Mitigating Bias in AI-supported Decision-making: Ophthalmologists' Perceptions of Bias Mitigation Strategies in Detecting Diabetic Retinopathy

Master's Thesis

June 2022

Anne Kathrine Petersen Bach, Jens Christian Brok and Trine Munch Nørgaard

Supervised by **Niels van Berkel**

Interaction Design
Department of Computer Science
Aalborg University

AALBORG UNIVERSITY

STUDENT REPORT

Resumé

Diabetes er en alvorlig sygdom som millioner af mennesker lider af verden over - og Danmark er ingen undtagelse. Faktisk er omkring 5% af den danske befolkning i øjeblikket diagnosticeret med sygdommen, hvilket forventes at stige til 8% i år 2030 [7]. Denne stigning vil uundgåeligt lægge et stort pres på det danske sundhedsvæsen, som allerede lider under manglende personale. I et forsøg på at mindske dette pres og øge lægers effektivitet ses der eksempler på brugen af *artificial intelligence* (AI) i forbindelse med udførelsen af screeninger for øjensygdommen *diabetisk retinopati* (DR) - en af mange komplikationer ved diabetes [8, 11, 13, 34]. Til at udføre screeninger benytter øjenlæger sig i forvejen af *clinical decision support systems* (CDSSs), og her kan AIs integreres for at understøtte screeningsprocessen [17]. Helt konkret kan AIs trænes til at opdage DR-relaterede forandringer i nethinden igennem billeder der tages af patienternes nethinde - også kaldet fundus billeder [23]. Disse billeder bliver, foruden brugen af AI, analyseret af øjenlæger manuelt, hvilket kan være en langsommelig og trættende proces i længden [27].

Selvom brugen af AI kan lede til en højere effektivitet [17] kan dén måde AI'ens resultater vises til øjenlægen på lede til forskellige former for kognitive bias, som kan resultere i fejlagtige diagnostiske beslutninger [3]. I dette speciale har vi undersøgt øjenlægerne på et dansk hospitals opfattelse af tre forskellige strategier til at mindske bias i deres nuværende arbejdsproces. Med vores arbejde håber vi på at kunne bidrage til forskningen inden for mindskning af bias i kliniske beslutningssituationer, som har målet at højne den diagnostiske nøjagtighed. I vores for-speciale undersøgte vi hvordan screeningsprocessen for DR foregår, hvilket vi gjorde igennem sessioner af contextual inquiry med øjenlægerne på et dansk hospital. Her fandt vi ud af, at måden hvorpå deres CDSS præsenterer sine resultater potentielt kunne introducere det kognitive bias *anchoring bias*, hvor man fejlagtigt baserer sin endelige diagnose på en initiel værdi, der kan vise sig at være irrelevant for beslutningen [3]. På den måde dannede vores for-speciale grundlaget for arbejdet i dette speciale, hvor vi har undersøgt seks øjenlægers opfattelse af tre strategier til at mindske *anchoring bias*. For at undersøge dette implementerede vi de tre strategier i øjenlægernes nuværende CDSS arbejdsgang ved designet og udviklingen af en interaktiv prototype, som vi kalder for the *debiasing workflow*. Den interaktive prototype brugte vi til at kommunikere, hvordan de tre strategier ville ændre deres nuværende arbejdsproces. Vi afholdte i alt seks evalueringer af the *debiasing workflow* med henblik på at forstå, hvordan øjenlægerne opfattede strategiernes potentielle brugbarhed i deres nuværende arbejdsproces.

Resultaterne af vores evalueringer viste at de tre strategier, ifølge nogle af øjenlægerne, havde potentialet til at øge den diagnostiske nøjagtighed, men de var for ingen et realistisk tiltag givet det vidtrækkende fokus på effektivitet. Derudover blev det klart at øjenlægerne opfattede vores *debiasing workflow* som værende en form for faktatjek af deres diagnoser, og at deres villighed til at overveje AI'ens forslag var påvirket af hvor meget de hver især stolede på dens nøjagtighed. Noget andet vi så var hvordan kompleksiteten af screenings-opgaven spillede en betydelig rolle for deres opfattelse af brugbarheden af vores *debiasing workflow*. Til sidst viste vores resultater at lægerne allerede tog forskellige bias-mindskende strategier i brug for at sikre overvejelsen af alternative diagnoser, hvilket bidrager til den diagnostiske nøjagtighed.

Mitigating Bias in AI-supported Decision-making: Ophthalmologists' Perceptions of Bias Mitigation Strategies in Detecting Diabetic Retinopathy

ANNE KATHRINE PETERSEN BACH, Aalborg University, Denmark
JENS CHRISTIAN BROK, Aalborg University, Denmark
TRINE MUNCH NØRGAARD, Aalborg University, Denmark

We are seeing an increase in the application of artificial intelligence (AI) into clinical decision support systems (CDSSs). Although this has the potential to alleviate clinicians' workload, the way AI output is presented to the users of AI-supported CDSSs potentially leads to cognitive biases that play a compromising role in the accuracy of diagnostic decisions. In this study, we investigate ophthalmologists' perceptions towards the implementation of three bias mitigation strategies into their current AI-supported CDSS workflow by designing an interactive prototype embodying these strategies. The prototype was evaluated through individual qualitative evaluations with six ophthalmologists to understand their view on the utility of the strategies in their current workflow. Our findings indicate that although the ophthalmologists saw potential in using the strategies to increase diagnostic accuracy, the necessity for efficiency, and the limited capabilities of the AI, rendered its use in practice unrealistic to some. Additionally, we found that varying task complexity levels had a substantial impact on the perceived usefulness of some of the strategies. Finally, the ophthalmologists perceived the strategies as precautionary measures, making some skeptical towards their use due to their self-perception as experts that outperform automated solutions.

Additional Key Words and Phrases: clinical decision support system, human-AI collaboration, artificial intelligence, interaction design, ophthalmology, diabetic retinopathy, cognitive reasoning, anchoring bias

ACM Reference Format:
Anne Kathrine Petersen Bach, Jens Christian Brok, and Trine Munch Nørgaard. 2022. Mitigating Bias in AI-supported Decision-making: Ophthalmologists' Perceptions of Bias Mitigation Strategies in Detecting Diabetic Retinopathy . In *CHI '22: ACM Conference on Human Factors in Computing Systems, April 30–May 06, 2022, New Orleans, USA.* ACM, New York, NY, USA, 25 pages.

