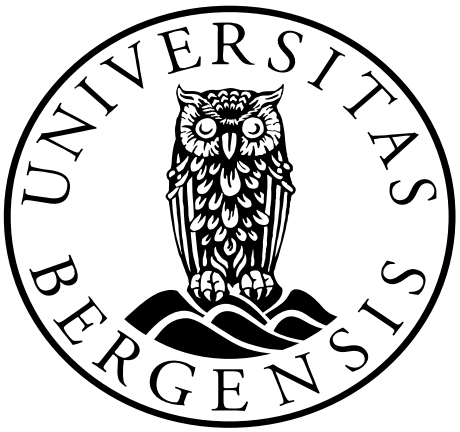
Biased Thinking in the Face of Uncertainty: Exploring the Process of Diagnostic Decision Making Within the Field of Psychiatry

Øystein Keilegavlen Bondevik og Thomas Færstad Frotvedt



MAPSYK360, Masterprogram i psykologi,

Studieretning: Psykologisk vitenskap

ved

UNIVERSITETET I BERGEN

DET PSYKOLOGISKE FAKULTET

VÅR 2018

Veileder: Bjørn Sætrevik, Institutt for samfunnspsykologi.

Abstract

A considerable amount of medical errors are indicated to occur during the diagnostic process. Various forms of faulty cognition on the part of the physicians have been identified as sources of diagnostic errors, including the susceptibility to several cognitive biases. It is therefore of interest to increase our knowledge of how biases may occur in diagnostic decision making. Our study aims to contribute to this knowledge by investigating the occurrence of anchoring bias, confirmation bias, and unwarranted confidence when making diagnostic decisions under uncertainty. For this purpose, we conducted two experiments on samples of Norwegian medical students (N = 128). The participants were presented with two hypothetical psychiatric cases. Anchoring bias was explored by manipulating presentation of the symptoms featured in the initial vignettes, investigating whether participants locked onto whichever symptoms were presented first when making a preliminary diagnosis. Confirmation bias and corresponding developments in diagnostic confidence were investigated through the participants' subsequent selections and interpretations of additional information. The results from each experiment did not indicate any occurrence of the three investigated phenomena. The non-findings are discussed in terms of the methodological aspects of the study, shedding light on challenges that may exist when investigating the different phenomena simultaneously. We also discuss the true prevalence and detectability of such biases. Based on our discussions, we make several proposals for how our design may be revised and expanded for future studies.

Keywords: confirmation bias, anchoring bias, confidence, diagnostics, psychiatry, medical decision making

Word count: 19136

Sammendrag

En betydelig andel feil innen medisin antydes å forekomme i den diagnostiske prosessen. Diverse former for feiltenkning blant leger har blitt identifisert som kilder til diagnostiske feil, deriblant sårbarhet for ulike kognitive bias. Det er derfor hensiktsmessig å øke vår forståelse for hvordan slike bias kan forekomme i diagnostisk beslutningstaking. Vår studie tar sikte på å bidra til denne forståelsen ved å undersøke forekomst av ankerbias, bekreftelsesbias og ubegrunnet sikkerhet i diagnostisering av usikre pasienttilfeller. Med dette formålet utførte vi to eksperimenter på utvalg av norske medisinstudenter (N = 128). Deltakerne ble presentert for to skrevne psykiatriske kasuistikker. Ankerbias ble testet ved å manipulere presentasjonen av symptomene i introduksjonen av pasienten, og å undersøke hvorvidt deltakere i sine valg av preliminære diagnoser tenderte til å vektlegge de symptomene de hadde blitt presentert for først. Bekreftelsesbias og korresponderende utviklinger i diagnostisk sikkerhet ble undersøkt i deltakernes påfølgende valg og tolkninger av ytterligere informasjon. Resultatene fra hvert eksperiment indikerte ingen forekomst av noen av de tre undersøkte fenomenene. Disse nullfunnene diskuteres i lys av de metodologiske aspektene ved studien, der det blir redegjort for diverse utfordringer som kan eksistere når de ulike fenomenene undersøkes samtidig. Vi diskuterer også den reelle forekomsten, såvel som identifiserbarheten, til fenomenene. Basert på diskusjonene kommer vi med ulike forslag til hvordan vårt design kan revideres og utvides for fremtidige studier.

Acknowledgements

First and foremost, we wish to thank our supervisor, first amanuensis Bjørn Sætrevik, for all the time and effort he has spent helping and guiding us in countless different ways throughout the process, including, but not limited to: Advising in the development of the topical focus and methodological features of our study, inviting us to partake in a workshop on medical decision making, facilitating our cooperation with the university hospital, monitoring parts of our data collection, developing a spreadsheet that considerably facilitated our analyses of the data, meeting us for discussions, reading and providing feedback to the various drafts of our thesis, and replying to an abundance of emails. His insight, enthusiasm and patience have been important ingredients in the completion of our work.

Additionally, we sincerely wish to thank professor Anders Bærheim for meeting us and for showing such interest in our work. We wish to thank him for allowing us to carry out the two experiments on different classes of medical students, and greatly appreciate the way he introduced us and our project to the classes, which aided our recruitment of participants. Furthermore, we wish to thank the two classes who participated for welcoming us and spending their breaks partaking in our experiments.

Finally, we want to thank all others who have helped and supported us during our work in various ways.

Table of Contents

A physician's workday is characterized by numerous decisions that must be made regarding one's patients, many of which concern plausible diagnoses. Among the components of medical practice, the skills and consequences associated with diagnostics may in many cases be the most critical (Croskerry, 2009a, 2009b). As diagnostic decisions are often directly connected to choices of treatment, getting the diagnosis right may often be pivotal to the patient's life, health and well-being (see Klein, 2005; Mendel et al., 2011; Parmley, 2006). Alas, due to the element of uncertainty that is always present to some degree in any medical case, erroneous diagnostic decisions are made fairly frequently (see Croskerry, 2009b). Various empirical contributions suggest that doctors get their diagnoses wrong 10-15% of the time (see Berner & Graber, 2008; Graber, 2013). Overall, diagnostic errors are indicated to represent the second most common class of errors in medicine, only surpassed by treatment errors (Graber, Gordon & Franklin, 2002; van den Berge, 2012).

While diagnostic errors may vary in terms of their nature and causes, a substantial amount of them is suggested to be rooted in the thinking processes of the physicians (Croskerry, 2009b; Graber, Franklin & Gordon, 2005; Graber et al., 2002). Cognitive psychology has therefore become a relevant field for understanding diagnostic errors, and may ultimately aid in discovering ways to reduce their occurrence (see Croskerry, 2009a, 2009b). In recent decades, the identification of cognitive biases that may hinder accurate diagnostic reasoning has been the focus of many empirical and theoretical contributions (Blumenthal-Barby & Krieger, 2015; Croskerry, 2003, 2009a, 2009b; Graber et al., 2005; Saposnik, Redelmeier, Ruff & Tobler, 2016). Understanding the situations in which various biases manifest themselves may aid in developing techniques to prevent them from adversely affecting diagnostic decisions (Blumenthal-Barby & Krieger, 2015; Crowley et al., 2013; Graber et al., 2002; Mendel et al., 2011; Parmley, 2006; Saposnik et al., 2016). Our study aims to contribute to such an understanding by exploring circumstances under which certain cognitive biases may plausibly occur in diagnostic reasoning. Specifically, we wish to investigate *anchoring/primacy bias*, *confirmation bias* and issues regarding diagnostic *overconfidence* among Norwegian medical students who make diagnostic assessments about hypothetical psychiatric cases.

Diagnostic Decision Making: Uncertainty and Error

Diagnostics involves the core operations of decision making as typically defined in cognitive psychology: It entails obtaining and assessing various information, before choosing among options (i.e. diagnoses) that are available in a particular situation (Matlin, 2013; see

also Croskerry, 2009a, 2009b; Klein, 2005). In recent decades, contributions from cognitive psychology have investigated how physicians arrive at diagnoses, examining the complex processes that may lead to both correct and incorrect diagnostic decisions (Croskerry, 2009a; Klein, 2005). According to the influential *hypothetico-deductive method*, diagnostic reasoning involves the generation of one or more diagnostic hypotheses, followed by searches for additional information to confirm or refute these (Norman, Young & Brooks, 2007). Factors from several interrelated sources contribute to the difficulty of diagnostic decision making. One such source is the information that is available and relevant in a given diagnostic situation, and how it is exchanged between the patient and the physician (see for example Croskerry 2009b; Payne, 2011). This information is, in turn, influenced by other sources, such as characteristics of the patients and the physicians, as well as the relationship between the two (see Croskerry, 2009a, 2009b; Parmley, 2006). Each of these elements may themselves be influenced by numerous situational factors, further adding to the complexity of a consultation process (Croskerry, 2009a, 2009b; Graber et al., 2002). Overall, the interplay of factors and circumstances that may be present in a given situation contribute to the inherent element of uncertainty that characterizes diagnostic decision making (see Croskerry, 2009a, 2009b). Whether prominent or obscure in a particular case, there is always some risk of arriving at incorrect diagnostic conclusions (see Graber et al., 2002). Even though the presence of diagnostic uncertainty and errors is indicated to be highest within the fields of internal medicine, family medicine and emergency medicine, errors may occur in any specialty (see Croskerry, 2003). While systematic investigations of different types and causes of diagnostic errors have long remained scarce (see Graber, 2013; Norman, 2009), an empirically supported taxonomy has received considerable attention in recent years. Developed by Graber and colleges (2005; see also Graber et al., 2002), this taxonomy and its underlying research indicate that many diagnostic errors are primarily cognitive in nature.

Graber and colleagues (2005) defined a diagnostic error as "a diagnosis that was unintentionally delayed (sufficient information was available earlier), wrong (another diagnosis was made before the correct one), or missed (no diagnosis was ever made), as judged from the eventual appreciation of more definitive information" (p. 1493). Their taxonomy distinguishes between three general classes of diagnostic errors (see also Graber et al., 2002). These classes can variously involve the aforementioned elements that influence and characterize a diagnostic situation. *No-fault* errors are errors that the physicians have very little or no chance in detecting and preventing, such as when the illness is very rare, silent, masked, or presented in an unusual fashion. *System errors* refer to latent flaws in the health

care system itself, spanning factors like policies, coordination of care, training and supervision, communication, distractions and workload. *Cognitive errors* encompass various types of errors in the thinking processes of the individual physicians. These may come in many different forms, such as faulty data gathering, inadequate knowledge and clinical reasoning, as well as faulty verification. Any number of the these classes may occur in a given diagnostic situation, and different erroneous elements may often be intertwined. In general, however, cognitive errors are indicated to occur the most frequently of the three (Croskerry, 2009a; Graber & Carlson, 2011; Graber et al., 2005), justifying further investigations of their manifestations as well as their underlying mechanisms.

Cognitive Errors in Diagnostics: Heuristics, Biases and Dual Process Theory

Within the aforementioned taxonomy, falling prey to heuristic biases is specified as one important form of cognitive errors. Such errors were detected in a substantial amount of the 100 clinical cases analyzed by Graber and colleagues (2005) when developing the taxonomy. This category of diagnostic errors had already received attention in various empirical works conducted prior to the development of the taxonomy. First popularized by Tversky and Kahneman in the 1970's as relevant to the understanding of judgments and decision making at large (Norman, 2009; Tversky & Kahneman, 1974), the constructs of heuristics and biases were soon applied to medical reasoning (see Crowley et al., 2013; Payne, 2011). In essence, heuristics refer to mental rules of thumb, or "shortcuts" that may be deployed in situations of uncertainty (Crowley et al., 2013). They simplify the operations required for making judgments, and often lead to adequate solutions. In general, they allow for decisions and solutions to be made efficiently and relatively effortlessly, and are therefore adaptive features of human cognition (Crowley et al., 2013; Norman 2009). Arguing for their usefulness in dynamic medical settings, Graber and colleagues (2002) stated that heuristics allow clinicians to navigate diagnostic challenges, specifying that heuristic solutions "free up cognitive resources so that they can be applied toward other demands" (p. 985). The price of using heuristics are their inherent susceptibility to systematic biases (see Crowley et al., 2013; Tversky & Kahneman, 1974). In their 1974 article, Tversky and Kahneman specified that heuristics may sometimes lead to misinterpretation and oversimplification of situations. In the case of diagnostics, erroneous decisions may occur through the failure to apply appropriate heuristic principles, or through overapplication of heuristics under inappropriate circumstances (see Graber et al., 2005). While cognitive biases have been investigated within diagnostics for many decades, a theoretical framework of the mechanisms involved in

diagnostic reasoning has long been lacking (see Croskerry 2009a, 2009b). However, a popular theory from cognitive psychology has recently been applied for this end, proposing how and when heuristics and biases may contribute to erroneous diagnostic decisions.

Developed alongside the "heuristics and biases" tradition, the *dual process theory*, has received considerable attention in recent decades (Croskerry, 2009a, 2009b; Kahneman, 2003; Saposnik et al., 2016). Within a medical perspective, some authors argue for its explanatory power with regard to the processes underlying diagnostic reasoning, including how inaccurate diagnostic hypotheses and decisions may be generated (see Croskerry, 2009a, 2009b; see also Payne, 2011; van den Berge, 2012). This framework views cognitive operations in terms of two general classes, commonly labeled *System 1* and *System 2* (Norman, 2009). Other labels are *non-analytical* or *intuitive* mode for System 1, and *analytical* mode for System 2 (Norman & Eva, 2010; Payne, 2011). The two classes are differentiated in terms of their general characteristics, summarized by Croskerry (2009a, 2009b): System 1 processing is fast, automatic, unconscious, highly context-bound and often governed by habit, (see also Kahneman, 2003; Payne, 2011). It is characterized by heuristics and other mental shortcuts, frequently producing adequate solutions without requiring much effort. Its automatic nature allows it to perform multiple operations in a parallel fashion, compatible with Graber and colleagues' (2002) description of heuristics as "freeing up cognitive resources" (p. 985). However, despite its efficient and often adequate performance, it occasionally fails, due to the susceptibility to biases associated with heuristic approaches. By contrast, System 2-processing is slow, analytical and effortful. It is more resource intensive, putting more strains on the already limited capacity of working memory. This means that the system operates in a serial fashion, rather than being able to perform multiple operations in parallel (see also Payne, 2011). However, the system is relatively flexible compared to System 1: Rather than being habitual and concrete, it operates on abstract concepts and rules that can be applied in many settings, including those in which we have little or no prior experience (see Norman, 2009; Payne, 2011). Its thorough, analytical nature also makes it more reliable in terms of producing correct solutions. Additionally, System 2 is able to monitor the operations of System 1, and to override system 1 responses when deemed necessary.

