al.* (2017) 'Avoiding premature closure and reaching diagnostic accuracy: some key predictive factors', *Medical Education*, 51(11), pp. 1127–1137. doi: 10.1111/medu.13382.

Kumar, B., Kanna, B. and Kumar, S. (2011) 'The pitfalls of premature closure: Clinical decision-making in a case of aortic dissection', *BMJ Case Reports*, pp. 8–11. doi: 10.1136/bcr.08.2011.4594.

Lambe, K. A., Hevey, D. and Kelly, B. D. (2018) 'Guided Reflection Interventions Show No Effect on Diagnostic Accuracy in Medical Students', *Frontiers in Psychology*, 9, pp. 1–9. doi: 10.3389/fpsyg.2018.02297.

Larrick, R. P. (2004) 'Debiasing', in Koehler, D. J. and Harvey, N. (eds) *Blackwell Handbook of Judgment and Decision Making*. Oxford, UK: Blackwell Publishing, pp. 316–337.

Lau, A. Y. S. and Coiera, E. W. (2008) 'Impact of web searching and social feedback on consumer decision making: A prospective online experiment', *Journal of Medical Internet Research*, 10(1), pp. 1–11. doi: 10.2196/jmir.963.

Leblanc, V., Brooks, L. R. and Norman, G. R. (2002) 'Believing is seeing: The

influence of a diagnostic hypothesis on the interpretation of clinical features', *Academic Medicine*, 77(10), pp. 67–69. doi: 10.1097/00001888-200210001-00022.

Leblanc, V., Norman, G. R. and Brooks, L. R. (2001) 'Effect of a diagnostic suggestion on diagnostic accuracy and identification of clinical features', *Academic Medicine*, 76(10), pp. 18–20. doi: 10.1097/00001888-200110001-00007.

Lewis, L. M. *et al.* (2007) 'Quantifying the usefulness of CT in evaluating seniors with abdominal pain', *European Journal of Radiology*, 61(2), pp. 290–296. doi: 10.1016/j.ejrad.2006.09.014.

Lidén, M., Gräns, M. and Juslin, P. (2019) 'From devil's advocate to crime fighter: confirmation bias and debiasing techniques in prosecutorial decision-making', *Psychology, Crime and Law*, 25(5), pp. 494–526. doi: 10.1080/1068316X.2018.1538417.

Ludolph, R. and Schulz, P. J. (2018) 'Debiasing Health-Related Judgments and Decision Making: A Systematic Review', *Medical Decision Making*, 38(1), pp. 3–13. doi: 10.1177/0272989X17716672.

Marewski, J. N. and Gigerenzer, G. (2012) 'Heuristic decision making in medicine', *Dialogues in Clinical Neuroscience*, 14(1), pp. 77–89. doi: 10.31887/DCNS.2012.14.1/jmarewski.

McKenzie, C. R. M. (1997) 'Underweighting alternatives and overconfidence', *Organizational Behavior and Human Decision Processes*, 71(2), pp. 141–160. doi: 10.1006/obhd.1997.2716.

McParland, C. R., Cooper, M. A. and Johnston, B. (2019) 'Differential Diagnosis Decision Support Systems in Primary and Out-of-Hours Care: A Qualitative Analysis of the Needs of Key Stakeholders in Scotland', *Journal of Primary Care and Community Health*, 10, pp. 57–61. doi: 10.1177/2150132719829315.

Medow, M. A., Arkes, H. R. and Shaffer, V. A. (2010) 'Are residents' decisions influenced more by a decision aid or a specialist's opinion? A randomized controlled trial', *Journal of General Internal Medicine*, 25(4), pp. 316–320. doi: 10.1007/s11606-010-1251-y.

Mello, J. *et al.* (2020) ‘The influence of physical attractiveness on attitude confidence and resistance to change’, *Journal of Experimental Social Psychology*, 90, pp. 1–12. doi: 10.1016/j.jesp.2020.104018.

Melone, N. P. *et al.* (1993) ‘The effect of decision support systems on managerial performance and decision confidence’, *Proceedings of the Twenty-sixth Hawaii International Conference on System Sciences*, 4, pp. 482–489. doi: 10.1109/HICSS.1993.284225.