1 INTRODUCTION

In parallel with the advancement within artificial intelligence (AI), clinicians in various medical fields are seeing an increase in its implementation into their workflows. In ophthalmology, AI technology has been implemented into clinical decision support systems (CDSSs) to assist ophthalmologists in their process of diagnosing eye disease in diabetes patients [8, 13, 34]. As of 2018, approximately 280.000 Danes were registered to have diabetes, making up 4,9% of the entire population of Denmark [7]. Diabetic Retinopathy (DR), an eye disease that causes damage to the blood vessels in the eye, is but one of several possible complications suffered by diabetes patients [11]. AI-supported CDSSs can aid ophthalmologists in reaching a decision, for instance pertaining to diagnosis or treatment, by giving actionable patient insights such as the location and nature of DR-related anomalies. In the past, AI-supported CDSSs have shown to contribute to the accuracy of diagnoses [19, 33].

Manuscript submitted to ACM

DR occurs as a result of high blood sugar levels damaging blood vessels in the eye and, if left untreated, can lead to severe loss of sight [15, 24]. As DR can be developing unbeknownst to the patient, it is in Denmark recommended that diabetics undergo regular retinal screenings to reduce the risk of losing sight [24]. Looking through various retinal images takes up a considerable part of the decision-making process that ophthalmologists go through when screening for DR. For instance, they examine fundus images to locate any potential retinal abnormalities related to DR [3]. Therefore, the introduction of the pattern recognition abilities of AI helps alleviate the ophthalmologists' workload by supporting them in the detection of these abnormalities [17].

Although the use of decision-supporting measures can contribute to more accurate diagnoses, other factors exist that play a role in diminishing this accuracy. In this paper, we focus on the cognitive bias, *anchoring bias*, that can arise as the result of using AI-supported CDSSs in the DR screening context. Cognitive biases are one example of the fallibility of human reasoning, and should therefore be acknowledged in this regard since they pose as compromising factors that can lead ophthalmologists to arrive at incorrect diagnoses [6]. Furthermore, despite technical advances of DR detection systems, the process of providing a holistic patient assessment requires the involvement of an ophthalmologist [4]. This underlines the relevance of investigating the aspects that make way for a successful collaboration between ophthalmologists and AI-supported CDSSs, in order to support a screening process that facilitates the skills of the ophthalmologist while mitigating potential cognitive biases at stake.

As a result of a pre-study carried out with seven ophthalmologists in the ophthalmology department of a Danish hospital, we discovered that the AI-supported CDSS used in their DR screenings presented its output in a way that could potentially lead to anchoring bias [3]. To reduce cognitive biases, such as anchoring bias, various mitigation strategies exist. In this study, we therefore investigate the implementation of three bias mitigation strategies into the DR screening workflow of the ophthalmologists who participated in our pre-study [3]. Our efforts will focus on taking steps towards successfully implementing bias mitigation strategies, as we believe that the willingness to use a system in which these strategies exist precedes the potential debiasing effects that the system may entail.

Specifically, we implemented the three bias mitigation strategies into the ophthalmologists' existing AI-supported CDSS workflow by designing an interactive prototype that embodies these strategies. This workflow will henceforth be referred to as the *debiasing workflow*. The prototype was evaluated with a total of six ophthalmologists in separate evaluation sessions to understand their perceptions of the debiasing workflow. The resulting data was analyzed using *Reflexive Thematic Analysis* [5]. Our efforts were guided by the following research question: "*What are the perceptions of ophthalmologists towards the incorporation of bias mitigation strategies into their AI-supported CDSS workflow in relation to the screening for DR?*".

Our findings firstly show that because the ophthalmologists are faced with the pressure of having to work as efficiently as possible, none of them thought the implementation of bias mitigation strategies into their workflow would be a realistic addition due to the extra steps brought about by these strategies. Despite this, we saw that some doctors envisioned the debiasing workflow having a positive effect on their diagnostic accuracy, while others thought that their level of diagnostic accuracy could not get any higher than it currently is. Furthermore, the doctors had different levels of openness towards the AI-generated insights, which was partly influenced by how they perceived the AI in the debiasing workflow to be fact-checking their diagnoses. For instance, some had more trust towards the AI's capabilities and were therefore in favor of the prospect of having the AI fact-check their diagnoses to potentially increase diagnostic accuracy. Contrarily, other doctors were convinced that the AI could not contribute with anything that would make them change their minds as for diagnoses. Additionally, the complexity levels of decision-making tasks seemed to influence the doctors' perceptions towards the usefulness of bias mitigation strategies in the given task. Specifically, they perceived the task of screening for DR as straightforward, resulting in them being more skeptical towards the need for bias mitigation in this case. Lastly, we found that the ophthalmologists work with different categories of diagnoses with different purposes in their effort towards making diagnostic decisions that are as accurate as possible.

The contributions of this work are: (1) insights into the context and workflow of ophthalmologists using an AI-supported CDSS to screen for DR, (2) concrete design suggestions for the implementation of bias mitigation strategies into a real-world AI-supported CDSS, and (3) an overview and assessment of the perceptions of ophthalmologists towards the application of bias mitigation strategies into their DR screening workflow.

The remainder of the paper is structured as follows: We start by covering related work within AI-supported CDSSs, cognitive reasoning, and bias in decision-making. Then, we present the Department of Ophthalmology and the pre-study from which this study takes departure. Following this, the three bias mitigation strategies used are presented along with how they are embodied in the debiasing workflow. We then present our methods for data collection and analysis, after which we present the resulting findings. Lastly, we discuss our findings and describe limitations and possibilities for future work.

2 RELATED WORK

In the following subsections, we firstly present research in the field of AI-supported CDSSs. Then, we outline relevant research efforts within cognitive reasoning, followed by an introduction to the role of biases in decision-making.