According to Croskerry (2009a, 2009b), various circumstances may lead either system to dominate the reasoning process and the resulting judgments and decisions. System 1 will typically engage automatically when the clinician recognizes a familiar pattern of symptoms and features in the patients, resulting in a quick, relatively effortless diagnostic decision. Croskerry (2009b) pointed out that many diagnostic decisions are based on such types of

automatic pattern recognition, and these may often work very well. If the patients' symptoms are not readily recognized as belonging to such a pattern, a more thorough, analytical System 2 response may occur. This system will attempt to make sense of the stimuli through objective and systematic examination of the information, and by applying accepted rules of reasoning and logic (Croskerry, 2009b). While either system may be appropriate under certain circumstances, System 1 responses may often take precedence (Croskerry, 2009b). This can happen for a number of reasons related to various factors surrounding the diagnostic situation, including the physician and the patient in question. One such factor concerns the training and experience of the physician: As specified by Croskerry (2009a; 2009b), repetitive System 2 processing of particular stimuli may eventually lead to a System 1 response when facing these stimuli later on. That is, through repeated exposure to situations that were initially unfamiliar (i.e. training), a physician may eventually be able to quickly and effortlessly identify salient features (symptoms) through pattern recognition. This may be regarded as the development of a heuristic judgment. It must be noted, however, that some authors argue that non-analytical processes of recognizing patterns and relying on prior experience may be used at any level of expertise (see Norman, 2009; Norman et al., 2007). According to their view, non-analytical processes are a component of diagnostic reasoning across all levels of experience, entailing processes of matching encountered stimuli to previously encountered examples or exemplars, similar to the process of pattern recognition described above. A key difference resulting from accumulated experience is the quality and quantity of the exemplars available for comparison. This view is supported by empirical findings within dermatology, electrocardiography and psychiatry (see Norman et al., 2007). In other words, essential features of diagnostic reasoning may be more similar across levels of experience than proposed by a typical dual process perspective.

Another suggested cause of System 1 dominance in diagnostic reasoning is *dysrationalia override* (see Croskerry, 2009a, 2009b): In some situations, clinicians may more or less deliberately choose to override strict principles of logic and follow intuitive feelings. According to Croskerry (2009b), this is not uncommon in medicine, possibly reflecting a failure to accept or incorporate sound and logic decision rules. Even thorough clinical decision guidelines may be neglected in some situations in which a physician acts on the irrational belief that he or she knows better. Croskerry (2009b) states that such departures from rationality may occur for historical, habitual, emotional and situational reasons, among others. However, the model also specifies how System 2 may sometimes override System 1 responses, labeled *rational override*. The aforementioned monitoring capacity of System 2

allows it to detect and reject System 1 responses when deemed appropriate, which may prevent flawed irrational responses like inappropriate use of heuristics from occurring. However, this controlling function is not infallible: As Croskerry (2009b) pointed out, the monitoring capacity of System 2 works best when the decision maker is well rested, well slept, free from distraction, and focused on the task at hand. However, these ideal conditions are often not met in real life (Croskerry, 2009b): Physicians may be hurried, distracted, tired, and limited by resource constraints. Workload may be high, dynamic and unpredictable, providing significant challenges to their processing abilities. Any resulting inattentiveness, fatigue and cognitive indolence may impair the monitoring process, allowing System 1 responses to dominate, despite the presence of uncertainty and potentially grave consequences associated with certain clinical decisions.

In sum, the framework suggests that System 1 responses may, due to various reasons, dominate many clinical decisions, even when thorough, analytical processing would be expedient. Such automatic responses may in turn represent a substantial, albeit complex, source of error through their inherent susceptibility to biases. Croskerry (2009b) did warn about the limited explanatory power of the framework, stating that not all reasoning and decision making will neatly fall into one of the two classes of processing. Other authors have also contested how the dual process framework is applied in terms of explaining diagnostic errors (see Norman, 2009; Norman & Eva, 2010). However, it remains a popular explanatory framework of diagnostic reasoning, and has been incorporated in numerous recent contributions within the field in recent years (see Croskerry, 2009a; 2009b; Payne, 2011; Saposnik, 2016; van den Berge, 2012). Furthermore, while the application of the framework in explaining the details of diagnostic reasoning is fairly new, making a debate of its applicability highly useful, the identification of cognitive biases in diagnostics has been going on for decades. As empirical contributions have produced numerous findings regarding cognitive biases in medicine, this remains an expedient field of inquiry with regard to understanding diagnostic errors. Among the biases detected in various studies are anchoring bias, confirmation bias and issues related to diagnostic overconfidence.

Cognitive Biases in Clinical Reasoning: Conceptualizations

Definitions of anchoring in previous literature vary slightly. An early conceptualization that has often been cited and used in empirical works is that by Tversky and Kahneman (1974). The authors state that for estimation tasks featuring elements of uncertainty, people often start “from an initial value that is adjusted to yield the final answer”

(p. 1128). The initial value thus acts as an "anchor" that affects subsequent estimates. Importantly, the authors claimed that the later adjustments of the estimates are often insufficient. That is, "different starting points will yield different estimates, which are biased towards the initial values" (p. 1128). The starting points may be suggested to the subject through the presentation of the problem itself (Tversky & Kahneman, 1974), such as symptoms initially presented in a clinical case. Coming from a clinical perspective, Crowley and colleagues (2013) conceptualized anchoring in a way that is compatible with that of Tversky and Kahneman (1974): In their view, anchoring refers to the process of locking on to salient evidence early in the diagnostic process, leading to an initial diagnosis. In their study, the authors made a distinction between an *anchoring heuristic* and an *anchoring bias*. The former referred to making an initial estimate, which, in an of itself, may be more or less correct. The latter entailed that the initial estimate is incorrect, and that the subsequent adjustments are insufficient or otherwise faulty (see also Croskerry, 2003; Pines, 2006). According to Parmley (2006), anchoring bias, or *primacy* bias, occurs "when people are exposed to identical information, but in varying order, resulting in significantly different judgments" (p. 47). Compatible with the conceptualizations recited previously, this definition captures the essence of how anchoring has been operationalized in several empirical contributions within diagnostics (see *Anchoring and primacy bias* for an overview).

Confirmation bias typically refers to seeking or interpreting evidence in ways that are partial to existing beliefs, expectations, or a hypothesis at hand (Nickerson, 1998). When applying this concept in clinical settings, authors have variously emphasized its manifestations, either in the process of information searching, or in the process of interpreting information at hand. Croskerry (2003) defined confirmation bias as the tendency to look for confirming evidence in order to support a diagnosis rather than looking for disconfirming evidence in order to refute it, despite the latter often being more persuasive and definitive. Klein (2005) stated that confirmation bias may manifest itself in the process of selecting tests and questions to use in a diagnostic situation. Focusing on the interpretation of information, Payne (2011) stated that confirmation bias may entail overly emphasizing clinical information that appears to support one's preliminary diagnosis, while not giving weight to information that goes against one's hypothesis and/or supports alternative diagnoses. Encompassing all the aforementioned features, Parmley (2006) defined confirmation bias as the process of knowingly or unknowingly searching and highlighting information that is consistent with an initial hypothesis or judgment, and ignoring or deemphasizing inconsistent information. Confirmation bias is associated with *premature closure*, the tendency to accept a particular

diagnosis before it has been fully verified, neglecting plausible alternatives in the process (see Croskerry, 2002, 2009a; Eva, 2001). According to several authors, anchoring and confirmation bias may often be closely related to each other (Croskerry, 2002; Pines 2006; see also Cunnington, Turnbull, Regehr, Marriott & Norman, 1997): A physician may "lock onto" salient symptoms and features of a patient early in an encounter or a presentation, whereupon a preliminary diagnosis will be generated. Subsequent processes of searching for and interpreting additional information may be biased towards this hypothesis. According to Croskerry (2002, 2003), confirmation bias may compound errors that result from anchoring bias. However, anchoring is not necessarily followed by any form of confirmation bias, and confirmation bias can occur without being preceded by anchoring. As will be shown, the two have often been investigated separately in empirical work.

In general, physicians's confidence in their diagnostic hypotheses may both influence and be influenced by their reasoning. As a physician considers information and chooses among diagnostic options, the degree of confidence in his or her choices may vary, depending on the perceived qualitative and quantitative aspects of available information, as well as the existence of plausible diagnostic alternatives (see Eva, 2001; Martin, 2001; Oskamp, 1965). Croskerry (2003) defined *overconfidence bias* as a universal tendency to believe that we know more than we do, and specifies that such biases may be augmented by anchoring. Indeed, "locking onto" salient information presented early may entail considerable confidence that this initial information is particularly important, which may, in turn, greatly affect the formation and rigidness of a diagnostic hypothesis. Berner and Graber (2008) described diagnostic overconfidence both in terms of more general attitudes, as well as situation-specific cognitions, depicted as failures to realize the limitations in one's state of knowledge. Arguing for overconfidence as an important source of diagnostic error, the authors link the construct with erroneous cognitive activities, such as faulty use of heuristics and biased confirmatory tendencies. As will be addressed, several empirical contributions within psychiatric diagnostics have indeed suggested a link between levels of confidence and confirmation bias. If clinicians strongly believe to have the right diagnostic hypothesis, they might view available information as confirmatory of this hypothesis, and to seek out additional supporting evidence, rather than pursuing possible alternatives (see Martin, 2001; Oskamp, 1965). On the contrary, it is also plausible that lower levels of confidence may be associated with less susceptibility to biased thinking (Martin, 2001). Croskerry (2003) specified that overconfidence entails tendencies to act on incomplete information, intuitions and hunches when making diagnostic decisions. In turn, such tendencies may produce erroneous decisions,

and thus put the patient's health and well-being at risk. Overall, these elements imply that levels of confidence may be an important element in diagnostics, possibly acting as both causes and effects in relation to the cognitive processes.

Cognitive Biases in Diagnostics Decision Making: Empirical Findings

Previous literature has demonstrated that the prevalence of diagnostic errors is considerable (Graber, 2013; Graber, et al., 2002; van den Berge, 2012), and that faulty cognitive processes, such as biased judgments and decisions, may represent a relevant source for such errors (see Graber et al., 2005; Graber et al., 2002; Croskerry, 2009a; 2009b; Saposnik et al., 2016). A few recent review articles summarize empirical findings regarding cognitive biases in medical reasoning (see Blumenthal-Barby & Krieger, 2015; Saposnik et al., 2016). While these reviews cite evidence of numerous cognitive biases, they also indicate that there are serious limitations to our knowledge with regard to the true prevalence of various biases in medical decision making. As a whole, they demonstrate the importance of increasing our understanding of biased thinking in medicine, including within the field of diagnostics. They also demonstrate that anchoring bias, confirmation bias and issues related to diagnostic confidence are relevant points of inquiry in this regard. In the following, findings related to each of these constructs within diagnostics are summarized.

Anchoring and primacy bias. As noted above, Parmley (2006) viewed the terms *anchoring* and *primacy* as synonymous with one another when referring to the phenomenon whereby the order of the information presented to an individual affects the decision he or she subsequently makes. We will henceforth treat them as overlapping, as they largely reflect the same phenomena, despite being labeled differently in various empirical contributions. Generally, anchoring bias has been investigated by presenting participants with hypothetical written cases featuring clinical information supportive of several different diagnoses. The order in which this information is presented has been varied, and participants have been expected to choose a diagnosis in line with whichever information is presented first. Such a pattern would indicate that the participant has locked on to this information and not given similar emphasis to later, equally relevant information. While some studies using variations of this design have demonstrated anchoring bias, findings have not been entirely consistent.

Following the aforementioned approach, Friedlander and Stockman (1983) investigated anchoring bias in the psychiatric domain, using cases featuring suicidality and anorexia nevrosa. In this study, they only found an anchoring effect in the former case. When attempting to replicate this study, Ellis, Robbins, Schult, Ladany & Banker (1990) did not

find any effect for either case. Following these results, they speculated that Friedlander and Stockman's (1983) results were subject to Type 1 error. Richards and Wierzbicki (1990) performed a study that consisted of 4 psychiatric cases featuring alcohol abuse, anxiety, depression and antisocial behaviour. The authors pointed out that earlier studies on anchoring had used small samples, that standardized methods were lacking, that the clinical cases used sometimes differed in their severity, and that the participants were given too much information to process at once. When attempting to correct most of these issues in their own experiment, Richards and Wierzbicki (1990) found an anchoring effect for the cases involving alcohol abuse, anxiety and antisocial behaviour, but only a modest effect for depression. Following a similar approach, Cunnington and colleagues (1997) made more consistent findings: Using 10 hypothetical clinical cases from internal medicine, the authors found clear tendencies for participants across the various conditions to favor the diagnosis congruent with the information presented first. Taking all these results together, findings regarding anchoring or primacy bias in medicine through manipulations of the symptom presentation have been somewhat mixed: Some studies have found a prevalence of such biases, whereas others have not. This provides incentives for further investigation.