Meloy, M. G. and Russo, J. E. (2004) ‘Binary choice under instructions to select versus reject’, *Organizational Behavior and Human Decision Processes*, 93(2), pp. 114–128. doi: 10.1016/j.obhdp.2003.12.002.

Mendel, R. *et al.* (2011) ‘Confirmation bias: Why psychiatrists stick to wrong preliminary diagnoses’, *Psychological Medicine*, 41(12), pp. 2651–2659. doi: http://dx.doi.org/10.1017/S0033291711000808.

Meyer, A. *et al.* (2013) ‘Physicians’ diagnostic accuracy, confidence, and resource requests: A vignette study’, *JAMA Internal Medicine*, 173(21), pp. 1952–1959. doi: 10.1001/jamainternmed.2013.10081.

Miller, D. J., Spengler, E. S. and Spengler, P. M. (2015) ‘A Meta-Analysis of Confidence and Judgment Accuracy in Clinical Decision Making’, *Journal of Counseling Psychology*, 62(4), pp. 553–567. doi: 10.1037/cou0000105.

Molenberghs, P. *et al.* (2016) ‘Neural correlates of metacognitive ability and of feeling confident: a large-scale fMRI study’, *Social Cognitive and Affective Neuroscience*, 11(12), pp. 1942–1951. doi: 10.1093/scan/nsw093.

Moore, D. A. and Healy, P. J. (2008) ‘The trouble with overconfidence.’, *Psychological Review*, 115(2), pp. 502–517. doi: 10.1037/0033-295X.115.2.502.

Moore, D. A. and Schatz, D. (2017) ‘The three faces of overconfidence’, *Social and Personality Psychology Compass*, 11(8), pp. 1–12. doi: 10.1111/spc3.12331.

Morgan, R. D. *et al.* (2000) ‘Do the DSM decision trees improve diagnostic ability’, *Journal of Clinical Psychology*, 56(1), pp. 73–88.

Morgan, T. J. H. *et al.* (2012) ‘The evolutionary basis of human social learning’, *Proceedings of the Royal Society B: Biological Sciences*, 279(1729), pp. 653–662. doi: 10.1098/rspb.2011.1172.

Mower, W. R. *et al.* (2019) 'Effect of Initial Bedside Ultrasonography on Emergency Department Skin and Soft Tissue Infection Management', *Annals of Emergency Medicine*, 74(3), pp. 372–380. doi: 10.1016/j.annemergmed.2019.02.002.

Mussweiler, T., Strack, F. and Pfeif, T. (2000) 'Overcoming the Inevitable Anchoring Effect: Considering the Opposite Compensates for Selective Accessibility', *Society for Personality and Social Psychology*, 26(9), pp. 1142–1150.

Nickerson, R. S. (1998) 'Confirmation Bias: A Ubiquitous Phenomenon in Many Guises', *Review of General Psychology*, 2(2), pp. 175–220. doi: 10.1037/1089-2680.2.2.175.

Nosic, A. and Weber, M. (2010) 'How risky do I invest? The role of risk attitudes, risk perceptions, and overconfidence', *Decision Analysis*, 7(3), pp. 282–301. doi: 10.1287/deca.1100.0178.

Nurek, M. and Kostopoulou, O. (2016) 'What you find depends on how you measure it: Reactivity of response scales measuring predecisional information distortion in medical diagnosis', *PLoS ONE*, 11(9), pp. 1–17. doi: 10.1371/journal.pone.0162562.

Nurek, M., Kostopoulou, O. and Hagmayer, Y. (2014) 'Predecisional information distortion in physicians' diagnostic judgments: Strengthening a leading hypothesis or weakening its competitor?', *Judgment and Decision Making*, 9(6), pp. 572–585.

Ossenkoppele, R. *et al.* (2013) 'Impact of molecular imaging on the diagnostic process in a memory clinic', *Alzheimer's & Dementia*, 9(4), pp. 414–421. doi: 10.1016/j.jalz.2012.07.003.

Pandharipande, P., Alabre, C. I., *et al.* (2016) 'Changes in physician decision making after CT: A prospective multicenter study in primary care settings', *Radiology*, 281(3), pp. 835–846. doi: 10.1148/radiol.2016152887.