2.1 AI-supported Clinical Decision Support Systems

In recent years, there has been a rise in the application of AI into the field of medicine [14]. Exploring how AI may be used to assist clinicians in fields such as cardiology, internal medicine, oncology, and ophthalmology has garnered the interest of researchers around the globe [17]. For instance, Tao et al. [33] conducted a study in which they implemented an AI-supported CDSS into the existing decision support system (DSS) of a Chinese hospital to see what effect it had on the diagnostic accuracy. Based on data such as patient history and test reports, the CDSS recommended 10 probable diagnoses to assist the doctors in settling on a final diagnosis. They found that using the AI-supported CDSS had a positive effect on both diagnostic accuracy as well as efficiency, namely the time it took for doctors to arrive at a diagnosis [33].

In ophthalmology, the current process of detecting DR is typically manual, and can be a strenuous task for doctors [27]. Therefore, researchers are exploring ways for AI to assist ophthalmologists in their process of diagnosing patients as well as deciding on the treatment of retinal ailments such as DR [4]. Specifically, AI has been utilized to analyze retinal images to find indications of DR [8, 13, 34]. The image-centered nature of ophthalmology goes well with the pattern recognition abilities of AI technology, allowing the clinician to reach efficiency levels that are otherwise impossible to achieve [17]. Specifically, it has been proven that retinal anomalies like macular edema, exudates, cotton-wool, neovascularizations, and microaneurysms are possible to detect using AI [23]. In a study by Gulshan et al. [13], 54 ophthalmologists were employed to grade fundus images that would train a deep learning model to automate parts of the DR detection process. While the model was successful in detecting DR, the authors call for further research on how a system as such would be implemented into a clinical context [13]. Chandakkar et al. [4] argue, in spite of promising advancements in DR-related detection systems [1, 8, 34], that we have yet to reach a point where a system can be left to its own devices and that it is therefore crucial that the ophthalmologist remains a deciding part of the evaluation process [4]. Additionally, how AI-supported CDSSs are actually used in their intended context is a topic that has seen little attention as of yet [35]. According to Wang et al. [35], investigating the implementation of CDSSs is important, as AI-supported CDSSs are historically difficult to implement in a way that the user finds acceptable.

Thus, several authors highlighted in this subsection call for further research on how CDSSs can be successfully implemented into their clinical context [13, 35]. Therefore, we build upon this prior work by taking a human-centered approach through which we investigate the perceptions of ophthalmologists towards an AI-supported CDSS used for the detection of DR-related anomalies.

2.2 The Dichotomy of Cognitive Reasoning

In the past, researchers have highlighted how cognitive biases are an inherent quality of the human mind [6, 29–32, 37]. According to Croskerry [6]: "*Cognitive failures are best understood in the context of how our brains manage and process information*", and further elaborates on how diagnostic reasoning is driven by a dual process system. This system divides the various approaches to thinking into two opposite groups: System 1 and System 2. System 1 thinking is defined as being fast, intuitive, and requiring low effort. During this type of thinking, unconscious cognitive shortcuts are used to make quick decisions. On the other end of the scale, System 2 thinking is described as slow, analytical, and requiring high effort, making it a more rational form of cognitive reasoning compared to System 1 thinking. [6].

In a study on how high-risk professionals approach decision-making in high-stakes situations under conditions such as limited time or uncertainty, Klein [18] found that decisions were often based on System 1 thinking, as its rapid nature allows for a higher efficiency. However, though System 1 thinking can speed up the decision process, and is therefore often subconsciously chosen by the decision-maker, it may also oversimplify the situation and ultimately lead to diagnostic errors [22, 28]. In addition, relying on System 1 thinking can lead to overestimating one's experience and underestimating uncertainty in cases where the decision-maker is overconfident [22]. Lighthall and Vazquez-Guillamet [22] therefore argue that shifting from System 1 thinking to System 2 thinking may prevent these errors on account of the evidence-based nature of System 2 thinking. Additionally, because System 1 thinking is unreliable for inexperienced practitioners, shifting to System 2 thinking can benefit both novices, by giving them the tools to make a more reliable decision, as well as experienced practitioners by supporting them in complex cases [22].

With our study, we aim to move doctors into System 2 thinking to avoid the potential disadvantages that come with System 1 thinking. To do this, we use bias mitigation strategies that have the potential to diminish cognitive biases often leading to System 1 thinking.

2.3 Bias in Decision-making

Even though the human mind is prone to errors when it comes to decision-making, due to factors such as biases, logical fallacies, and assumptions, we still rely on medical practitioners to be able to make accurate decisions [6]. Critical care practitioners, for instance, need to be able to make these decisions in stressful and uncertain environments. Making wrongful diagnostic decisions can have severe impacts on patients and, in worse cases, have harmful consequences [22].

Although the process of deciding on a diagnosis is central to the medical field, the cognitive processes that underlie them are, most likely, the aspect of medical care we understand the least [22]. Many of the various types of biases resulting from error-prone cognitive patterns have become well-known in literature, describing more than 100 biases that affect clinical decision-making [6]. Anchoring bias is one example of a cognitive bias that may affect clinical decision-making. This specific bias occurs when the decision-maker has a distorted perception due to an anchor that fixates them on a decision close to said anchor [36]. In other words, it is the tendency to rely on the first piece of information received as a reference point, thus skewing the final decision [21]. This reference point may appear as a result of a computation, a starting point, or some other presented value from which the decision-maker typically makes insufficient adjustments [9]. Anchoring bias has shown to be common in many decision-making contexts and is particularly difficult to separate from, which has become evident in the seldom successful attempts at mitigating the bias [2].

With the existence of an anchor comes the risk of adjusting the initial diagnosis according to that anchor as a means to reach the final diagnosis, which is problematic as the anchor may be irrelevant. However, there is also the closely related risk of taking in the anchor as the final diagnosis and thereby ruling out any alternative diagnoses altogether, which is often referred to as premature- or early closure. In a study determining the relative contribution of system-related and cognitive components to diagnostic error, it was found that the most common cognitive cause of diagnostic error was premature closure, i.e. "*the failure to continue considering reasonable alternatives after an initial diagnosis was reached*" [10].

Hence, in this study, we investigate the perceptions of ophthalmologists towards the inclusion of three bias mitigation strategies into their AI-supported CDSS used in the DR screening process.

3 CONTEXTUALIZATION

In the following subsections, we elaborate on the context of focus in this study, namely the Department of Ophthalmology in a Danish hospital. First, we provide an overview by detailing the process of conducting screenings for DR, and the role that AI plays in the ophthalmologists' CDSS. Next, we describe the methods used as well as key findings from our pre-study.