Confirmation bias and confidence. Confirmation bias has been documented within several medical domains, including psychiatry. In an experiment, Mendel and colleagues (2011) sought to explore the occurrence of confirmation bias among psychiatrists and medical students, as well as investigating whether confirmation bias leads to poorer diagnostic accuracy. The experiment began with case vignette describing a patient with symptoms compatible with both Alzheimer's disease and a severe depressive episode. The participants were then instructed to select one of these as a preliminary diagnosis. The vignette was presented in such a way that a majority of the participants would initially choose the incorrect option (depression). They could then access various pieces of follow-up information about the patient by selecting from a list of options. The options were presented as short summaries of the information that could be accessed, and participants could select as many as they wanted. Half of the options in the list were worded so that they appeared to indicate support of Alzheimer's disease as the correct diagnosis, while the other half appeared to indicate depression. This allowed for various degrees of confirmatory, neutral and disconfirmatory searching strategies to be deployed by the participants. Upon selection, the participants received the complete texts containing the requested information. Crucially, these texts would, as a whole, support Alzheimer's disease, regardless of what diagnosis had been indicated by the corresponding options in the list. This should lead participants who had

chosen the incorrect option for their preliminary diagnosis to realize the need to change it based on the additional information, unless they showed confirmation bias in their search for, and interpretation of, this information. At the end of the case, the participants were to make their choice of a final diagnosis, choosing one of the two aforementioned options. The experiment found confirmation bias for 13% of the psychiatrists and for 25% of medical students. Thus neither group of participants were immune to confirmation bias, but one even less so than the other. For the effects on diagnostic accuracy, the experiment indicated that participants who deployed a confirmatory information search more often chose the wrong diagnostic option compared to those who showed a more balanced or disconfirmatory information search.

Parmley (2006) investigated the occurrence of confirmation bias in clinician's psychodiagnostic assessments, using an online experimental design. Participants in the experiment received two case vignettes, and were asked to state a diagnostic hypothesis for each case. Each vignette were supportive of a particular diagnosis. A week later, the participants were presented with further information about each case, which was either consistent or inconsistent with the previously indicated diagnosis. Following this, the participants were asked once again to state a diagnostic hypothesis. Failure to adjust their hypotheses when receiving inconsistent information would indicate confirmation bias. Overall, this bias was found in 33% of the responses. Years in practice played no significant role in terms of exhibiting confirmation bias.

Oskamp (1965) made an early contribution to the relationship between confirmatory tendencies and the level of confidence in one's clinical decisions. He sought to investigate developments in accuracy and confidence in one's diagnostic hypotheses, and the relationship between the two, among clinical psychologists and students of psychology and personality. Through four stages, participants were presented with written information about a psychological case, each stage revealing more details about the patient. At each stage, after reading the information, participants were to answer numerous questions concerning psychology and personality judgments, and to rate their level of confidence in these judgments on a scale, before moving on to the next stage. As the participants progressed through the stages, the accuracy in their judgments seized to increase relatively quickly. By contrast, the levels of confidence in their judgments continued to rise throughout the entire procedure, to the point where nearly all participants became more confident than their performances warranted. In other words, confidence often became overconfidence in the latter stages of the experiment. Regarding the final stage of reading, Oskamp remarked that

participants in the final stage primarily seemed to attempt to confirm their existing assessments. Parmley (2006) noted that such overconfidence may lead to an excessive focus on information that appears to confirm one's hypotheses, and to potential neglect or misinterpretation contradictory information.

Since Oskamp's (1965) publication, few contributions have explicitly elaborated on the relationship between confirmatory tendencies and stated confidence in a clinical context (Parmley, 2006). One exception is Martin (2001), who investigated this relationship in a sample of students, with or without clinical training, using a simulated 15-minute consultation with a mock therapy client. At three instances throughout the session, participants were to state a diagnostic hypothesis, rate their confidence in this hypothesis, and list questions they subsequently wanted to ask the patient. Through subsequent coding of the various questions that were submitted by the participants, distinctions were made between confirmatory, neutral and disconfirmatory questions. Generally, the participants tended to start off with a neutral approach. However, over time, they tended to become more confirmatory and less neutral in their questions. Additionally, participants who were more disconfirmatory in their approach reported lower levels of confidence in their diagnoses compared to those who were more confirmatory. In general, these results are compatible with those found by Oskamp (1965). Taking all these studies into account, there is evidence that confirmation bias may occur in psychiatric diagnostics, and that being confident in one's diagnostic hypothesis may be related to such confirmatory tendencies when making diagnostic assessments.

Aim of the Current Study

Theoretical and empirical literature suggests how investigating cognitive biases can make up a primary component in the work to reduce diagnostic errors. As shown above, many diagnostic errors appear to derive from erroneous cognitive processes, which encompass susceptibility to various biases. Increasing our knowledge and understanding of how and when such biases occur help set the stage for investigating how they can be reduced or even eliminated (see for example Croskerry, 2003). Above, we cited evidence that anchoring bias and confirmation bias may be relevant points of inquiry for such a purpose within psychiatric diagnostics. We also recited findings indicating that confirmatory tendencies may be related to the physician's level of confidence in his or her diagnostic hypotheses. The various studies described above differ in their thematic and methodological focus when studying cognitive biases, inspiring us to combine investigative features in a way that, to our knowledge, has not previously been attempted within psychiatric diagnostics.

Using written cases, each featuring two plausible diagnoses, we wished to study anchoring bias as defined by Parmley (2006), using a similar approach to that of earlier studies: Through manipulating the presentation of a certain set of initial symptoms, primarily the order in which they appear to subjects, we wished to investigate whether such variations would be sufficient to affect the preliminary diagnostic hypotheses of the subjects. However, our study would combine such an inquiry with multiple investigations related to confirmation bias, which would make up the core of our study. Although anchoring and confirmation bias may co-occur in diagnostics (see Croskerry, 2002, 2003; Pines, 2006), a simultaneous investigation of the two appears to be lacking in psychiatry. Based on Parmley's (2006) conceptualization of confirmation bias, and primarily inspired by the experimental design used by Mendel and colleagues (2011), our investigation would span both searches for and interpretations of information. We would allow participants to pursue additional information regarding each diagnostic option by selecting from a list of relevant follow-up questions, each of which would indicate support of a particular diagnosis. Measurements of diagnostic confidence preceding these pursuits would also be taken, to investigate relations between confidence and style of information gathering. Biased confirmatory tendencies in the interpretation of information would be investigated by examining developments in confidence among participants who consistently pursued a particular diagnosis, despite receiving no information that objectively provided any conclusive support for it. The combination of phenomena investigated meant that there would be no "correct" diagnoses, but that varying degrees of ambiguity had to be present throughout the cases.

The theoretical and empirical literature presented indicates that medical students, like physicians, may be susceptible to biased thinking in diagnostics. We therefore wished to investigate the aforementioned phenomena in samples of advanced medical students from a Norwegian university hospital. To our knowledge, this population has not been previously explored in such a study. In order to investigate the combination of biases and confidence in the diagnostic process, we designed a series of two experiments, each featuring two hypothetical cases developed specifically for this end. The two experiments were conducted in a classroom setting, and were largely similar in terms of structure and content. Both were pre-registered at the Open Science Framework (OSF). Below we state our general research questions and expectations.

Research Questions and Hypotheses

1: Will the order in which the symptoms are presented in a case vignette affect the choice of a preliminary diagnosis? Our corresponding hypothesis (H1) was as follows: *Participants will be more likely to select the preliminary diagnosis congruent with the symptoms presented first in a vignette, rather than selecting the diagnosis congruent with symptoms presented later*. Such a pattern would indicate anchoring bias as previously defined. For Experiment 2, our investigation of symptom presentation had a slightly broader focus, not only encompassing variations in the order of initial symptoms, but also variations in the number of symptoms appearing to favor different diagnoses. This was done as a means to induce stronger diagnostic hypotheses in our participants, furthering our emphasis on investigating confirmatory tendencies.

2: Will participants primarily seek out information that appears to confirm their existing diagnostic hypothesis, rather than seeking out information that appears to favor an alternate diagnosis? Our corresponding hypothesis (H2) was as follows: *In their requests of additional information, participants will more often select items that appear to support their preceding preliminary diagnosis, than they will select those that appear to favor an alternate diagnosis.*

3: Will higher confidence in the diagnosis correspond to more "confirmatory" information gathering? Our corresponding hypothesis (H3) was as follows: *When participants are given the opportunity to request additional information about a clinical case, requests for "confirmatory" information will be preceded by higher levels of confidence in the previously selected diagnosis, than requests for "dissenting" information.*

4: Will participants who maintain the same diagnostic hypothesis throughout the case, and only pursue "confirmatory" information, become more confident in their diagnosis over time, when the actual information received is inconclusive? Our corresponding hypothesis (H4) was as follows. *In instances in which participants have stated the same diagnosis throughout the case evaluation, and have also exclusively requested "confirmatory" information throughout the case, the participants will report an increase in confidence in their diagnostic choice at the end of the case.* Such a pattern would indicate biased confirmatory tendencies when interpreting information that in reality neither confirms nor refutes any of the diagnostic options.

The methods and confirmatory analyses from each experiment followed the specifications made in the corresponding pre-registrations at OSF. As Experiment 2 was developed following the completion and analysis of Experiment 1, the methods, results and

specific topics for discussion regarding each experiment will be presented sequentially. Data files, flow charts of the case structure and translated case materials from each experiment is available in a Google project folder which can be accessed through the OSF homesite for the project (*Online Experiment on Anchor and Confirmation Bias in Setting Diagnoses*; https://osf.io/dn4rv/).

Methods and Results

Methods - Experiment 1

Experiment 1 was pre-registered as *Anchoring and Confirmation Bias in Diagnostics - an Online Experiment* (https://osf.io/rmgdy/register/565fb3678c5e4a66b5582f67).

Participants. 71 advanced medical students from a university hospital in Norway participated in the experiment. The students were approaching the completion of their medical degree, and had thus undergone extensive education regarding the diagnostic processes within the various fields of medicine, including psychiatry. We therefore considered them to be sufficiently experienced to provide us with relevant data concerning our research questions. Typical demographic variables such as gender and age were not central to our investigations, and were therefore not collected. Participation in the experiment was voluntary and anonymous. No compensation was given to each individual participant, but 10 randomly selected participants were awarded with a gift card valid for a lunch meal at a local café.

Procedure. Prior to the administration of the experiment to the sample of medical students, a pilot version was administered to three students of psychology, in order to test the technical components of the forms, and to make sure that the content was understandable. The actual experiment was conducted in an auditorium during a break between two lectures. The professor briefly introduced the experimenters and the general topic of the experiment, conveyed as "decision making under uncertainty", and encouraged participation. The experimenters then explained the procedure, specifying that participation was voluntary and anonymous, and that all participants could withdraw from the experiment at any time without any consequences. Participation was done through laptop computers, tablets and smartphones, but did not require registration of any personal information, such as names or email addresses. As participation would occur through the internet connection of the hospital, the IP-address would be the same for all participants. Any incidentally stored data (IP-address, device or browser used) would be unavailable to any individual who could understand the collected data. Thus, the anonymity of the individual responses was deemed to be sufficiently secured.

As the experiment featured two "quasi conditions" (see *Materials* for an elaboration of this term) that were distinguished by receiving different variations of the case vignettes, two online forms were used. This was done because the online format was unable to randomly assign participants into separate versions of the forms. In order to assign each participant to one particular form, the link to a basic website was first presented to the assembly on a blackboard. On this website, there were two new links, each leading to one of the forms. The textual content of the links presented to the participants revealed nothing about the content of the corresponding forms, nor about any differences between the two. Based on where they were seated, the approximate left half of the classroom were instructed to follow the first link, while the right half were instructed to follow the second link. This division achieved a pseudo form of randomization of the participants into the two conditions. By clicking on the links to which they were assigned, the participants were then lead to the actual forms, featuring the two cases. Completion of the experiment took about 10 minutes. Afterwards, the participants were debriefed about the purpose and content of the experiment. Although the results of the experiment would not be known until the data had been analyzed, the experimenters offered to present the results to the class at a later time. In agreement with the class representative, a more thorough presentation of the goals and expectations of the study, as well as a summary of the results, was later distributed to the participants via email (see Appendix A).

Materials. The experiment consisted of two hypothetical psychiatric cases (labeled Case 1 and 2) developed by the authors. The cases were strictly based on diagnostic criteria from an ICD-10 manual (World Health Organization, 1999), and were presented in the same order for all participants. At the start of each case, the participants were informed that they would be presented with a fictional patient, and that they were to decide on the most probable diagnosis. This decision would be made by choosing one of two plausible diagnostic options. In Case 1, these options were: A) dementia (by Alzheimer's disease) or B) depressive episode (subsequently labeled *depression*). The participants were informed that they were to state which of these two diagnoses they thought would be most likely to be correct at three different time points throughout the case. They were also informed that, while assessing the case, they would be able select a follow-up question to ask the patient on two occasions. Following this introductory information, the participants were presented with lists of diagnostic criteria for dementia and depression, as described in the ICD-10 manual (World Health Organization, 1999). The participants were instructed to base their decisions on these diagnostic criteria, rather than any prior knowledge they had about the diseases, such as their respective base rates in the general population. Following the presentation of the diagnostic

criteria, the participants were presented with a short vignette describing the patient in question.

The case vignette began with some general information about the patient (i.e. age, sex, occupation). Both quasi-conditions were then presented with a single, coherent paragraph, featuring six pieces of information that described various symptoms and details about the state of the patient. All participants received identical pieces of information, presented through the exact same sentences: Two of the pieces featured clinical details that favored diagnosis A, two favored diagnosis B, while the remaining two were generally compatible with both diagnoses (i.e. *neutral*). The experimental manipulation that distinguished the two conditions came in form of the order in which these six pieces were presented. In one condition, symptoms compatible with dementia were presented earlier in the vignette description, whereas symptoms compatible with depression were presented later. The opposite was done for the other condition. The two neutral pieces were spread out in the paragraph, in order to make the contrast between the symptoms supporting either diagnosis less prominent. However, these pieces appeared in the same positions across both conditions. The label "quasi-condition" entails that the two groups received identical treatment throughout the experiment, with the exception of the initial symptom presentation in each case. The manipulation had no relevance when testing the hypotheses not involving anchoring bias (H2-H4). However, the two groups will henceforth be referred to as conditions. A complete list of all the clinical content presented in the cases, including the order in which it was presented for the two conditions, is presented in Appendix B.

After the vignette presentation, the participants were asked to state a preliminary diagnostic hypothesis (*T1 diagnosis and certainty*). This was done by selecting a value on a horizontal 10-point scale. The extreme left end of the scale (1) represented the highest level of certainty that diagnosis A was most likely for this patient. Similarly, the extreme right (10) represented the highest level of certainty that diagnosis B was the most plausible diagnosis. The closer to the mid-point the participants checked, the lower degree of certainty they expressed in their selected diagnosis. Using an even number of points in the scale meant that the participants were unable to give an entirely neutral response, and had to state a diagnostic hypothesis, even though they may have been only marginally more confident in this option than in the other. This was done in order to obtain valid responses with regard to our research questions.