Pandharipande, P., Reisner, A. T., *et al.* (2016) 'CT in the emergency department: A real-time study of changes in physician decision making', *Radiology*, 278(3), pp. 812–821. doi: 10.1148/radiol.2015150473.

De Paola, M., Gioia, F. and Scoppa, V. (2014) 'Overconfidence, omens and gender heterogeneity: Results from a field experiment', *Journal of Economic*

Psychology, 45, pp. 237–252. doi: 10.1016/j.joep.2014.10.005.

Pfister, M. *et al.* (1999) 'Judgment analysis in clinical nephrology', *American Journal of Kidney Diseases*, 34(3), pp. 569–575. doi: 10.1016/S0272-6386(99)70088-0.

Phillips, B., Prybutok, V. and Peak, D. (2014) 'Decision Confidence, Information Usefulness, and Information Seeking Intention in the Presence of Disconfirming Information', *Informing Science*, 17, pp. 1–24. doi: 10.28945/1932.

Phua, D. H. and Tan, N. C. (2013) 'Cognitive aspect of diagnostic errors', *Annals of the Academy of Medicine Singapore*, 42(1), pp. 33–41.

Rabin, M. and Schrag, J. L. (1999) 'First Impressions Matter: A Model of Confirmatory Bias', *The Quarterly Journal of Economics*, 114(1), pp. 37–82. doi: 10.1162/003355399555945.

Relling, A. E. and Giuliodori, M. J. (2015) 'Effect of peer instruction on the likelihood for choosing the correct response to a physiology question', *Advances in Physiology Education*, 39(3), pp. 167–171. doi: 10.1152/advan.00092.2014.

Roets, A. *et al.* (2014) 'Effects of dispositional need for closure and training on medical decision making', *Medical Decision Making*, 34(2), pp. 144–146. doi: 10.1177/0272989X13499449.

Roets, A. and Van Hiel, A. (2011) 'Item selection and validation of a brief, 15-item version of the Need for Closure Scale', *Personality and Individual Differences*, 50(1), pp. 90–94. doi: 10.1016/j.paid.2010.09.004.

Rogers, E. M., Singhal, A. and Quinlan, M. M. (1983) *Diffusion of innovations*. 3rd edn. New York, NY: The Free Press.

Rollwage, M. *et al.* (2020) 'Confidence drives a neural confirmation bias', *Nature Communications*, 11(2634), pp. 1–11. doi: https://dx.doi.org/10.1038/s41467-020-16278-6.

Rowe, G. and Wright, G. (1996) 'The impact of task characteristics on the performance of structured group forecasting techniques', *International Journal of Forecasting*, 12(1), pp. 73–89. doi: 10.1016/0169-2070(95)00658-3.

Rowe, G., Wright, G. and McColl, A. (2005) 'Judgment change during Delphi-like procedures: The role of majority influence, expertise, and confidence', *Technological Forecasting and Social Change*, 72(4), pp. 377–399. doi:

10.1016/j.techfore.2004.03.004.

Roy, D. *et al.* (2014) 'Metacognitive scaffolds improve self-judgments of accuracy in a medical intelligent tutoring system', *Instructional Science*, 42(2), pp. 159–181. doi: 10.1007/s11251-013-9275-4.

Russo, J. E. *et al.* (2008) 'The Goal of Consistency as a Cause of Information Distortion', *Journal of Experimental Psychology: General*, 137(3), pp. 456–470. doi: 10.1037/a0012786.

Russo, J. E., Medvec, V. H. and Meloy, M. G. (1996) 'The Distortion of Information during Decisions', *Organizational Behavior and Human Decision Processes*, 66(1), pp. 102–110. doi: 10.1006/obhd.1996.0041.

Saito, M. (1998) 'Fluctuations of answer and confidence rating in a general-knowledge problem task: Is confidence rating a result of direct memory-relevant output monitoring?', *Japanese Psychological Research*, 40(2), pp. 92–103. doi: 10.1111/1468-5884.00079.

Saito, M. (2003) 'Two modes of confidence rating: An effect of episodic information of participant's own past responses in a repeated-question paradigm', *Japanese Psychological Research*, 45(2), pp. 94–99. doi: 10.1111/1468-5884.t01-1-00034.