3.1 The Department of Ophthalmology

Our data-gathering efforts were conducted at the premises of the Department of Ophthalmology in a Danish hospital. One of the core tasks of the department is for its ophthalmologists to carry out screenings for DR in diabetes patients. Each day, one ophthalmologist is on shift to watch for incoming patient cases from all over the region that need to be screened. To get screened for DR, diabetes patients visit a nurse at their local hospital to have various images taken of their retina, including fundus images. The images are then passed on to the ophthalmologist on shift for them to review remotely. When a patient's case is received in their system, the screening must be completed within one hour, ensuring that the patient receives their screening results quickly. Thus, the doctor carrying out the screening does not see the patient in person, but only handles the patient data that was collected at the patient's local hospital. To complete a screening, the ophthalmologist decides on the stage of DR present in a patient, ranging from no DR to a severe stage where treatment is needed. In addition, they decide on the next course of action depending on the stage of DR, ranging from initiation of treatment in worse cases to deciding on when the patient should be screened again. Screening intervals vary from just a couple of months for patients with severe DR, and up to two years for patients with no indications of DR.

To carry out DR screenings, the ophthalmologists use three separate applications. Firstly, they use an application that facilitates the incoming patient cases along with the patient journals where the resulting diagnosis is entered into. This application will henceforth be referred to as the *patient journal*. Secondly, they use a browser-view to inspect optical coherence tomographies (OCTs), which is a series of images also taken by nurses for the ophthalmologists to inspect during screenings for DR. Lastly, they use an application to view and assess the fundus images, henceforth referred to as the *fundus image application*. Upon entry into this application, all fundus images are automatically processed by an AI algorithm looking for two specific abnormalities in the retina; microaneurysms and hemorrhages, which are both

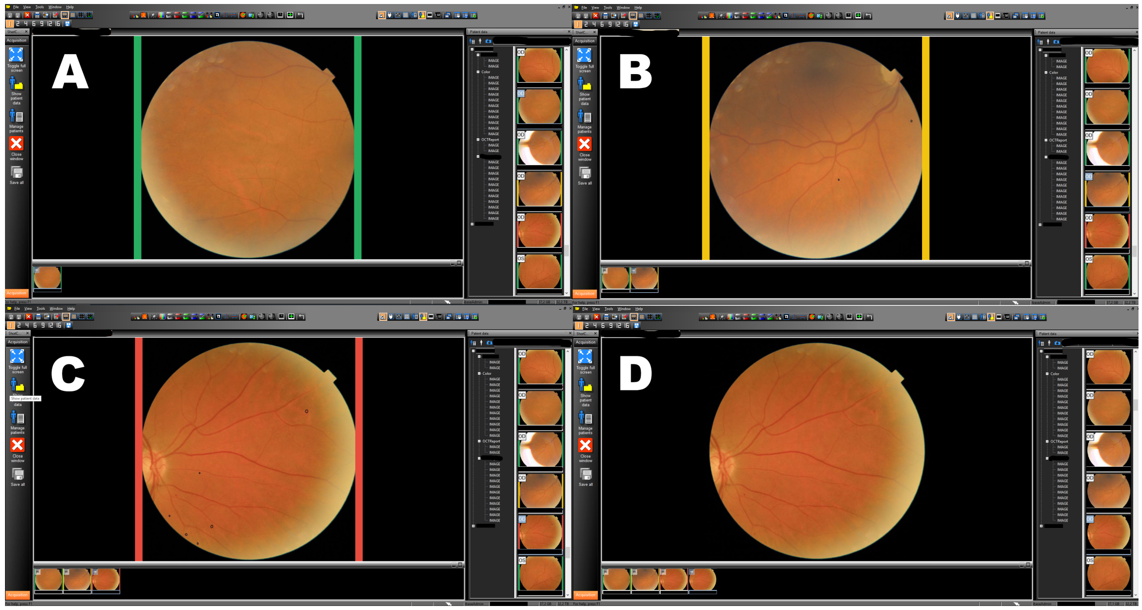


Fig. 1. The fundus image application, where the AI system A) has detected no abnormalities, labeling the image green, B) has detected between 1-3, labeling the image yellow, and C) has detected above 3, labeling the image red. D) shows an original fundus image without the AI system's visualizations. [3].

indicators of DR. This AI algorithm will henceforth be referred to as the *AI system*. All fundus images are labeled by the AI system with either green, yellow, or red, depending on the number of abnormalities detected in that fundus image. These color labels are visualized to the user through an outline of each fundus image with one of the three colors, as is viewed in Figure 1. The green label is applied whenever the AI system has not found any abnormalities in the retina; yellow when it has found one to three, and red is for when more than three abnormalities were detected. To visualize the specific abnormalities detected in each fundus image, the AI system creates a black outline directly on top of the microaneurysms or hemorrhages detected, which can be viewed in Figure 1. Additionally, doctors are also able to view the original fundus images without the AI system's visualizations, giving them the option of which images to use in screenings. By using the AI system, the department hopes to make the screening process more efficient by bringing down the time spent by ophthalmologists to complete a screening. Altogether, these three applications constitute the CDSS used by the ophthalmologists in their DR screening process.

3.2 Pre-study: Investigating the AI-supported DR Screening Workflow

To understand the role of the AI-supported CDSS in the DR screening practice of ophthalmologists at the hospital, we conducted a pre-study. The remainder of this section expands on the methods of the pre-study as well as its key findings.

3.2.1 Pre-study Methods. The pre-study firstly entailed individual interviews with seven ophthalmologists at the department, each lasting approximately one hour. In addition, three out of the seven ophthalmologists agreed to participate in a session of contextual inquiry, which resulted in a total of three sessions each lasting approximately four hours. Each session of contextual inquiry consisted of two of the authors sitting next to an ophthalmologist in their office whilst screening patients for DR. The ophthalmologist was asked to explain their process while going through it, while one of us was taking notes and the other was asking occasional questions. Each of the three sessions were audio-recorded and later transcribed. Likewise, all seven interviews were audio-recorded and transcribed for subsequent analysis. To analyze the transcriptions and contextual inquiry notes, we used Reflexive Thematic Analysis [5]. The outcome of the analysis were several prominent themes from the data, some of which are elaborated on in the following subsection [3].