Following the selection of a preliminary diagnosis, the participants were presented with the forced opportunity to select one of four relevant follow-up questions that they could

ask the patient (*T1 request*). This was done by presenting a list of four options. The first two options featured questions that would investigate clinical details indicating support of dementia (labeled *Request A1* and *Request A2* in the analyses), while the latter two options featured questions that would investigate clinical details indicating depression (labeled *Request B1* and *Request B2)*. The participants were subsequently led to a new slide displaying the patient's answer to the question they had selected. Crucially, regardless of the question selected, the corresponding answers were designed to be vague and ambiguous to such a degree that objectively, they did not conclusively support or rule out either of the diagnoses. However, the information could still be interpreted by the participants as to support the diagnosis featured in the question. Following this section, the participants were once again asked to state their diagnostic hypothesis and their certainty in this hypothesis (*T2 diagnosis and certainty*), by checking on a 10-point scale, identical to that at T1. The participants were then able to select an additional follow-up question to ask the patient (*T2 request*). The options and the order in which they were presented were identical to those of T1. Although it was possible for the participants to select the same option at both T1 and T2, we did not expect that many would do this. Finally, after reading the answer to the selected question, the participants were to state their final diagnosis and level of certainty (*T3 diagnosis and certainty*) on a scale identical to those used at T1 and T2. Figure 1 demonstrates the structure of the cases featured in Experiment 1.

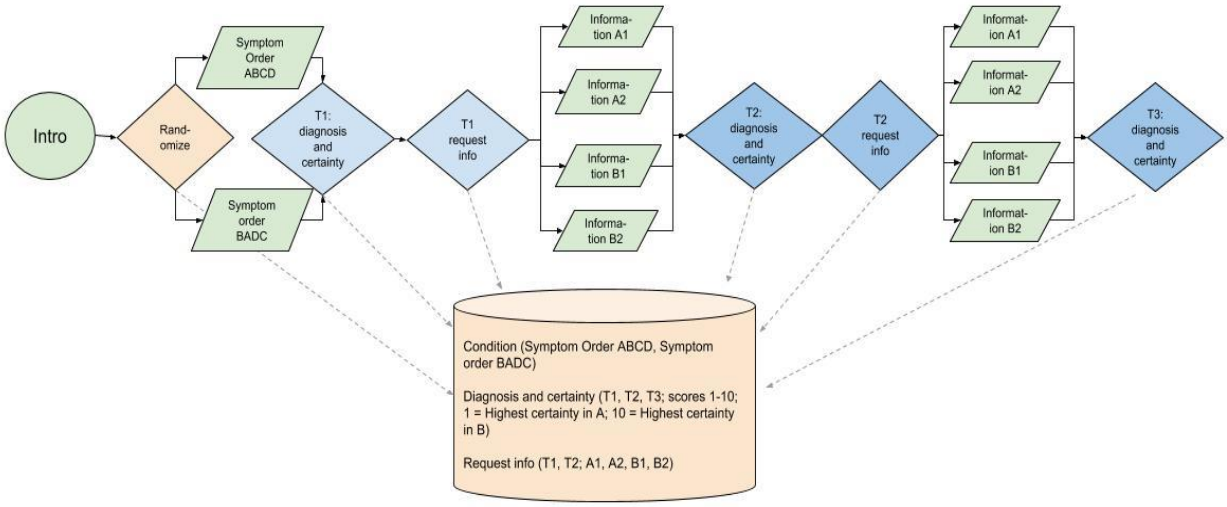


Figure 1. The structure of the cases in Experiment 1, as exemplified by Case 1. A large version is available at the OSF website for the research project (https://osf.io/dn4rv/)

After completing Case 1, the participants were introduced to Case 2, which followed the same structure. However, the diagnoses used in this case were C) bipolar mood disorder and D) borderline personality disorder. Importantly, the order in which the two diagnoses in each case were presented was always the same for all participants, with the exception of the aforementioned manipulation of the order of symptoms in the vignettes. That is, in Case 1, elements related to diagnosis A consistently appeared before those related to diagnosis B, both in the case instructions, in the lists of ICD-10 criteria related to each diagnostic option, in all selections of diagnostic options, and in all selections of follow-up questions. Similarly, in Case 2, all such elements related to diagnosis C were consistently presented before those related to diagnosis D. As participants in one of the conditions would read a vignette in Case 1 in which the first symptoms would match the diagnostic option presented first everywhere else in the case, the conditions were intended to be counterbalanced, as specified in our pre-registration for the experiment. This step was expedient in order to balance out any unintended effects of consistently being presented with elements related to one particular diagnostic option before the other. Practically, it would entail that for Case 2, participants in the same condition would read a vignette in which the first symptoms presented were consistent with the diagnostic option appearing *last* everywhere else in the case (i.e. case instructions, etc.). The other condition would follow the opposite pattern across the two cases.

However, due to an error in the development of the online forms, this counterbalancing was not implemented as intended: In Experiment 1, participants in one condition read vignettes in which the first symptoms matched the diagnostic options listed first everywhere else, not only in one of the cases, but in both. This condition was thus labeled *AB-CD*, rather than *AB-DC,* as had been intended. Conversely, participants in the other condition only read vignettes in which the first symptoms matched the diagnostic options listed last everywhere else in the cases. This condition was thus labeled *BA-DC* for Experiment 1. Despite this error, it was still possible to investigate the effects of symptom presentation (H1) as intended, as the conditions still received different vignette variations in each case. While we did not expect that this lack of counterbalancing would produce any systematic differences in the responses across the two conditions with regard to any of our investigations (H1-H4), we would perform additional examinations of the results to check for any indications of such unintended effects.

Data analysis. Our data collection was done through Google Forms. Google Spreadsheet and Microsoft Excel was used to prepare the data for analysis. The statistical analyses were performed using SPSS Statistics 24 and 25. All our hypotheses were tested by

investigating the participants' averaged responses across both cases. Recoding of the data for this purpose was done where necessary.

***Recoding of the data*.** In order to test our hypotheses, for each case, we recoded all *diagnosis and certainty* variables (*T1*, *T2* and *T3*) into six variables. Deriving from the scores on the 10-point scales, three of them were dichotomous variables categorizing the participants by choice of diagnosis, while the other three were continuous variables indicating the reported confidence in the selected diagnosis: *T1 diagnosis*, *T2 diagnosis* and *T3 diagnosis* were calculated by transforming the values of the corresponding *diagnosis and certainty* scores: Scores of 1-5 were given the value *diagnosis A* (*diagnosis C* for case 2), while scores 6-10 were given the value *diagnosis B* (*diagnosis D for* case 2). *T1 Confidence*, *T2 Confidence* and *T3 Confidence* ranged from 1 to 5, with 1 representing the lowest level of confidence in the selected diagnosis, and 5 representing the highest (original scores 1 and 10 = 5; 2 and 9 = 4; 3 and 8 = 3; 4 and 7 = 2, 5 and 6 = 1).

To simplify our test of H1, the responses to *T1 diagnosis* in each case were recoded into two indices based on the condition to which the participants were assigned. The resulting indices, *Initial diagnosis matches first symptoms*, were dichotomous, with 0 and 1 as possible values, corresponding to the categories *No* and *Yes*, respectively. The values from these two indices were then combined into the index *Initial diagnoses indicate anchoring*, with possible values ranging from 0 to 2, representing the number of times this occurred for each participant across the two cases. To simplify our tests of H2, H3 and H4, the responses to *T1 request* and *T2 request* in each case were recoded in order to distinguish between confirmatory and dissenting information requests, relative to the preceding diagnosis selected (*T1 diagnosis* and *T2 diagnosis*, respectively). *T1 request* and *T2 request* were therefore respectively recoded into *T1 request confirming info* and *T2 request confirming info*, each with the possible values of 0 and 1, corresponding to the categories *No* and *Yes*, respectively.

In order to test H2, an index was calculated based on the four *request confirming info* indices, which aggregated the number of instances across both cases in which the information requests aimed to confirm the preceding diagnostic selections. The resulting index, *Instances of seeking confirming info*, had possible values ranging from 0 to 4. To test H3, average scores were calculated for all confidence ratings (using the *T1* and *T2 Confidence* scores described above) that preceded confirmatory information requests (all instances in which *request confirming info = Yes*) across the two cases. Similarly, an average confidence score was calculated for all confidence ratings preceding dissenting information requests (*request confirming info = No*) across the cases. This generated the test indices *Average confidence*

stated on diagnosis preceding confirmatory info request and *Average confidence stated on diagnosis preceding dissenting info request*, each with a possible range from 1 to 5 (and missing values for participants with no responses matching the particular pattern).

To test H4, an index was calculated, detecting all instances in which participants selected the same diagnosis at all time points (T1, T2 and T3) and exclusively requested confirming information at both time points (T1 and T2) in a particular case. For all such instances (*Same diagnosis T1, T2, T3 and confirming info on T1, T2 = Yes*), we subtracted the *T1 Confidence* value from the *T3 Confidence* value, in order to investigate the developments in confidence for the participant in question. For participants who showed this consistent response pattern in either of the cases or in both of them, we calculated an average score: The test index *Average change in certainty for consistent diagnosis and confirmatory information request* had a possible range from -4 to 4 (or missing values for participant who did not show this pattern in any of the cases). Thus, this variable would measure the changes in confidence for participants who showed this pattern in one of the cases, and the average changes in confidence across the two cases for those who showed this pattern in both of them. The file *Data File - Both Experiments - Complete*, available at the OSF homesite for the project, includes all the recodings done for the responses in the experiments.

Statistical tests. No specific steps were taken to exclude outliers, as our data generally rendered little room for outlier values. The inference criteria followed the specifications made in the pre-registration: A standard alpha value of <.05 was used as the cutoff value. All tests were one-tailed, as we investigated effects in particular directions. Calculations of Cohen's *d* would be reported for statistically significant results.

H1: *Participants will be more likely to select the preliminary diagnosis congruent with the symptoms presented first in the vignette, rather than selecting the diagnosis congruent with symptoms presented later.* Specifically, in Case 1, participants in condition AB-CD, who read the *Symptoms A first* vignette variation were expected to more often choose preliminary diagnosis A than B at T1. On the contrary, participants in condition BA-DC, who read the *Symptoms B first* vignette variation, were expected to more often choose preliminary diagnosis B than A at T1. For Case 2, H1 predicted similar patterns with regard to diagnosis C and D. Our expectations for participants to generally select the diagnosis suggested by the vignette variations would be supported by a mean level of *Initial diagnoses indicate anchoring* for all participants to significantly exceed the level of 1. Thus, a one-sample t-test was performed, comparing the number of initial diagnoses matching the first symptoms for all participants (*Initial diagnosis matches first symptoms*) against a reference constant of 1. The

reference constant reflected the null-hypothesis, which predicted that participants would select an initial diagnosis that matched the symptoms listed first in only one of the two cases (50%). This prediction implied that the effect of the symptom order on diagnostic choice would be no stronger than mere chance (50/50).

H2: *In their requests of additional information, participants will more often select items that appear to support their preceding preliminary diagnosis, than they will select those that appear to favor an alternate diagnosis.* Specifically, in Case 1, we expected participants who selected diagnosis A at the preceding opportunity (T1 and T2) to more often seek to confirm this diagnosis by choosing *Request A1* or *Request A2* in the following information gathering phase. Similarly, participants who selected diagnosis B were expected to more often select *Request B1* or *Request B2*. In Case 2, we expected similar relationships between selecting diagnosis C and *Request C1* and *C2*, and between selecting diagnosis D and *Request D1* and *D2*. In sum, H2 predicted that, across both cases, participants would significantly more often than not (i.e. more than two times out of the possible four) seek to confirm their preceding diagnostic hypothesis. Thus, a one sample t-test was performed, comparing the of the number of confirmatory information requests for all participants (*Instances of seeking confirming information*) against a reference constant of 2. The reference constant reflects the value predicted by the null-hypothesis, which implied that participants would be no more likely to select confirmatory information than they were to select dissenting information.

H3: *When participants are given the opportunity to request additional information about a clinical case, requests for "confirmatory" information will be preceded by higher levels of confidence in the previously selected diagnosis, than requests for "dissenting" information.* Specifically, we expected requests for confirmatory information (at T1 and T2 in both cases) to be preceded by significantly higher levels of confidence in the selected diagnosis, than requests for dissenting information. By contrast, the null-hypothesis predicted that there would be no difference between the levels of confidence preceding confirmatory and dissenting information requests, respectively. We therefore performed a t-test for dependent samples, comparing the participants' scores on *Average confidence stated on diagnosis preceding confirmatory info request* with their scores on *Average confidence stated on diagnosis preceding dissenting info request*. Cases with missing scores in either of these indices were excluded from the analysis.

H4: *In instances where participants have stated the same diagnosis throughout the case evaluation, and have also exclusively requested "confirmatory" information throughout the case, the participants will report an increase in confidence in their diagnoses at the end of*

the case. More specifically, in instances where participants had stated the same diagnosis throughout the case evaluation (at T1, T2 and T3), and had requested confirmatory information at both instances (T1 and T2), we expected that the participants would report a significant increase in confidence in their diagnoses after the information gathering (stating higher confidence at T3 than at T1). We thus performed a one-sample t-test of the *Average change in certainty for consistent diagnosis and confirmatory information request* variable against a reference constant of 0, which would indicate no change, as predicted by the null-hypothesis. Cases with missing scores in this index were excluded from the analysis.

Results - Experiment 1

Out of the 71 students participating in the experiment, 37 completed the form corresponding to condition AB-CD, while the remaining 34 completed the BA-DC form. As our analyses required data from both cases for each participant, only complete responses were included.

H1. A one-sample t-test was conducted on *Initial diagnosis indicate anchoring* against a reference constant of 1, in order to determine whether participants significantly tended to favor the preliminary diagnoses suggested by the symptoms presented first across the two cases, rather than selecting the alternate diagnoses. The average occurrences of initial diagnosis matching the first symptoms (M = 1.00, SD = .74) was identical to the reference constant of 1, a statistically non-significant mean difference of 0, 95% CI [-0.17 to 0.17], $t(70) = 0$, $p = .5$, one-tailed.