Schwartz, B. (2005) *The Paradox of Choice: Why More is Less*. New York, NY: HarperCollins.

Short, D., Frischer, M. and Bashford, J. (2003) 'The development and evaluation of a computerised decision support system for primary care based upon "patient profile decision analysis"', *Journal of Innovation in Health Informatics*, 11(4), pp. 195–202. doi: 10.14236/jhi.v11i4.567.

Shynkaruk, J. M. and Thompson, V. (2006) 'Confidence and accuracy in deductive reasoning', *Memory & Cognition*, 34(3), pp. 619–632. doi: 10.3758/BF03193584.

Sibbald, M. *et al.* (2021) 'Should electronic differential diagnosis support be used early or late in the diagnostic process ? A multicentre experimental study of Isabel', *BMJ Quality & Safety*, (0), pp. 1–8. doi: 10.1136/bmjqs-2021-013493.

Sieck, W. R. and Arkes, H. R. (2005) 'The recalcitrance of overconfidence and its contribution to decision aid neglect', *Journal of Behavioral Decision Making*, 18(1), pp. 29–53. doi: 10.1002/bdm.486.

Simon, D. *et al.* (2001) 'The Emergence of Coherence over the Course of Decision Making', *Journal of Experimental Psychology: Learning Memory and Cognition*, 27(5), pp. 1250–1260. doi: 10.1037/0278-7393.27.5.1250.

Simon, M. and Kim, J. (2017) 'Two sources of overconfidence: Incorporating disconfirming feedback in an entrepreneurial context', *Journal of Small Business Strategy*, 27(3), pp. 9–24. doi: 10.5465/AMBPP.2016.12598abstract.

Simpkin, A. L. and Schwartzstein, R. M. (2016) 'Tolerating Uncertainty — The Next Medical Revolution?', *New England Journal of Medicine*, 375(18), pp. 1713–1715. doi: 10.1056/NEJMp1606402.

Singh, H. *et al.* (2017) 'The global burden of diagnostic errors in primary care', *BMJ Quality & Safety*, 26(6), pp. 484–494. doi: 10.1136/bmjqs-2016-005401.

Sirriyeh, R. *et al.* (2012) 'Reviewing studies with diverse designs: The development and evaluation of a new tool', *Journal of Evaluation in Clinical Practice*, 18(4), pp. 746–752. doi: 10.1111/j.1365-2753.2011.01662.x.

Sloman, S. A. (1996) 'The empirical case for two systems of reasoning', *Psychological Bulletin*, 119(1), pp. 3–22. doi: 10.1037/0033-2909.119.1.3.

Smith, D. and Dumont, F. (1997) 'Eliminating overconfidence in psychodiagnosis: Strategies for training and practice', *Clinical Psychology: Science and Practice*, 4(4), pp. 335–345. doi: 10.1111/j.1468-2850.1997.tb00125.x.

Spaanjaars, N. L. *et al.* (2015) 'Experience and diagnostic anchors in referral letters', *European Journal of Psychological Assessment*, 31(4), pp. 280–286. doi: 10.1027/1015-5759/a000235.

Spallazzi, M. *et al.* (2019) 'The Incremental Diagnostic Value of [F-18]Florbetaben PET and the Pivotal Role of the Neuropsychological Assessment in Clinical Practice', *Journal of Alzheimer's Disease*, 67(4), pp. 1235–1244. doi: 10.3233/JAD-180646.

Stanovich, K. E. and West, R. F. (1997) 'Reasoning Independently of Prior Belief and Individual Differences in Actively Open-Minded Thinking', *Journal of Educational Psychology*, 89(2), pp. 342–357. doi: 10.1037/0022-0663.89.2.342.

Stone, D. N. (1994) 'Overconfidence in Initial Self-Efficacy Judgments: Effects on Decision Processes and Performance', *Organizational Behavior and Human*

Decision Processes, pp. 452–474. doi: 10.1006/obhd.1994.1069.

Stoner, J. (1961) *A comparison of individual and group decisions involving risk*, *Unpublished Master Thesis*. MIT.