3.2.2 Pre-study Findings. **AI-generated Labels Causing Anchoring Bias:** One of the most prominent findings coming from our pre-study was how the AI system's way of visually presenting its results affects the ophthalmologists while screening for DR. We found that whenever all fundus images taken of a patient's retina were labeled green, doctors would spend less time going through those images. This was both an observation made by the authors while conducting sessions of contextual inquiry but was also mentioned explicitly by several of the ophthalmologists during interviews. The tendency among doctors to spend less time reviewing green images compared to yellow or red ones may partly be due to them trusting in the AI system's capabilities to detect microaneurysms and hemorrhages. Therefore, their logic when seeing all green fundus images for a patient is that the patient does not have DR, which, according to the ophthalmologists, is true in most cases. However, the tendency may also be due to the connotations inherent in the green color, symbolizing to the doctor that the fundus images of that color are 'all fine'. This connotation is reinforced when put into the context of the other two colors, yellow and red, resembling the color scale of traffic lights and its respective color symbolism. The potential problem with applying this logic to green images is that many other abnormalities can develop in a patient's retina that are not captured by the AI system, and this means that other, potentially life-threatening, diseases can go unseen if green images are looked over too quickly. In fact, one ophthalmologist shared a personal experience with a patient's fundus images being labeled all green by the AI system, but after taking a closer look at the images she discovered that the patient had a tumor in their retina. [3].

Taking this finding into consideration, we believe that the color labels used by the AI system act as potential anchors in the ophthalmologists' decision-making process, which, in the worst case, can lead to severe diseases being overlooked. Other potential consequences of the color labels could be for the ophthalmologist to believe that a diabetic without DR has DR because of a yellow or red label, when in fact the AI system was wrong [3]. In either case, with their use of an AI-supported CDSS, doctors risk getting either false negative or false positive results from the AI system, which emphasizes the need to assess which bias mitigation strategies have the potential to be adopted by doctors into their screening workflow.

Differing Uses of and Attitudes Towards the AI System: Another finding that we want to highlight from our pre-study revolves around how the ophthalmologists would compare their assessments to that of the AI system, and whenever they matched the doctors would get a sense of comfort and security in their diagnosis. This was especially true when the ophthalmologist had not found any microaneurysms nor hemorrhages, and was met by solely green color labels [3]. Contrarily, other ophthalmologists that participated in our pre-study did not trust the outputs of the AI system, which was partly due to an error in the system. Specifically, the system frequently mistakes retinal pigment for DR-related abnormalities, resulting in false positive results. Ultimately, these doctors mostly disregarded the AI system's outputs and rarely used them.

Therefore, in the present study, we want to maintain the sense of security that some doctors experience when the AI system's output corroborates their own findings. In addition, we explore the doctors' reactions to when the opposite

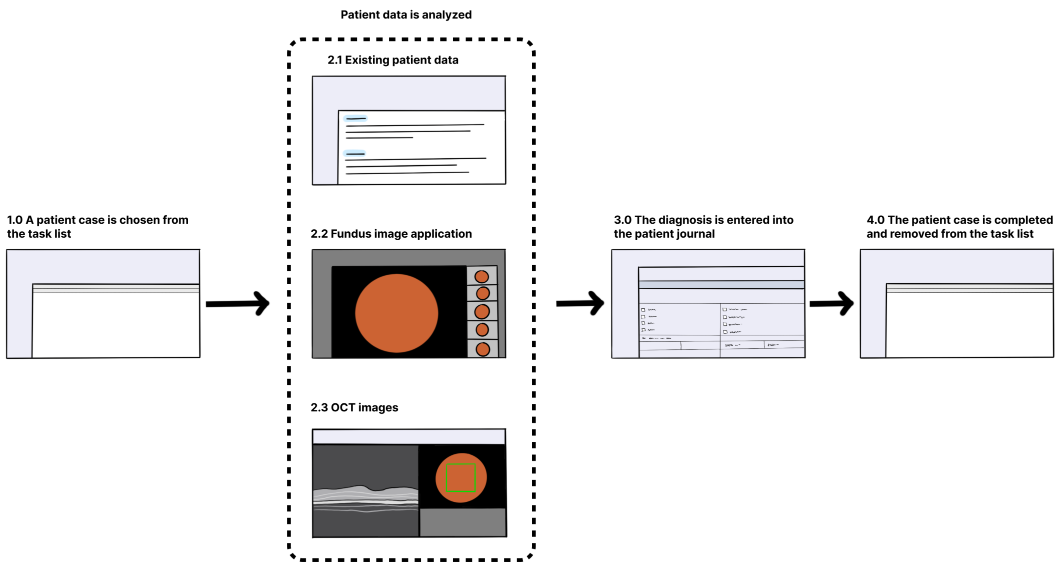


Fig. 2. A visualization of the current DR screening workflow of ophthalmologists at the department.

happens, namely when there is a mismatch between their assessment and the AI system's output. Furthermore, we acknowledge that some of the ophthalmologists' disregard for the AI system in general may remain and impact their perceptions towards the modifications we suggest to their workflow with our debiasing workflow.

The DR Screening Workflow: The final pre-study finding we want to highlight is the existing DR screening workflow. Earlier in this section, we explained how ophthalmologists use a CDSS consisting of three separate applications to support them in their DR screening process. In this subsection, we describe the general workflow of the doctors while screening a given patient for DR, which can be viewed in Figure 2. First, the doctor opens up the application for viewing incoming patient cases to see if any new screening tasks are available to them (1.0). Then, they go on to look at the patient data and images taken by nurses, which is viewed in the OCT image browser view, the fundus image application, and the patient journal (2.1-2.3). These applications are all used in an arbitrary sequence, since doctors have different preferences as to what to look at first. Then, they enter their diagnosis into the patient journal (3.0). Some of the doctors do this gradually as they gain new knowledge from patient data. Lastly, the doctor goes back to the task list to mark the screening of the patient case as completed (4.0).

Generally, the ophthalmologists view the task of screening patients for DR as relatively straightforward, and they are all confident that they can reach a precise diagnosis within the one hour that is available to them. Since a large proportion of the diabetes patients that get screened do not turn out to have DR, screenings often take much less than an hour for the doctors to complete. If a patient case is complicated, e.g. with many retinal abnormalities or a long history of health issues, the screening often takes longer to complete compared to cases with no abnormalities or previous health issues. Nevertheless, screening patients repetitively is a tedious task, and the ophthalmologists have other pressing responsibilities, making it important that the screening process can be completed as quickly as possible.