H2. A one-sample t-test was conducted on *Instances of seeking confirmatory info* against a reference constant of 2, in order to determine whether participants tended to request confirmatory follow-up information significantly more often than requesting dissenting information across the two cases. The one-sample t-test showed that the average number of requests for confirming information (M = 1.93, SD = 0.76) was close to the reference constant of 2, a non-significant mean difference of -0.07, 95% CI [-0.25 to 0.11], $t(70) = -0.78$, $p = .22$, one-tailed.

H3. A paired-samples t-test was conducted to investigate whether participants' levels of confidence preceding confirmatory information requests significantly exceeded their levels of confidence preceding dissenting information requests. Only scores for participants who had selected at least one of each type of information throughout the two cases were included (n = 67). The levels of *Average confidence stated on diagnosis preceding confirmatory info request* (M = 2.43, SD = 0.85) were somewhat lower than the levels of *Average confidence*

stated on diagnosis preceding dissenting info request (M = 2.64, SD = 0.88). The direction of the discrepancy was therefore the opposite of that predicted by H3, with a mean difference of -.21, 95% CI [-0.45 to 0.03], $t(66) = -1.76$, $p = .96$, one-tailed.

H4. A one-sample t-test was conducted on *Average change in certainty for consistent diagnosis and confirmatory information request* against a reference constant of 0, to determine whether there was any significant increase in stated confidence from T1 to T3 for participants who had been entirely consistent throughout any or both of the cases (n = 12). The one-sample t-test showed that the mean score for these participants (M = -0.17, SD = 1.34) was close to the reference constant of 0, a non-significant difference of -0.17, 95% CI [-1.02 to 0.68], $t(11) = -0.43$, $p = .34$, one-tailed.

Investigating effects resulting from the lack of counterbalancing. As described under *Materials*, our design intended to counterbalance the conditions so that, across the two cases, participants in each condition would read one vignette in which the first symptoms matched the diagnostic option appearing first everywhere else in the case, and one in which the first symptoms matched the diagnosis presented last. As noted, an error led to one condition (AB-CD) being presented with vignettes in which the first symptoms matched the diagnoses presented first everywhere else in both Case 1 and Case 2 (instead of only in Case 1). Consequently, across both cases, the other condition (BA-DC) was only presented with vignettes in which the first symptoms matched the diagnosis presented *last* everywhere else. It was plausible that the absence of such a counterbalancing could contribute to systematic differences in the responses between the conditions. Follow-up analyses were therefore conducted, comparing the two conditions in terms of their scores on the variables used in the statistical tests of the hypotheses. These analyses did not indicate that the two conditions were systematically affected by the unintended differences in the setups to which the they were subjected. That is, the response patterns across the two conditions were largely similar, implying that the absence of counterbalancing had no unintended effects on the participants' performances. The results of these analyses are therefore not reported in detail.

Discussion - Experiment 1

Our statistical tests showed that the null-hypotheses could not be rejected for any of our predictions. While there may be several possible explanations for our non-findings, examinations of the responses indicated certain unintended imbalances in the case material. The results for the selections of preliminary diagnoses (at T1), showed that more participants tended to select diagnosis B (depression) in Case 1 and diagnosis D (borderline personality

disorder) in Case 2, regardless of the order of symptom presentation. In other words, many participants who had been intended to be led towards diagnoses A and C through the vignette manipulations diverted from the expected responses, contributing to the null-finding. These unexpected skewnesses in the responses imply that the symptoms indicating depression and borderline personality disorder were somehow seen as stronger indicators for their respective diagnoses, compared to the symptoms intended to indicate dementia (A) and bipolar disorder (C). By overriding any potential effects of the order in which the symptoms were presented, such imbalances may thus have undermined our investigations of anchoring bias (H1). Further signs of imbalances were found in the selections of follow-up questions in each case: While intended to be equally relevant, certain options were much more commonly selected than others. Some appear to have been so "attractive" that they may have overridden any confirmatory tendencies that could otherwise have existed, thus possibly undermining our investigations of H2.

As imbalances in the follow-up questions may have undermined our attempts to investigate confirmatory tendencies, they may also have affected the corresponding investigations of confidence in several ways. By overriding any real distinctions between confirmatory and dissenting information requests, it is conceivable that the skewnesses in the perceived relevance of certain symptoms featured in the follow-up questions may have led to the null-finding for H3. If the symptoms featured in a "dissenting" follow-up question (relative to the selected diagnosis) somehow appear more informative than those featured in the other options, a high level of confidence in one's existing diagnostic hypothesis should not rule out selecting this question. Such imbalances may thus help to explain the slightly higher levels of confidence preceding the selection of dissenting information, which went against the predicted direction. Very few participants qualified for inclusion in the test for H4, undermining our investigation of this prediction. It was not possible to predict beforehand how many would remain entirely dedicated to a particular diagnosis throughout a case. However, the low number of eligible participants may partially have been caused by imbalances in the perceived relevance of certain symptoms featured in the follow-up questions. As addressed above, this may have led many to pursue options that diverted from their stated hypotheses, thus reducing the number participants who would stay entirely consistent throughout a case.

In short, various potential imbalances in the case material may have affected the results of all the tests. A follow-up experiment was therefore constructed, aiming to correct

for some of the issues discussed above. This was administered to a new sample of medical students.

Goals - Experiment 2

The goals and methods for Experiment 2 were largely similar to those of Experiment 1. Our expectations regarding confirmation bias and diagnostic confidence (H2-H4) also remained unaltered. Aside from attempting to improve the balance in the case material, the main difference in Experiment 2 was a slight broadening of our investigations regarding the effects of symptom presentation on the participants' choice on a preliminary diagnosis (H1). In Experiment 1, anchoring bias was operationalized strictly in terms of the order in which the symptoms were presented. In Experiment 2, however, we would also implement slight variations in the number of symptoms indicative of different diagnoses across the vignette variations. This was done as a means to induce stronger diagnostic beliefs in the participants in a controlled fashion, without eliminating the element of uncertainty. Inducing stronger diagnostic beliefs could possibly facilitate our investigations of confirmation bias, which remained the main focus of our experiment. Practically, the change in the symptom presentation entailed that in each vignette, a particular diagnosis would be favored, not only by the order (primacy) of certain symptoms, but also by their slight majority, compared to the alternate diagnosis. Succeeding in improving the balance between various symptoms would entail that such quantitative and positional properties of the symptoms, rather than any qualitative differences (i.e. in strength) would suggest a particular diagnosis. Corresponding to these changes, H1 now predicted that participants would be more likely to select the preliminary diagnosis indicated by both the order (earlier) and the number (higher) of symptoms, than to select the alternate diagnosis. Selecting a preliminary diagnosis favored by the symptom presentation could still be regarded as to reflect a type of anchoring, as it could entail "locking on to salient features presented early" (see *Cognitive Biases in Clinical Reasoning: Conceptualizations*). However, as a particular diagnosis would now be indicated by both primacy and a slight majority of symptoms, any effects found could not safely be judged to reflect anchoring bias as defined in previous studies.

Methods - Experiment 2

Experiment 2 was registered as *Follow-up experiment on anchor and confirmation bias in setting diagnoses* (https://osf.io/s8nwf/register/565fb3678c5e4a66b5582f67).

Participants. 57 advanced medical students participated in this experiment. While belonging to a different class than the participants in Experiment 1, these students were also approaching the completion of their medical degree at the time of the experiment. The compensation arrangements were the same as in Experiment 1. No demographical data was collected.

Procedure. The experiment was administered in the same fashion as Experiment 1. The only important difference with regard to the procedure, was that the subsequent debriefing would have to occur entirely via email due to time constraints (see Appendix A).

Materials. The psychiatric cases used in Experiment 2 were generally very similar to those of Experiment 1. No changes were made with regard to the general structure of the cases, nor to the diagnoses featured. The experiment once again used 2 conditions, differentiated by the presentation of the initial symptoms. For this round, the two conditions were counterbalanced as intended (see *Materials* under Experiment 1), and were thus labeled AB-DC and BA-CD. A slight change was implemented in the order in which the follow-up questions were presented. The questions investigating each diagnosis were now presented in an alternated order (i.e. *Request A1*; *B1*; *A2*; *B2*) rather than sequentially (i.e. *Request A1*; *A2*; *B1*; *B2*). This was done as a means to further balance the perceived relevance of each question as well as the relevance of the corresponding diagnoses they appeared to support. Figure 2 demonstrates the structure of the cases featured in Experiment 2.

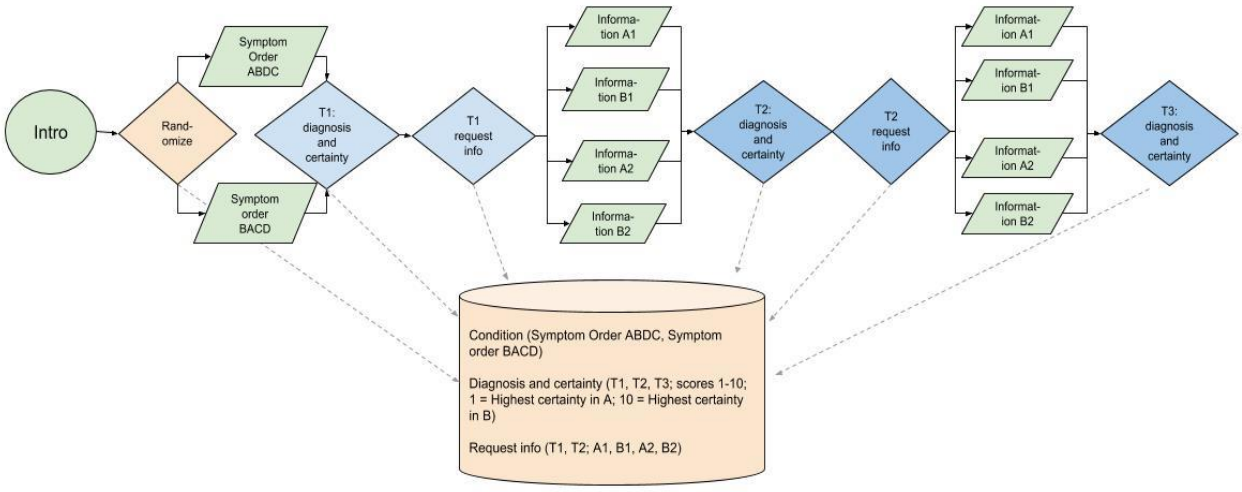


Figure 2. The structure of the cases in Experiment 2, as exemplified by Case 1. A large version is available at the OSF website for the research project (https://osf.io/dn4rv/).

The main difference between Experiment 2 and Experiment 1 concerned the vignette variations. In each case vignette, a neutral symptom was presented first, followed by two symptoms indicating a particular diagnosis. However, unlike in Experiment 1, these would now only be followed a single symptom indicating the other diagnosis. Aside from the changes in the number of symptoms, there were slight modifications in the wordings of certain symptoms, questions and answers. This was done in order to better balance the strength of the elements related to the two diagnostic options, and to avoid the unintended skewnesses in preferences found in Experiment 1. Additionally, a few modifications were made in the initial lists of ICD criteria presented prior to each vignette, in order to make certain relevant symptoms less obscure. The case material used in Experiment 2 is presented in Appendix C.

Data analysis. All recoding of the data was done in an identical fashion to that in Experiment 1, resulting in the same indices used for the pre-registered tests. The sole difference from Experiment 1 with regard to expectations, was that the testing of H1 now encompassed manipulations of both the order and the number of symptoms. However, the statistical procedure used to test this hypothesis was identical to that in Experiment 1. The remaining hypotheses and tests (H2-H4) remained unaltered. As in Experiment 1, a standard alpha value of <.05 was used as the inference criterium for all statistical tests. Additionally, as all hypotheses predicted specific directions regarding the effects under inquiry, all statistical tests were one-tailed. Calculations of Cohen's *d* would be reported for statistically significant results.

Results - Experiment 2

Out of the 57 students participating in the experiment, 25 completed the form corresponding to condition AB-DC, while 31 completed the BA-CD form. One participant in condition BA-DC only competed Case 1. As our statistical tests required complete data from each participant, this response was excluded from the analyses, leaving 56 complete responses to be subjected to the tests. As in Experiment 1, typical demographic variables such as gender and age were not central to the research questions, and were thus not collected.

H1. A one-sample t-test was conducted on *Initial diagnosis indicate anchoring* against a reference constant of 1, in order to determine whether participants across the two cases (at T1) favoured the diagnoses indicated by two symptoms presented earlier in the vignettes, rather than the alternate diagnoses indicated by a single symptom presented afterwards. The average occurrences of initial diagnosis matching the first symptoms (M = 1.00, SD = 0.71)

was identical to the reference constant of 1, a non-significant mean difference of 0, 95% CI [-0.19 to 0.19], $t(55) = 0$, $p = .5$, one-tailed.

H2. A one-sample t-test was conducted on *Instances of seeking confirmatory info* against a reference constant of 2, in order to determine whether participants tended to request confirmatory follow-up information significantly more often than requesting dissenting information across the two cases. The average number of requests for confirming information (M = 2.23, SD = 1.04) was close to the reference constant of 2, a non-significant mean difference of 0.23, 95% CI, [-0.05 to 0.51], $t(55) = 1.66$, $p = .05$, one-tailed.

H3. A paired-samples t-test was conducted to investigate whether participants' levels of confidence preceding confirmatory information requests significantly exceeded their levels of confidence preceding dissenting information requests. Only scores for participants who had selected at least one of each type of information throughout the two cases were included (n = 47). The levels of *Average confidence stated on diagnosis preceding confirmatory info request* (M = 2.38, SD = 0.91) were almost identical to the levels of *Average confidence stated on diagnosis preceding dissenting info request* (M = 2.38, SD = 1.08), a non-significant mean difference of <.01, 95% CI [-0.30 to 0.30], $t(46) = .01$, $p = .49$, one-tailed.