Thomas, R. P. *et al.* (2008) 'Diagnostic hypothesis generation and human judgment.', *Psychological Review*. Thomas, Rick P.: Department of Psychology, University of Oklahoma, Norman, OK, US, 73019, rthomas@psychology.ou.edu Thomas, Rick P.: rthomas@psychology.ou.edu, Dougherty, Michael R.: mdougherty@psyc.umd.edu, 115(1), pp. 155–185. doi: 10.1037/0033-295X.115.1.155.

Thomas, R. P., Dougherty, M. R. and Buttaccio, D. R. (2014) 'Memory Constraints on Hypothesis Generation and Decision Making', *Current Directions in Psychological Science*, 23(4), pp. 264–270. doi: 10.1177/0963721414534853.

Thompson, V. (2009) 'Dual-process theories: A metacognitive perspective', in Evans, J. and Frankish, K. (eds) *In two minds: Dual processes and beyond*. Oxford, UK: Oxford University Press, pp. 171–195.

Thompson, V., Prowse Turner, J., *et al.* (2013) 'The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking', *Cognition*, 128(2), pp. 237–251. doi: 10.1016/j.cognition.2012.09.012.

Thompson, V., Ackerman, R., *et al.* (2013) 'The role of answer fluency and perceptual fluency in the monitoring and control of reasoning: Reply to Alter, Oppenheimer, and Epley (2013)', *Cognition*, 128(2), pp. 256–258.

Thompson, V., Evans, J. S. B. T. and Campbell, J. I. D. (2013) 'Matching bias on the selection task: It's fast and feels good', *Thinking and Reasoning*, 19(3–4), pp. 431–452. doi: 10.1080/13546783.2013.820220.

Thompson, V. and Johnson, S. C. (2014) 'Conflict, metacognition, and analytic thinking', *Thinking and Reasoning*, 20(2), pp. 215–244. doi: 10.1080/13546783.2013.869763.

Thompson, V., Prowse Turner, J. and Pennycook, G. (2011) 'Intuition, reason, and metacognition', *Cognitive Psychology*, 63(3), pp. 107–140. doi: 10.1016/j.cogpsych.2011.06.001.

Trouche, E., Sander, E. and Mercier, H. (2014) 'Arguments, more than confidence,

explain the good performance of reasoning groups', *Journal of Experimental Psychology: General*. United States, 143(5), pp. 1958–1971. doi: 10.1037/a0037099.

Tullis, J. G. and Goldstone, R. L. (2020) 'Why does peer instruction benefit student learning?', *Cognitive Research: Principles and Implications*, 5(15), pp. 1–12. doi: 10.1186/s41235-020-00218-5.

Vega, S. *et al.* (2020) 'Metacognition in moral decisions: judgment extremity and feeling of rightness in moral intuitions', *Thinking and Reasoning*, 27(1), pp. 124–141. doi: 10.1080/13546783.2020.1741448.

Voytovich, A. E., Rippey, R. M. and Suffredini, A. (1985) 'Premature conclusions in diagnostic reasoning', *Journal of Medical Education*, 60(4), pp. 302–307. doi: 10.1097/00001888-198504000-00004.

Wallach, M. A., Kogan, N. and Bem, D. J. (1962) 'Group influence on individual risk taking', *Journal of Abnormal and Social Psychology*, 1962(1), pp. 1–39.

Wallach, M. A., Kogan, N. and Bem, D. J. (1963) 'Diffusion of responsibility and level of risk taking in groups', *Journal of Abnormal and Social Psychology*, 1963(1), pp. 1–31.

Wallach, M. A., Kogan, N. and Burt, R. B. (1965) 'Can group members recognize the effects of group discussion upon risk taking?', *Journal of Experimental Social Psychology*, 1(4), pp. 379–395. doi: 10.1016/0022-1031(65)90016-8.

Wang, S. and Thompson, V. (2019) 'Fluency and feeling of rightness: The effect of anchoring and models', *Psychological Topics*, 28(1), pp. 37–72. doi: 10.31820/pt.28.1.3.

Wang, X. and Du, X. (2018) 'Why does advice discounting occur? The combined roles of confidence and trust', *Frontiers in Psychology*, 9(2381), pp. 1–8. doi: 10.3389/fpsyg.2018.02381.