Therefore, since there is consensus among the ophthalmologists that their diagnoses are as accurate as possible, they are generally more concerned with increasing efficiency over accuracy when it comes to optimizing their workflow.

4 INTEGRATING BIAS MITIGATION STRATEGIES INTO THE CLINICAL DECISION SUPPORT SYSTEM

In this section, we present the interactive prototype used to communicate the debiasing workflow to the ophthalmologists during evaluations, into which the three bias mitigation strategies are implemented.

4.1 Bias Mitigation Strategies

Since high-risk professionals have a tendency to use System 1 thinking in decision-making, which makes them more vulnerable to cognitive biases such as anchoring bias, we use bias mitigation strategies in an effort to shift ophthalmologists into System 2 thinking with the aim of increasing diagnostic accuracy. In this study, we take departure in the three bias mitigation strategies of *hear the story first*, *decision justification*, and *consider the opposite*, which are described in detail in the subsections below.

4.1.1 Hear the Story First. As described earlier, anchoring bias has to do with the cognitive tendency to make estimates based on an initial value, resulting in the final estimate being a product of an adjustment towards this initial value [22]. A suggested approach to mitigating anchoring bias is therefore to look at the facts before being presented with others' diagnoses [22]. In other words, an approach could simply be to 'remove the anchor' or the initial value that often ends up biasing the final estimate. This notion is also reflected in Groopman's book [12] about the thought processes that underlie the decisions that doctors make on behalf of patients, where he conducted an interview with Dr. James Lock, a pediatric cardiologist who specializes in the diagnosis of heart conditions [25]. In this interview, when discussing the avoidance of bias in the diagnosis of patients, Dr. James Lock stated: "*When a case first arrives, I don't want to hear anyone else's diagnosis. I look at the primary data*" [12].

Although the effectiveness of this strategy has not, to our knowledge, been verified by other researchers in the past, we see potential in its ability to mitigate anchoring bias. Therefore, this is the first bias mitigation strategy that we have focused on in this study, which will be referred to as *hear the story first*.

4.1.2 Decision Justification. The next bias mitigation strategy is that of *decision justification*, which entails prompting the decision-maker to justify the reasoning behind their decision as a means to activate reflective thinking. Isler et al. [16] argues that guiding decision-makers explicitly through how they should reflect may improve their cognitive performance [16]. Doing so has shown to not only increase the quality of clinicians' diagnoses but also to function as a mitigator of cognitive biases, consequently improving diagnostic accuracy [30]. Recent research examining the effectiveness of bias mitigation strategies showed that *decision justification* was one of the most consistently successful strategies in improving diagnostic accuracy [16, 20].

Though *decision justification* is not targeted specifically towards anchoring bias, but cognitive biases in general, we believe that it may be a suitable means to shifting ophthalmologists into System 2 thinking, as it entails a mental walkthrough of the evidence that leads the decision-maker to their decision.

4.1.3 Consider the Opposite. Considering possible alternatives before committing to a decision is a bias mitigation strategy that has been proven effective at mitigating anchoring bias [2, 26]. The strategy, aptly named *consider the opposite*, has also previously been targeted towards the fallacy of overtly relying on heuristic knowledge and the overconfidence of the decision-maker in their decision [26]. To combat anchoring bias, *consider the opposite* prompts the decision-maker to consider information that contradicts their current beliefs, thus challenging their potential adjustment towards the anchor [26]. Simply presenting decision-makers with multiple alternatives to consider has shown to reduce cognitive biases in general [2]. Therefore, *consider the opposite* seems an appropriate strategy to apply to the workflow of the ophthalmologists, as it targets both anchoring bias as well as other cognitive biases [2, 26], and is easily transferred to computer-based platforms [2].

4.2 The Debiasing Workflow

As described in the previous section, we have established that the AI-generated color labels applied to each fundus image may act as anchors, introducing risks that can compromise doctors' decisions. In that regard, we found three mitigation strategies that show potential for diminishing anchoring bias. To assess how doctors perceive these strategies, we designed an interactive prototype that embodies all three, which can be viewed in Appendix A. The user interface (UI) design style of the prototype mimics that of the patient journal and the fundus image application to avoid diverting the doctors' focus towards aesthetics.

Fundamentally, there exists four possible outcomes of a screening when combining the diagnosis reached by the ophthalmologist with the output of the AI system:

(1) Neither the ophthalmologist nor the AI system find DR-related abnormalities
(2) Both the ophthalmologist and the AI system find DR-related abnormalities
(3) The ophthalmologist finds DR-related abnormalities while the AI system does not
(4) The AI system finds DR-related abnormalities while the ophthalmologist does not

Each of these four outcomes branch off into individual flows in the prototype, which can be viewed in Figure 3. For the 'system' to decide which flow to branch off into, the doctor's diagnosis is compared to the AI system's output after the diagnosis has been entered into the patient journal. This check of the two assessments against each other is only simulated by our prototype.

The starting point in the debiasing workflow is the same as in the current DR screening workflow; the ophthalmologists begin by opening up a patient case from the task list, which takes them to the respective patient journal. Then, they begin the process of looking at data that helps them form their decision, including OCT- and fundus images as well as other patient data in the patient journal. To avoid the creation of an anchor due to the color labels, we have removed all AI system output from the fundus image application to introduce the bias mitigation strategy *hear the story first*, which is viewed in Figure 3 (0.1). This means that in the fundus image application, the doctors are now only able to view the original fundus images that have not been processed by the AI system. Then, after having looked at all the relevant data, the doctors return to the patient journal (0.2) to fill in their diagnosis, which they submit by pressing the 'OK' button. This decision then triggers one of the four flows in Figure 3, depending on which of the four outcomes is the case.