H4. A one-sample t-test was conducted on *Average change in certainty for consistent diagnosis and confirmatory information request* against a reference constant of 0, to determine whether there was any significant increase in stated confidence from T1 to T3 for participants who had been entirely consistent throughout any or both of the cases (n = 20). The one-sample t-test showed that the mean score for these participants (M = 0.48, SD = 1.36) was fairly close to the reference constant of 0, a non-significant mean difference of 0.48, 95% CI [-0.16 to 1.11], $t(19) = 1.56$, $p = .07$, one-tailed.

Discussion - Experiment 2

Experiment 2 was designed as a follow-up to Experiment 1, largely examining the same effects while attempting to correct some of the detected shortcomings. Once again, our statistical tests showed that the null-hypotheses could not be rejected for any of the research questions. Despite the stronger attempts to lead participants in particular diagnostic directions through manipulations of the case vignettes, they did not follow the predicted patterns with regard to selections of preliminary diagnoses (H1). Among the diagnostic options available in each case, depression (diagnosis B) and borderline personality disorder (diagnosis D) remained the most commonly selected diagnoses, even though half of the participants were only presented with a single symptom favoring these particular diagnoses. This indicates that

imbalances in the vignette material remained. The general levels of diagnostic confidence were similar to those of Experiment 1, implying that we did not manage to induce stronger diagnostic hypotheses in the participants in this round. As for requesting additional information (H2), we did not detect any response patterns indicating confirmation bias. Certain imbalances in the "popularity" of the various follow-up questions remained, despite our attempts to correct these for Experiment 2. Even though the popularity of the various questions in Case 2 were somewhat more equal than their counterparts in Experiment 1, this was not true for Case 1. We thus appear to only have been partially successful in improving the balance of various questions in terms of perceived relevance and importance.

As participants appeared rather unbiased in their preferences of additional information with regard to their selected diagnostic hypotheses, the absence of any clear differences in confidence levels preceding confirming and dissenting information gathering (H3) is logical. Indeed, the respective levels of confidence preceding confirmatory and dissenting information gathering were practically identical in Experiment 2. Even though mean levels of diagnostic confidence among "entirely consistent" participants increased from T1 to T3 (H4), this change did not reach statistical significance. While higher than in Experiment 1, the number of participants who followed such a pattern in any or both cases remained quite low (n = 20). Given the relatively large increase in diagnostic confidence (0.48) among the participants who did stay consistent, it is possible that a larger sample could have rendered a significant result for this test. However, as in Experiment 1, imbalances in the perceived relevance and strength of the follow-up questions may also have contributed to the low number of participants qualified for inclusion in the test. Once again, this may have led more participants to stray from their stated diagnosis than what would have been the case if the follow-up questions had been equally strong. Additionally, the non-significant developments in confidence among those who did qualify for the test may have been partly caused by the follow-up information being excessively ambiguous, as will be addressed in our general discussion of this hypothesis.

Summary of the Results

Below we summarize the findings from both experiments regarding our pre-registered hypotheses and statistical tests.

H1. In both experiments, we tested whether varying the presentation of symptoms in somewhat ambiguous psychiatric case vignettes would affect the choice of a preliminary diagnosis. In Experiment 1, this was done by varying the order in which symptoms indicative

of different diagnoses appeared. We expected that participants would favor the diagnosis suggested by the symptoms presented first. As the total amount of information presented in each vignette variation was identical, such a pattern would indicate anchoring bias. In Experiment 2, symptom presentation in the vignettes not only favored a particular diagnostic option through primacy (order), but also through a higher number of symptoms indicating this diagnosis. Despite these manipulations, the tests for the effect of symptom presentation did not yield significant results in either experiment. That is, across the two cases in the experiments, participants were no more likely to select a diagnosis favored by the symptom presentation than they were to select the alternate diagnosis.

H2. We investigated whether participants were significantly more likely to select follow-up questions that appeared to support their previously stated diagnostic hypothesis than they were to pursue information regarding the alternate diagnosis. While the follow-up questions in both experiments concerned the same symptoms, some of the questions available in Experiment 2 featured slightly different wordings. Nevertheless, the findings in both experiments were the same: In each experiment, participants were no more likely to select questions indicating support of their previously selected diagnoses than they were to select concerning the alternate diagnoses.

H3. We tested our prediction that the levels of diagnostic confidence preceding selections of confirmatory information would be higher than those preceding selections of dissenting information. The results from the experiments failed to support this hypothesis. In Experiment 1, the responses showed a slight tendency to follow the opposite pattern, entailing slightly higher levels of confidence preceding selections of dissenting information than those preceding selections of confirmatory information. In Experiment 2, we found no difference in levels of confidence between the two styles of information gathering.

H4. Finally, we investigated how confidence in a diagnosis developed throughout a case for participants who did not change their diagnosis or explore dissenting information at any point during the case. We expected that the reported levels of diagnostic confidence for these participants would have significantly increased by the end of a case compared to their initial levels. Such tendencies would indicate that these participants had interpreted ambiguous and inconclusive follow-up information in a confirmatory manner. This prediction, however, was not supported in either experiment, as neither result reached statistical significance. That is, there were no clear developments in diagnostic confidence among such “entirely consistent” participants.

Follow-up Analyses

For the pre-registered analyses, we recoded and averaged the participants' responses across the two cases. These steps were done in order to simplify the statistical analyses, and to provide more reliable measurements. The recoding was based on the assumption that the two cases were practically equivalent in terms of their potential to induce the effects under inquiry. In other words, there should not be any systematic differences in the results across the two cases. However, there existed a possibility that the two cases used in the experiments were somehow unequal in terms of inducing the various effects. Differences between the cases would not necessarily be detected by the confirmatory analyses, and could result in a misleading null-finding, based on the averaged responses. For instance, in each experiment it was possible that participants would generally choose a total of two confirmatory follow-up questions across the two cases, but that these selections would generally occur more often in one of them (e.g. Case 1). The test of H2, using a reference constant of 2, could then yield a non-significant result, despite signs of confirmation bias in this one case. The two cases could also differ in terms of their potential to induce the other effects under inquiry (H1, H3 and H4).

As the two cases may have differed in some unforeseen way, follow-up analyses were also conducted, investigating the predicted effects in each individual case from both experiments. Overall, these analyses yielded the same null-findings as our main tests. An exception was the test for H2 in Case 1 in Experiment 2. As there were two possibilities for selecting follow-up information per case, the pooled responses from these two instances were tested against a reference constant of 1, which reflected an equal tendency to select confirmatory and dissenting information in a case. The mean number of confirmatory information requests in this case ($M = 1.21$, $SD = 0.76$) exceeded the reference constant of 1, entailing a statistically significant mean difference of 0.21, 95% CI [0.01 to 0.42], $t(55) = 2.12$, $p = .02$, one-tailed). The calculated Cohen's d value was .28, indicating a small effect, according the standards recited by Mulhern and Greer (2011). While this one test provided some evidence for H2, none of the other cases in either experiment produced any significant results regarding this hypothesis. Given the large amount of tests performed to investigate for each effect in each individual case across both experiments, it is not very surprising that a single test yielded a significant result, which may be a product of mere chance. Furthermore, while reaching statistical significance, the effect was not very substantial. H2 therefore largely remains unsupported. In general, the results of the individual cases indicate that our choice to average responses was unproblematic, as they were mostly similar across both cases in each

experiment. While we cannot conclude that the two cases were completely equal, it appears that neither case succeeded in inducing the effects in question.

As the sample size in each experiment was not very large, this may have reduced the statistical power of our tests, possibly contributing to our null-findings (see Mulhern & Greer, 2011). In an effort to increase the statistical power, exploratory analyses of the hypotheses were also performed using the pooled responses from both experiments, increasing the sample size to 127. This was possible as there were generally only minor differences in the case material used in the two experiments. Still, as the material was not identical, great caution would be taken when interpreting any significant results. The statistical analyses used for testing the hypotheses on the combined sample were the same as for the individual experiments. These analyses revealed no significant results with for any of our four hypotheses. Given the similarity in the responses between the individual experiments, and the generally high *p*-values from the confirmatory tests, the non-findings in these follow-up analyses were unsurprising. However, these tests do to some extent pre-empt the notion that the null findings of the individual experiments resulted from low statistical power.

General discussion

A substantial amount of research has demonstrated how diagnostic decisions can be affected by various cognitive biases, which may contribute to occurrences of diagnostic errors. Previous findings indicate that slight variations in the presentation of symptoms favoring different diagnoses can influence a clinician's diagnostic decisions, even when the presentation only differs in terms of the order in which these symptoms appear. This may manifest itself through the occurrence of anchoring bias, whereby information presented early is given more weight than information presented later, even though all the information presented may be equally relevant. Erroneous decisions may also be a result of confirmation bias, which refers to searching and/or interpreting information in ways that primarily supports one's existing beliefs, rather than being open to alternate views and explanations that are equally plausible. Furthermore, literature indicates that unwarranted confidence in one's diagnostic beliefs and decisions may be related to biased confirmatory tendencies when dealing with information.

In a series of two classroom experiments, we sought to investigate the occurrence of anchoring bias and confirmation bias in two samples of advanced medical students, using two written psychiatric cases. We also wished to explore whether high levels of diagnostic confidence were primarily related to confirmatory styles of seeking and interpreting

information about the patients. Across the two experiments, we found no evidence for anchoring bias, nor for confirmation bias. Furthermore, we did not find evidence for systematics relationships between higher levels of diagnostic confidence and confirmatory selections and interpretations of follow-up information. The similarity in the results between the experiments occurred despite attempts to induce stronger diagnostic hypotheses and correcting imbalances in the second experiment.

Investigating Anchoring Bias

The investigations of anchoring bias in Experiment 1 used a design that largely resembled those used in previous literature. While failing to support our hypothesis, the results in this experiment are not incompatible with those of earlier contributions: Some prior studies using hypothetical cases have found evidence for anchoring bias, whereas others have not (see Cunnington et al, 1997; Ellis et.al, 1990; Friedlander & Stockman, 1983; Richards & Wierzbicki, 1990). The reasons for these mixed findings are not entirely clear. Richards and Wierzbicki (1990) discussed the lack of coherence, based on existing literature, as well as the equivocal findings from their own studies. They speculated that there could be certain aspects or factors in some clinical cases previously used that modified the effect of anchoring. However, the authors pointed out that earlier attempts had not controlled for levels of severity in their clinical material. In other words, featuring sufficiently balanced case material may represent a recurring issue in the investigations of anchoring bias. If symptoms supporting various diagnoses are not equally strong, participants may easily end up favoring one diagnosis over the other, relatively independent of the order in which the symptoms are presented. As noted, this appears to have been an issue in both of our experiments: Participants tended to favor depression (diagnosis B) in Case 1 and borderline personality disorder (diagnosis D) in Case 2 when selecting a preliminary diagnosis. Even in Experiment 2, in which we had attempted to correct the apparent imbalances, and only presented one symptom indicating these diagnoses to half of the participants, such skewnesses persisted.

Achieving sufficient balance in the vignette material is a complex task. In our experiments, a distinction can be made between balance *within* symptoms and balance *between* symptoms. The former entails that all symptoms in a case vignette have to maintain certain degrees of ambiguity, allowing them to be interpreted as to support a particular diagnosis without objectively allowing for any diagnosis to be confidently selected or discarded. For the neutral symptoms in the vignettes, the ambiguity needs to be evenly balanced between the diagnoses. Symptoms intended to indicate support of a diagnosis, on the

other hand, have to be sufficiently strong to achieve this without entirely eliminating the plausibility of an alternate diagnosis. “Between-symptoms” balance entails that all such ”weighted” symptoms featured in a case vignette should be equally potent in this regard. The complexity in making the vignette material balanced in these different ways may have been augmented by the restricted length of the vignettes, which was a necessity brought on by practical circumstance, as addressed under *Limitations*. While we cannot say exactly what made the symptoms favoring diagnoses B and D more convincing to our participants, it is evident that we did not entirely succeed in our manipulation attempts.

Investigating Confirmation Bias in the Search for Information

The two experiments were similar in terms of goals and procedure used for investigating confirmation bias in the search for information (H2), and featured only minor differences in the clinical material. In a similar fashion to our investigations of anchoring bias (H1), our tests for H2 may have been somewhat undermined by unintended imbalances in the case material, here in the follow-up questions available to the participants. In principle, when investigating confirmation bias in the selection of follow-up questions, all questions should appear to be equally informative and relevant to the participants. As addressed, for reasons not clear to us, certain questions in our experiments appear to have been particularly attractive to the participants in some way, relatively independent of the participants’ preceding diagnostic hypotheses. For some participants, these questions may have overridden confirmatory preferences that would otherwise have guided their choices. Such departures from their stated diagnostic hypotheses may have contributed to the null-findings.

Compared to previous studies, our results appear to represent something of an anomaly. Several earlier studies have found evidence of confirmation bias in the search for additional information (Martin, 2001; Mendel et al., 2011; see also Payne, 2011). One notable difference between our study and those conducted previously pertains to the amount of material included in the experiments. Practical concerns (see *Limitations*) led us to include less clinical information than what was featured in other studies. For instance, we only gave the participants 4 options of follow-up information to pursue in each case, and only allowed them to select 2 of these. By contrast, Mendel and colleagues (2011) gave the participants a total of 12 options, 6 favoring each diagnosis, and also allowed them to select as many as they wanted. In his simulated therapy sessions, Martin (2001) allowed participants 15 minutes to ask the patient whichever questions they wanted, rather than forcing them to select different options. Including fewer options increases the necessity for each option to be equally strong.

The restrictions in the amount of follow-up material that could be pursued may have accentuated any imbalances that may have existed in the strength and perceived importance of each element.