Wason, P. C. (1960) 'On the failure to eliminate hypotheses in a conceptual task', *Quarterly Journal of Experimental Psychology*, 12(3), pp. 129–140. doi: 10.1080/17470216008416717.

Wason, P. C. (1968) 'Reasoning about a rule.', *The Quarterly journal of experimental psychology*, 20(3), pp. 273–281. doi: 10.1080/14640746808400161.

Wason, P. C. and Shapiro, D. (1971) ‘Natural and contrived experience in a reasoning problem’, *Quarterly Journal of Experimental Psychology*, 23(1), pp. 63–71. doi: 10.1080/00335557143000068.

Webster, D. M. and Kruglanski, A. W. (1994) ‘Individual differences in Need for Cognitive Closure’, *Journal of Personality and Social Psychology*, 67(6), pp. 1049–1062. doi: 10.1037/0022-3514.67.6.1049.

Webster, P. (2020) *Virtual health care in the era of COVID-19*. Available at: www.thelancet.com.

Weston, P. S. J. *et al.* (2016) ‘Diagnosing Dementia in the Clinical Setting: Can Amyloid PET Provide Additional Value Over Cerebrospinal Fluid?’, *Journal of Alzheimer’s Disease*, 54(4), pp. 1297–1302. doi: 10.3233/JAD-160302.

Wood, T. J. *et al.* (2017) ‘The influence of first impressions on subsequent ratings within an OSCE station’, *Advances in health sciences education : theory and practice*. Netherlands, 22(4), pp. 969–983. doi: http://dx.doi.org/10.1007/s10459-016-9736-z.

Woolley, A., Kostopoulou, O. and Delaney, B. C. (2016) ‘Can Medical Diagnosis Benefit from “‘Unconscious Thought’”?’, *Medical Decision Making*, 36(4), pp. 541–549. doi: 10.1177/0272989X15581352.

World Health Organization (2016) *Diagnostic Errors: Technical series on safer primary care*, *World Health Organization*. Geneva.

Yaniv, I. (2004) ‘Receiving other people’s advice: Influence and benefit’, *Organizational Behavior and Human Decision Processes*, 93(1), pp. 1–13. doi: 10.1016/j.obhdp.2003.08.002.

Yélamos, O. *et al.* (2020) ‘Improvement of diagnostic confidence and management of equivocal skin lesions by integration of reflectance confocal microscopy in daily practice: Prospective study in 2 referral skin cancer centers’, *Journal of the American Academy of Dermatology*. doi: 10.1016/j.jaad.2019.05.101.

Zwan, M. D. *et al.* (2017) ‘Diagnostic impact of [F-18] flutemetamol PET in early-onset dementia’, *Alzheimer’s Research & Therapy*, 9(2), pp. 1–8. doi: 10.1186/s13195-016-0228-4.

Acknowledgements

First and foremost, I am extremely grateful to my primary supervisor, Dr Olga Kostopoulou for her continuous support, both personal and academic, her guidance over the years, the countless meetings, and the amount of time and effort dedicated to helping me complete this PhD. I would also like to thank my secondary supervisor, Professor Brendan Delaney for his precious clinical input and advice. I am also deeply thankful to Dr Martine Nurek for her tireless support and guidance.

I would also like to thank my examiners, Professor Valerie Thompson and Dr George Mylonas, who took the time to examine this work and provide their valuable and insightful feedback.

Many thanks to Imperial College London and the National Institute for Health and Care Research (NIHR) Patient Safety Translational Research Centre (PSTRC) for funding this research. Thanks to the NIHR Clinical Research Network North West London for their valuable help in recruiting participants for the studies and to the physicians who took part in them.

I would also like to thank my family and friends for their continuous support. Special thanks to my wife Rania, for being always there, supporting me to achieve my goals and making me a better person. Rania, we have been through a lot but, in the end, God always makes a way.

Finally, my deep appreciation to Dr Angeliki-Maria (Angie) P. Palaiologou, Associate Professor at the University of Ioannina, who decisively contributed to my inspiration for my academic pursuits. Angie is one of the most intellectually charismatic academics I have ever met, a gifted mind, and one of the most influential persons in my life, without even being aware of it.

This work is therefore dedicated to Dr Angie Palaiologou.