If neither the ophthalmologist nor the AI system found DR-related abnormalities, the ophthalmologist is presented with a banner in the bottom of the screen stating that the AI system did not find any microaneurysms or hemorrhages (1.1). This is introduced to give the doctor a sense of security in their diagnosis, drawing on the pre-study finding of how they felt a sense of comfort whenever their assessment matched that of the AI system [3]. Contrarily, if both parties have discovered DR-related abnormalities, the ophthalmologist is presented with a window requiring them to justify their answer by writing it out in text (2.1). This is done to prompt the ophthalmologist to use System 2 thinking, sparking an analytical mindset, which is our implementation of the *decision justification* strategy. After writing their justification, they are then presented with the AI system's output to give them the opportunity to compare their

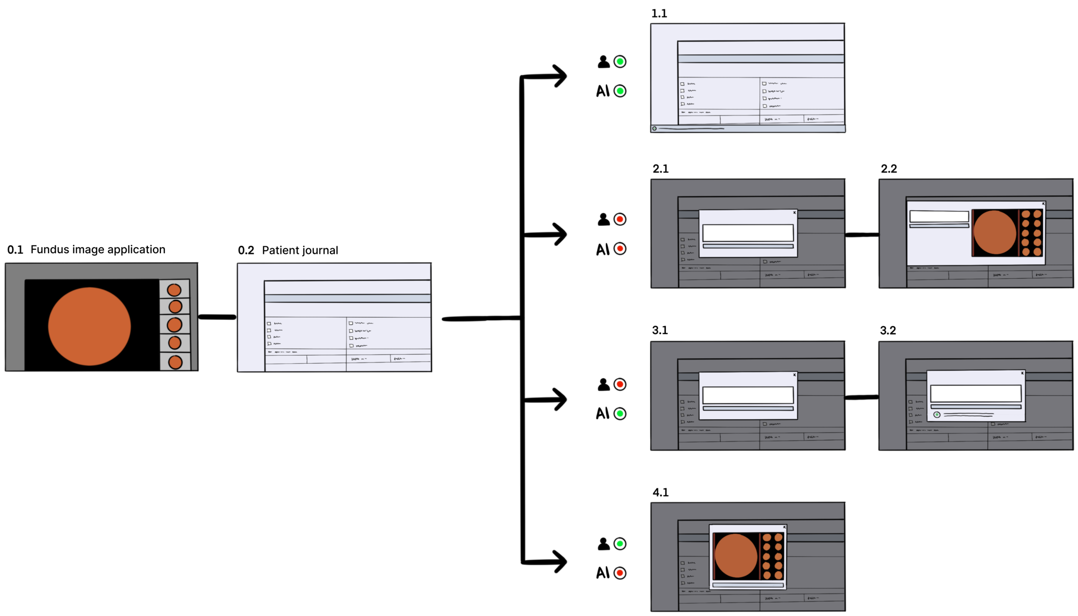


Fig. 3. A flowchart depicting the debiasing workflow, where a green dot symbolizes that no DR-related abnormalities were found and a red dot that DR-related abnormalities were found. Images of the prototype screens depicted in this figure can be viewed in Appendix A

diagnosis with it (2.2). Similarly, if the ophthalmologist finds DR-related abnormalities but the AI system does not, they are presented with the same window as in 2.1 (3.1). In this case, when the doctor has written their answer, they are then presented with a notice similar to the banner in 1.1, stating that the AI system found nothing, thereby using the strategy of *consider the opposite* (3.2). Lastly, if the ophthalmologist finds nothing but the AI system does, the doctor is presented with an opposing answer from the AI system once they have entered their diagnoses into the patient journal (4.1). This presentation of an opposite answer by the AI system, another use of the *consider the opposite* strategy, lets the ophthalmologist consider whether or not they have missed anything they wish to include in their diagnosis.

5 METHOD

In the ensuing subsections, we describe our process of participant recruitment and data gathering, followed by how the evaluation data was analyzed using Reflexive Thematic Analysis.

5.1 Data Gathering

To investigate doctors' perceptions towards the debiasing workflow, we conducted evaluations with the ophthalmologists that participated in our pre-study [3]. Through a liaison within the department, we were able to recruit six ophthalmologists to participate in our study, the details of whom can be viewed in Table 1. The study participants will henceforth be referred to individually using the letter 'P' followed by a number between 1 and 6, e.g. 'P3'. The participants were between the ages of 35 and 51 and had years of experience as specialists in ophthalmology ranging

Participant	Primary job title	Sex	Age	Years of experience
P1	Chief physician	F	47	14.5
P2	Senior registrar	M	44	4
P3	Chief physician and surgeon	F	51	16.5
P4	Chief physician and surgeon	M	40	12
P5	Specialist	M	38	4.5
P6	Senior registrar	M	35	1

Table 1. Details of the participating ophthalmologists.

from 1 to 16.5. Our data gathering process began by conducting a pilot evaluation with our liaison at the department prior to conducting evaluations with each of the recruited ophthalmologists. The purpose of the pilot evaluation was to become aware of any potential inconsistencies between the debiasing workflow and the domain of ophthalmology to ensure that those would not become the focus of the evaluations. A total of six evaluations were held, all of which were audio-recorded and transcribed. During the evaluations, participants were presented with the debiasing workflow prototype and each of the four different flows were explained in detail, after which questions pertaining to the specific flows were asked.

5.2 Data Analysis

Following the evaluations, all transcriptions were analyzed using Reflexive Thematic Analysis, which is a qualitative method for data analysis that is easily adapted to different contexts [5]. A core aspect of this method is that the researcher's own subjectivity is used as an advantage rather than something to avoid, as opposed to other analysis methods where the aim is to be as objective as possible. Thus, an integral part of Reflexive Thematic Analysis is for the data to be viewed from the perspective of the analyzing researcher wherein their own culture, social position, and academic background shape the interpretation of the data [5].

All three authors took part in analyzing the transcriptions. Initially, one was chosen to be analyzed individually by all. Then, we compared our resulting codes and themes, sharing our understandings to establish alignment moving forward. The remaining transcriptions were then divided among all three authors to be analyzed individually. When all six transcriptions had been analyzed, we went through the themes and codes, which had been conceptualized by each author individually, in plenum.

6 FINDINGS

In the following subsections, we present and elaborate on the findings resulting from evaluations of the debiasing workflow. As a result of Reflexive Thematic Analysis, the findings have been categorized into five themes. Specifically, they entail how doctors envision that the diagnostic accuracy may be affected by the debiasing workflow, how efficiency and the complexity levels of decision-making tasks affect doctors' perceptions of the workflow, how open doctors are to AI-generated insights, and the role bias currently plays in the ophthalmologists' daily practice.