Investigating Anchoring Bias and Confirmation Bias Simultaneously

A distinct feature of our experiments was the combined investigation of anchoring and confirmation bias within the same cases. By testing for both at the same time, we designed the cases in such a way that no correct diagnoses could ultimately be identified. While participants could believe certain diagnoses to be correct, the case material would not objectively warrant any definitive conclusions to be made at any point during the cases. This lack of genuinely "right answers" sets our approach apart from previous studies on confirmation bias within the psychiatric field (Martin, 2001; Mendel et al., 2011; Parmley, 2006; see also Oskamp, 1965), which had also measured diagnostic accuracy in various ways, through the implementation of correct and incorrect diagnostic options. Both Mendel and colleagues (2011) and Parmley (2006) investigated confirmation bias by initially leading some or all participants towards the incorrect diagnoses through the opening vignettes. In order to ultimately arrive at the correct diagnoses, participants were forced to be sufficiently open to alternative explanations and options. While such an approach would be possible to combine with a simultaneous investigation of anchoring bias, the limited sample sizes in our experiments would have made such a combination problematic: That is, in principle, one could lead participants in one condition in the "wrong" direction by presenting symptoms supporting the incorrect diagnosis earlier in the vignette variation, in accordance with our operationalization of anchoring bias. However, such a design would entail that the confirmation bias could only have been realistically investigated in this condition, with the other condition, having been led in the "right" direction, representing a control group. As we only expected to recruit a maximum of 80 participants for Experiment 1, and only later knew for sure that we would be able to conduct a second experiment on a new sample, we opted not to reduce the investigations of confirmation bias to just one of the conditions in our original design. As the sample size expected for Experiment 2 was the same as for Experiment 1, and our investigative goals mostly remained the same, we chose not to alter the design by implementing correct diagnoses.

While our approach differed from previous investigations by not including objectively correct diagnoses, the participants in our study were still instructed to state which diagnoses they thought most likely to be correct. As participants did not know that there were no actual

"right answers", we assume that they would still attempt to identify a correct diagnosis, thus still possibly being susceptible to the effects under inquiry. However, the need to maintain varying, yet ever-present, degrees of uncertainty throughout the cases brought on a risk that certain elements in the case material could become more ambiguous than what was optimal with regard to investigating biased thinking. As discussed below, excessive ambiguity does not seem to have acted as a primary cause for the majority of our null-findings. However, it may have contributed to the null-findings for H4.

Confirmatory Tendencies and Diagnostic Confidence

Previous findings in psychiatry gave us some reason to expect a positive relationship between confidence and confirmatory tendencies in the diagnostic process. Our experiments tested two predictions (H3 and H4) concerning such relationships. H3 predicted that confirmatory tendencies in the search for additional information would be preceded by higher levels of confidence in one's existing diagnostic hypothesis, compared to searches for dissenting information. While this expectation stands to reason, it is somewhat troubling that there were no clear tendencies for participants to predominantly search for information classified as confirmatory, as demonstrated by our non-findings for H2. All questions featured in the cases were worded in ways that clarified which diagnoses they would potentially provide support for, and could thus technically be classified as either confirmatory or dissenting relative to each of the possible diagnostic options. However, we cannot be certain to what extent selecting a "confirmatory" detail actually implied a particular motivation to confirm one's existing hypothesis. For instance, participants may have selected questions, not necessarily in order to confirm the indicated diagnosis, but also to see if it could be refuted. Pursuing certain symptoms could reflect an interest, not only in their potential presence, but also in their absence. As noted, the skewnesses in the selections of follow-up questions in the experiments do indicate that some symptoms were in some way more interesting than others. While we attempted to make all follow-up options equally "attractive" to the participants, in order to avoid that the perceived informative value of particular questions overrode any biased motivation to confirm one's existing beliefs, it seems that we did not entirely succeed in achieving such balances. Instead, participants may still have pursued certain follow-up questions that they perceived to be the most informative, regardless of their stated diagnostic hypothesis. Given the rather neutral and unbiased diagnostic beliefs indicated by our results, the lack of support for H3 seems reasonable.

Findings in both experiments also went against the hypothesized relationship between exclusively pursuing a particular diagnostic option and experiencing increasing levels of confidence in this diagnosis (H4). An issue present in both experiments was that the number of eligible participants was probably too low for this hypothesis to be properly tested. As noted previously, the low number of participants qualified for inclusion in these tests may partly be caused by the imbalances in the perceived importance of certain follow-up question, restricting the number of entirely consistent participants in each case. For those who did stay consistent, the lack of significant developments in diagnostic confidence may be related to the high degree of ambiguity in the follow-up information. Following a dual process perspective, excessive ambiguity may have undermined biased interpretations of information by prompting a more analytical System 2 approach (see Croskerry, 2009b). While our design necessitated some uncertainty to be present throughout the cases, the vignettes and the follow-up questions were designed to be more or less weighted towards certain diagnostic options. As implied in our discussion of H1 through H3, unintended imbalances in the case material seem more likely to have been the most central challenge to our investigations, rather than excessive ambiguity. It is not clear how the aforementioned skewnesses in the choices of preliminary diagnoses and follow-up questions should systematically relate to ambiguity in the case material, for instance by being the product of analytical processing triggered by such ambiguity. For H4, however, ambiguity may have played a more central role.

Unlike the follow-up questions, the corresponding answers were designed to be so vague that they did not clearly point in any diagnostic direction, in accordance with the aforementioned absence of correct and incorrect answers. At the same time, the information had to maintain a potential to be erroneously interpreted as supportive of a particular diagnostic option. To balance these two demands was challenging, which, in combination with the limited length of the pieces of follow-up information (see *Limitations*), may have made the information too ambiguous, given our purpose. Even though some participants stuck to the same diagnostic hypothesis throughout a case, they received no information that objectively warranted an increase in their confidence. In accordance with the connection between excessive ambiguity and analytical processing indicated by the dual process perspective (see Croskerry, 2009b), it is plausible that the participants realized that the information contained no genuine support for the diagnosis in question. Remaining analytical, these participants may thus more easily have stayed unconvinced by the information received. It is possible that the ambiguous information could still have been interpreted in a confirmatory manner, if the genuine confidence in the in the corresponding diagnosis had

been sufficiently high prior to receiving it. However, as evident from our results, participants did generally not have such strong diagnostic beliefs upon receiving the follow-up information. Furthermore, if excessive ambiguity in the follow-up information meant that high levels of preceding confidence was necessary for biased confirmatory interpretation to occur, it would still be unlikely for us to find clear support for H4, which predicted *increases* in confidence. While this hypothesis could well be true in certain real-life situations (see below), stronger case-material would be necessary in order to properly test for such tendencies in an experimental setting.

The Prevalence and Detectability of Cognitive Biases in Diagnostics

Many recent publications acknowledge that phenomena such as anchoring bias and confirmation bias may occur in medical settings (see Blumenthal-Barby & Krieger, 2015; Croskerry, 2002, 2003; Klein, 2005; Saposnik et al., 2016; see also Crowley et al., 2013, Norman, 2009). While neither of these constructs stem from medical literature (see Tversky & Kahneman, 1974; Hahn & Harris, 2014), their relevance in diagnostic settings is logically sound: It makes sense that a clinician may lock onto early stimuli during a patient encounter, even though these may be no more important than stimuli revealed later. Similarly, it seems perfectly plausible that clinicians may deal with information in ways that primarily aim to confirm their existing beliefs. Additionally, the notion of diagnostic confidence being related to such tendencies stands to reason. However, to investigate such biases empirically also reflects an acceptance of certain assumptions regarding their true prevalence in medicine, as well as their experimental detectability. These interrelated assumptions are also relevant to discuss, given our non-findings.

The real prevalence and impact of anchoring bias and confirmation bias in psychiatric diagnostics may be very difficult to unearth. This is partly because susceptibility to such biases will likely be highly dependent on the individual case, due to the aforementioned complexity of factors that characterize a given diagnostic situation. As a consequence, it is also difficult to estimate how likely it is for such biases to actually be found in a given experimental investigation of diagnostic decision making. On one hand, Norman and Eva (2010) argued that many biases have been established experimentally through clever manipulations intended to "induce error for the sake of determining if the biases exist" (p. 97). If such artificial contexts are necessary in order for certain biases to be detected, one might suspect that the biases could be less common in real life than certain experimental studies suggest. On the other hand, it is also possible that biases such as anchoring and confirmation

bias are very real phenomena in diagnostics, but that it is sometimes difficult to detect them experimentally. Just as the chances for biases to occur in a real life may be highly dependent on the characteristics and context of the particular diagnostic situation, this may also be true for the occurrence of biases in various experimental situations.

While we have recited experimental contributions within psychiatry that indicate support for the existence of the various effects under inquiry in our study, the investigations conducted so far are quite sparse. This is especially true for the investigations of developments in diagnostic confidence. One may wonder whether a considerable amount of non-findings regarding all of the investigated phenomena remain unpublished, which appears to be a common issue in psychology at large (see Ferguson & Brannick, 2012). While it is only possible to speculate, the apparent acknowledgement of such biases as genuine phenomena in diagnostics may make it less tempting to report on failures to detect them. Unpublished non-findings may thus contribute to the impression that biases such as those investigated in our study are quite common when, in fact, they are not. In addition to the sparsity, the variability in the findings made so far is relevant to consider. The existing findings regarding anchoring bias are mixed, and the occurrence of confirmation bias also varies between studies, even within psychiatry. For instance, Mendel and colleagues (2011) found indications of confirmation bias in just 13% of the clinicians featured in the study, whereas up to a third of the clinician's in Parmley's (2006) study showed such tendencies. While relevant when questioning the real prevalence of such phenomena, the variability may also reflect the different approaches used to investigate them, with different designs and populations yielding different results.

Although relevant when examining our null-findings, we cannot answer the complicated questions regarding the actual prevalence and detectability of the various investigated phenomena. To what degree our null-findings stem from methodological or practical characteristics of our experiments, and to what degree they reflect a legitimate part of the variability in the occurrence of the phenomena in real diagnostic settings, is impossible to know for sure at this point. Our approach has been topically and methodologically inspired by previous experiments that have actually indicated the existence of the various phenomena. The relative scarcity of existing empirical literature, as well as the variability in the findings, make inquiries such as ours useful in ultimately increasing our understanding of the true roles of cognitive biases in diagnostic decision making. Through our work, we have identified some challenges for such investigations which, in turn, may provide useful guidelines for later research.

Limitations

Our study attempted to investigate a combination of effects that to our knowledge remained unexplored in psychiatric diagnostics. While making a unique contribution to the field, the combination of effects tested simultaneously in our study, combined with limited resources for testing them, provided certain challenges. There were considerable restrictions in the time allowed for the completion of each experiment, particularly Experiment 2, that were brought on by conducting them during breaks between lectures. Our reasons for carrying out the experiments in these settings had to do with certain inherent advantages. Namely, utilizing such natural breaks for the students, aided by encouragements from the lecturer to partake in the experiments, allowed for a fair number of participants to be recruited with ease. Furthermore, time-pressure may exist in many real diagnostic situations, and may contribute in provoking biased thinking (see Croskerry, 2009b), which was the general topic of our investigations.

However, in the case of our experiments, the time limitations may have affected other aspects of the ecological validity of the experiments. The short time alloted to introducing, carrying out and concluding the experiments implied a need to limit the amount of information and options that could be included in the cases. In addition to the inherently limited realism in investigating psychiatric diagnostics in an experimental setting with hypothetical written cases (see for example Mendel et al., 2011), these restrictions may have further decreased the realism of the cases. In a real diagnostic process, physicians would be able to obtain a lot more information than our experiments allowed for. They could also choose freely what questions to ask and what symptoms to pursue. In our experiments, the vignette material and the follow-up questions featured had to be carefully selected and worded. The number and types of details that could be included was restricted, not only by the need to keep the cases fairly short, but also by the inherent properties of certain symptoms indicative of the various diagnoses: Many symptoms that would be of interest to a diagnostician in a real situation were simply too characteristic of a particular diagnosis to be allowed for inclusion in the cases. That is, certain key symptoms would too easily warrant the selection or rejection of particular diagnosis, thus undermining the uncertainty of the case. It is likely that such inclusions could result in even more considerable issues regarding certain diagnoses and follow-up questions being more commonly selected. While our aims made it necessary to exclude certain key symptoms from the case material, this further detracted from the realism of the cases, thus providing another limitation concerning ecological validity.

As noted previously, restrictions in the material included in the experiments also posed a challenge to inducing the effects under inquiry in a controlled fashion. The main challenge of investigating symptom presentation (H1) in the way we intended arguably laid in balancing the symptoms properly in several different ways at the same time, as elaborated previously. While short vignettes facilitated efficient completion of the experiments, featuring a low number of symptoms put more pressure on each detail in terms of being perfectly balanced, and left less room and fewer opportunities for "fine tuning" the strength of the vignette material at large. This is not to say that longer vignettes are inherently superior: While shorter vignettes could augment any imbalances due to the relative salience of each detail, longer vignettes would entail more possibilities for imbalances to occur overall. Additionally, including more ambiguous text in a vignette could possibly weaken any effects beyond what would be expedient, by excessively "burying" the prominence of certain details. Nevertheless, given the non-findings in our investigations, it is possible that insufficient amounts of vignette material may have limited the potential of the effects in question to be induced.

Similar practical factors may have had an impact on our investigation of confirmatory tendencies and diagnostic confidence. Much like using a short case vignette, a low number of available follow-up questions may have accentuated any imbalances in the strength and perceived importance of each option, producing unintended effects on the participants' selections of follow-up information (H2). This may, in turn, have affected our measurements of confidence preceding these selections (H3), as well as reducing the number of participants who stayed devoted to a particular diagnosis throughout a case (H4). Furthermore, as noted, keeping the corresponding follow-up answers short may, when combined with the need for sufficient balance and ambiguity, have increased the vagueness of the content beyond what was expedient, with regard to measuring developments in diagnostic confidence (H4). As with more substantial vignette material, it is possible that longer pieces of follow-up information could increase the opportunities for "fine tuning" the material in an expedient fashion. However, this is not to say that the challenge of achieving the right degree of balance would necessarily be easier, due to similar potential issues as those addressed regarding longer vignettes.

Another limitation concerns the development of the clinical content featured in the experiments. Due to practical circumstances, the material was primarily developed by the authors, who lack clinical education and experience. In this regard, it must be noted that the cases were developed in coordination with a trained clinical psychologist, and were strictly

based on ICD-10 criteria for the featured diagnoses. These factors partially compensate for our lack of experience in psychiatry. However, it is possible that trained psychologists or psychiatrists could have produced more realistic, balanced or effective cases, even with the practical limitations surrounding our experiments. In each experiment, we attempted to balance the foundation for the participants' responses by asking the them to base their decisions and assessments on the clinical material in the cases, as well as the presented ICD criteria corresponding to the various diagnoses. Another potential issue concerns the fact that the participants only saw the list of ICD criteria once at the start of each case, in order to allow for quick completion. It is possible that the participants were presented with too much information at once, and that his may have affected their responses in ways that we cannot detect. For instance, some may have forgotten about notable symptoms in the lists, and opted to rely on what they had previously learned about the various conditions. While not possible to detect, such occurrences could contribute to unintended variety in the participants' responses.

Strengths, Implications and Future Directions

To our knowledge, our work represents the first attempt to simultaneously investigate anchoring bias, confirmation bias and developments in diagnostic confidence using written psychiatric cases. Furthermore, prior to our study, these phenomena do not appear to have been thoroughly investigated using Norwegian medical students. By pre-registering our experiments before collecting any data, we facilitated transparency of the research process by making a priori commitments to investigate a specific set of hypotheses, through methods that were clarified in advance. While we did not find support for our hypotheses, and have discussed the prevalence and detectability of the phenomena in question, we still regard them as relevant for achieving an increased understanding of cognitive errors in diagnostic decision making. For this purpose, our experiments may represent a template for future investigations. Their online format makes them easy to administer to large samples. Furthermore, their general structure is easy to replicate, revise and broaden. In the following, we propose several ideas for potential refinements and expansions, many of which derive from challenges and limitations addressed previously.

As evident from our discussions, the time allotted to conducting the experiments and the amount of content in the clinical cases are interrelated elements that could both be expanded, with several potential benefits for the investigations. Longer vignettes, along with greater opportunities for selecting and obtaining additional information, could increase the

resemblance of real patient consultations, thus improving the ecological validity. Furthermore, as implied previously, expansions of the various elements could also facilitate in making the material sufficiently balanced, given similar research goals. Concerning the vignettes and the follow-up information, expansions could entail greater opportunities for "fine tuning", as discussed under *Limitations*. Giving participants more follow-up questions, and allowing them to select more of them, could reduce the potential effects of imbalances by making each option less prominent. Furthermore, even if certain options would somehow remain particularly attractive to all participants regardless of their stated diagnostic hypotheses, such selections could be identified and controlled for in subsequent analyses. This could potentially allow for confirmatory tendencies to still be investigated among the selections of the remaining options. Expansions of the cases may also be made with regard to the number of times in which diagnostic hypotheses and follow-up information can be selected (i.e. T3, T4, T5). This can possibly allow for more thorough and nuanced investigations of the participants' diagnostic reasoning and decision making, as well as corresponding developments of diagnostic confidence.

In addition to increasing the scope of each case, it is possible to increase the number of cases featured in an experiment, potentially including cases from other fields of medicine, such as internal medicine. Such expansions could also yield more output and insight into the participant's thinking, spanning across a greater variety of clinical situations. Increasing the duration and the scope as suggested here could require a more substantial and thorough recruitment process, as it would demand more of the participants' time and effort. In contrast to our approach, by which both recruitment and testing was done within a very short time, it could be beneficial to recruit participants further in advance, allowing them to partake in the study at agreed-upon times. Additionally, for longer procedures, it would probably be expedient to provide compensation for each individual participant. Taking steps such as these may lead to the recruitment of a considerable amount of motivated subjects, even for a larger and more demanding experiment.

Our results indicate that it may be difficult to simultaneously investigate anchoring bias and confirmation bias, due to the inherent challenges in achieving sufficient balance in the case material without making it excessively ambiguous. However, we believe that it is possible to combine the investigations of these phenomena, provided sufficient amounts of participants, time and material. Given a sufficiently large sample, it would be possible to combine the investigations by implementing objectively correct diagnoses, as discussed under *Investigating Anchoring Bias and Confirmation Bias Simultaneously*. A simple design in this

regard would still only include two diagnostic options per case. However, in order to properly test all the intended effects, it would probably be beneficial to somewhat increase the materials featured in each case, compared to what was included in our experiments. In each case, one condition could initially be led towards an incorrect option through manipulations of the order of symptoms in the vignette. Confirmation bias could then be investigated in these participants' selections and interpretations of additional information, which would, at large, suggest that the other option was correct, similar to the design of Mendel and colleagues (2011). The other condition would act as a control in this case with regard to confirmation bias, having been led towards the right diagnosis in the vignette. However, anchoring bias could still be investigated in both conditions. By including an even number of cases, the two conditions could be led in the correct and the incorrect direction an equal number of times throughout the experiment, resulting in considerable output from each participant regarding both anchoring bias and confirmation bias. Additionally, implementing correct and incorrect diagnoses would allow for investigations of relationships between cognitive biases and diagnostic accuracy. Such a design would also easily allow for the inclusion of confidence measurements, enabling investigations of developments in diagnostic certainty similar to those featured in our study, as well as diagnostic overconfidence based on the actual accuracy of the diagnostic assessments.

It is also plausible that qualitative approaches could add valuable insight to participants' thinking when faced with such clinical cases. Examples include think-aloud protocols, structured or semi-structured interviews, focus groups, and questionnaires. These could all shed more light on how participants provide meaning to the clinical information encountered in a case, on which information they base their diagnostic hypotheses on at various times, on their reasons for their stated levels of certainty, as well as on why and how they make their choices when selecting follow-up questions. Another possibility for expansion concerns the comparison of individuals with different levels of expertise, such as novice medical students, advanced medical students, residents and certified doctors with various degrees of seniority.

Conclusions

Our investigations of susceptibility to anchoring bias and confirmation bias in psychiatric diagnostics on samples of Norwegian medical students did not detect any occurrence of such biases. Nor did they reveal any corresponding tendencies to be overly confident in one's diagnostic hypotheses. Our non-findings do not necessarily imply that

Norwegian medical students are not susceptible to errors related to such biases. Still, the actual prevalence of these phenomena in real medical setting, as well as their experimental detectability, are worth discussing. It is conceivable, however, that the non-findings in our study at least partially stem from methodological and practical characteristics of our design. Based on our work, we have made some proposals for future research, encompassing various refinements and expansions of our general approach. We believe that an expanded study could provide interesting and important insight into diagnostics, and the cognitive processes that influence diagnostic outcomes.

References

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a Cause of Diagnostic Error in Medicine. *The American Journal of Medicine, 121*(Suppl. 5), S2-S23. doi:10.1016/j.amjmed.2008.01.001

Blumenthal-Barby, J. S., & Krieger, H. (2015). Cognitive Biases and Heuristics in Medical Decision Making: A Critical Review Using a Systematic Search Strategy. *Medical Decision Making, 35*(4), 539-557. doi:10.1177/0272989X14547740

Croskerry, P. (2002). Achieving Quality in Clinical Decision Making: Cognitive Strategies and Detection of Bias. *Academic Emergency Medicine, 9*(11), 1184-1204. doi:doi.org/10.1197/aemj.9.11.1184

Croskerry, P. (2003). The Importance of Cognitive Errors in Diagnosis and Strategies to Minimize Them. *Academic Medicine, 78*(8), 775-780. doi:10.1097/00001888-200308000-00003

Croskerry, P. (2009a). Clinical Cognition and Diagnostic Error: Applications of a Dual Process Model of Reasoning. *Advances in Health Sciences Education, 14*(Suppl. 1), 27-35. doi:10.1007/s10459-009-9182-2

Croskerry, P. (2009b). A Universal Model of Diagnostic Reasoning. *Academic Medicine, 84*(8), 1022-1028. doi:10.1097/ACM.0b013e3181ace703

Crowley, R. S., Legowski, E., Medvedeva, O., Reitmeyer, K., Tseytlin, E., Castine, M., . . . Mello-Thoms, C. (2013). Automated detection of heuristics and biases among pathologists in a computer-based system. *Advances in Health Sciences Education, 18*(3), 343-363. doi:10.1007/s10459-012-9374-z

Cunnington, J. P., Turnbull, J. M., Regher, G., Marriott, M., & Norman, G. R. (1997). The Effect of Presentation Order in Clinical Decision Making. *Academic Medicine*, 72(10 Suppl. 1), S40-S42. Retrieved from https://journals.lww.com/academicmedicine/Abstract/1997/10001/The_effect_of_presentation_order_in_clinical.14.aspx

Ellis, M. V., Robbins, E. S., Schult, D., Ladany, N., Banker, J. (1990). Anchoring Errors in Clinical Judgments: Type I Error, Adjustment, or Mitigation? *Journal of Counseling Psychology, 37*(3), 343-351. doi:10.1037/0022-0167.37.3.343

Eva, W. K. (2001). *The influence of differentially processing evidence on diagnostic decision-making* (Doctoral dissertation, McMaster University). Retrieved from http://hdl.handle.net/11375/7187

Ferguson, C. J., & Brannick, M. T. (2012). Publication Bias in Psychological Science: Prevalence, Methods for Identifying and Controlling, and Implications for the Use of Meta-analyses. *Psychological Methods, 17*(1), 120-128. doi:10.1037/a0024445

Friedlander, M. L., & Stockman, S. J. (1983). Anchoring and publicity effects in clinical judgment. *Journal of Clinical Psychology, 39*(4), 637-644. doi:10.1002/1097-4679(198307)39:4<637::AID-JCLP2270390433>3.0.CO2-Q

Graber, M. L., & Carlson, B. (2011). Diagnostic error: the hidden epidemic. *Physician executive, 37*(6), 12-14, 16, 18-19. Retrieved from https://www.researchgate.net/publication/51921525_Diagnostic_error_the_hidden_epidemic

Graber, M. L. (2013). The incidence of diagnostic error in medicine. *BMJ Quality & Safety, 22*(Suppl. 2), ii21-ii27. doi:10.1136/bmjqs-2012-001615

Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic Error in Internal Medicine. *Archives of Internal Medicine, 165*(13), 1493-1499. doi:10.1001/archinte.165.13.1493

Graber, M., Gordon, R., & Franklin, N. (2002). Reducing Diagnostic Errors in Medicine: What's the Goal? *Academic Medicine, 77*(10), 981-992. Retrieved from http://journals.lww.com/academicmedicine/fulltext/2002/10000/reducing_diagnostic_errors_in_medicine__what_s_the.9.aspx

Hahn, U. & Harris, A.J. (2014). What Does It Mean to be Biased: Motivated Reasoning and Rationality. In Ross, B. H. (Ed.), *Psychology of Learning and Motivation*, (Vol 61, pp. 41-102). London, The United Kingdom: Elsevier. doi:10.1016/B978-0-12-800283-4.00002-2

Kahneman, D. (2003). A Perspective on Judgment and Choice: Mapping Bounded Rationality. *The American Psychologist, 58*(9), 697-720. doi:10.1037/0003-066X.58.9.697

Klein, J. G. (2005). Five pitfalls in decisions about diagnosis and prescribing. *British Medical Journal, 330*(7494), 781-183. doi:10.1136/bmj.330.7494.781

Martin, J. M. (2001). *Confirmation bias in the therapy session: The effects of expertise, external validity, instruction set, confidence and diagnostic accuracy* (Doctoral dissertation). Available from ProQuest Dissertations. (UMI No. 9978910)

Matlin, M. W. (2013). *Cognitive psychology* (8th ed. International Student Version). Hoboken, New Jersey, USA: Wiley.

Mendel, R., Traut-Mattausch, E., Jonas, E., Leucht, S., Kane, J. M., Maino, K., . . . Hamann, J. (2011). Confirmation bias: why psychiatrists stick to wrong preliminary diagnoses. *Psychological Medicine, 41*(12), 2651-2659. doi:10.1017/S0033291711000808

Mulhern, G., & Greer, B. (2011). *Making Sense of Data and Statistics in Psychology* (2nd ed.). Basingstoke, England: Palgrave MacMillan.

Nickerson, R. S. (1998). Confirmation bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology, 2*(2), 175-120. doi:10.1037/1089-2680.2.2.175

Norman, G. (2009). Dual processing and diagnostic errors. *Advances in Health Sciences Education, 14*(Suppl. 1), 37-49. doi:10.1007/s10459-009-9179-x

Norman, G., Young, M., & Brooks, L. (2007). Non-analytical models of clinical reasoning: the role of experience. *Medical Education, 41*(12), 1140-1145. doi:10.1111/j.1365-2923.2007.02914.x

Norman, G. R., & Eva, K. W. (2010). Diagnostic error and clinical reasoning. *Medical Education, 44*(1), 94-100. doi:10.1111/j.1365-2923.2009.03507.x

Oskamp, S. (1965). Overconfidence in Case-study Judgments. *Journal of Consulting Psychology, 29*(3), 261-265. Retrieved from https://www.stuttercut.org/detective/oskamp.pdf

Parmley, M. (2006). *The Effects of the Confirmation Bias on Diagnostic Decision Making* (Doctoral dissertation, Drexel University). Retrieved from https://idea.library.drexel.edu/islandora/object/idea%3A1164

Payne, V. L. (2011). *Effect of a Metacognitive Intervention on Cognitive Heuristic Use During Diagnostic Reasoning* (Doctoral dissertation). Available from ProQuest Dissertations. (UMI No. 3471984)

Pines, J. M. (2006). Profiles in Patient Safety: Confirmation Bias in Emergency Medicine. *Academic Emergency Medicine, 13*(1), 90-94. doi:10.1197/j.aem.2005.07.028

Richards, M. S., & Wierzbicki, M. (1990). Anchoring errors in clinical-like judgments. *Journal of Clinical Psychology, 46*(3), 358-365. doi:10.1002/1097-4679(199005)46:3<358::AID-JCLP2270460317>3.0.CO;2-7

Saposnik, G., Redelmeier, D., Ruff, C. C., & Tobler, P. N. (2016). Cognitive biases associated with medical decisions: a systematic review. *BMC Medical Informatics and Decision Making, 16*(138), 1-14. doi:10.5167/uzh-127746

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*(4157), 1124-1131. doi:10.1126/science.185.4157.1124

Van den Berge, C.K.A. (2012). *Cognitive Diagnostic Error in Internal Medicine* (Doctoral thesis, Erasmus University Rotterdam, the Netherlands). Retrieved from http://hdl.handle.net/1765/31617

World Health Organization (1999). *ICD-10: Psykiske lidelser og atferdsforstyrrelser: Kliniske beskrivelser og diagnostiske retningslinjer.* Oslo, Norway: Universitetsforlaget.