6.1 Envisioned Effect of the Debiasing Workflow on Diagnostic Accuracy

A core theme identified in our data relates to diagnostic accuracy. Some ophthalmologists (P1, P2, P5) did not imagine that the debiasing workflow would heighten the accuracy of their diagnoses, as they generally did not think that the diagnoses could get any more accurate than they are now. Others (P3, P4, P6) thought the debiasing workflow would be a definite improvement in terms of diagnostic accuracy, but at the cost of efficiency. For instance, P6 stated: "*I think it would be really great to have periodically [...] to get back into it and to start thinking for yourself, and not fall asleep completely*", implying that screening for several hours at a time can exhaust their ability to stay focused. When asked to pretend that they could have as much time as they wanted to conduct a screening, and thereby disregard efficiency, P3 stated: "*If I had all the time in the world, this [the debiasing workflow] would be the optimal way to do it*".

In relation to the implementation of *hear the story first*, P3 found that removing the AI system's output from the fundus image application would likely increase the accuracy of diagnoses. Specifically, she stated: "*The bias has been removed, I arrive at my own conclusion, and then I compare my conclusion with the [AI system's] conclusion*". Furthermore, the majority of the ophthalmologists (P2, P3, P4, P6) found that *consider the opposite* would be the most helpful strategy in making sure they avoid missing important details. For instance, as emphasized in the following statement by P3: "*Sometimes you get tired, so if the program said 'Something [an abnormality] is there', and I scale up [the image] [...] and then I think 'Yes, I didn't notice that myself'*", later highlighting how the third and fourth flow (Figure 3), in which the strategy of *consider the opposite* is embodied, are the most useful ones. Though more skeptical, P4 stated, when referring to *consider the opposite*: "*It is absolutely a more useful solution. But it wouldn't make our work any faster, but from a quality perspective, then yes*". With regard to the strategy of *decision justification*, P6 theorized that by sharing one's justifications with each other, the diagnostic accuracy could be heightened collectively, as they would gain insight into each other's thought processes.

All ophthalmologists expressed how they thought the debiasing workflow could, with advantage, be used as a learning tool for inexperienced screeners to ultimately heighten their diagnostic accuracy. As inexperienced ophthalmologists can be more unsure of their decisions, the debiasing workflow could serve as a tool to become more skilled and achieve more certainty in their diagnostic reasoning. For instance, P2 commented that it could "*[...] act as a kind of safety measure*". In that regard, P3 stated: "*When you are training, you have a need for someone else to check what you are doing, 'Is this right or is it wrong?'. This is what the program does the way you have arranged it*", and P4 also noted: "*It would work well as a learning tool since you have to reflect on what you have found - but to those of us that are as obdurate as we are, I don't think it will change very much*".

6.2 The Restraining Role of Efficiency

Another important theme identified in our data has to do with the doctors' focus on efficiency in their work practice. For instance, P1 commented that "*in our everyday work, we have to focus on efficiency mostly, we have to be quick and view many images because the patient numbers keep growing. So we need a solution that helps us become quick*". Similarly, P3 added that "*efficiency is king today, and that is what has become problematic. Often you consciously compromise*", indicating that even though it may be at the expense of diagnostic accuracy, the doctors are under pressure to put efficiency first. Additionally, the same participant explained how the amount of tests performed on each patient has increased, resulting in more patient data to consider for each patient while the expectation of high efficiency remains. Therefore, none of the ophthalmologists were in favor of the alternate screening approach introduced with the debiasing workflow.

One of the aspects that provoked this standpoint was the implementation of *hear the story first*, in which the AI-generated color labels, as viewed in Figure 1, were removed from the fundus image application. Specifically, some of the ophthalmologists (P1, P3) commented on how this would mean that they would have to spend more time examining each fundus image themselves, as opposed to getting an instantaneous overview of all fundus images, as is provided by the color labels in the existing workflow. In this regard, P3 stated that "*[...] if I already trust what the AI system says* *and I have its answer from the start then I can shortcut looking through all the images. I can't do that if I get them [the processed fundus images] afterwards. Then I have to analyze every image as new*". In addition, P1 elaborated: "*There are very many patients that have nothing on their retinas*", referring to how she is currently able to quickly identify the cases with no DR as a result of the AI system labeling all of these patients' fundus images green.

Other ophthalmologists (P2, P4, P5, P6) were more indifferent to the removal of the color labels. This was mainly because they either did not use the AI system's output at all, or did not look at it until they had gone through the original images themselves. P2 saw potential in using *hear the story first* to avoid being biased at the beginning of screenings, and elaborated that he hopes his colleagues using the color labels do not take them too literally. When presented with the second and third flow (Figure 3), in which *decision justification* is implemented, all ophthalmologists once again expressed concern when it comes to efficiency. For instance, P4 thought it was counter-productive to provide a justification whenever he found any retinal abnormalities, and stated: "*[...] from an efficiency point of view - no thanks. It would just prolong the process completely*".

6.3 Effects of Decision Complexity Levels on Perceptions Towards the Debiasing Workflow

Our participants' general attitude towards the debiasing workflow seemed to be influenced by their view on the complexity level of the decision-making task. Since the purpose of the DR screening is solely to identify patients that need further diagnosing or treatment, the ophthalmologists agree that this task is straightforward compared to other decision-making scenarios that they encounter in their daily work. For instance, when explaining her approach to complex decision-making tasks, such as those pertaining to patients with several concurrent illnesses, P3 stated: "*It's a bit more artistic and you are way more free in what you think or what you say*". On the contrary, when describing the DR screening process, the same participant stated: "*This [the artistic approach] is not the case here, because this is done according to a form*". Here, P3 commented on how DR evolves through different stages, which have specific criteria that determine whether a patient is in one stage or the other. Therefore, P3 argues: "*I have no possibility to have my own opinion about it. So it [the diagnosis] is only based on whether I find all the things [criteria] that belong to the diagnosis*", later explaining how she would not know what to specify in the text input field asking her to justify her decision, as she believes the patient journal reflects the entirety of her answer. To this, P6 commented: "*I don't think anyone wants to write the same thing in two different places*".

6.4 Openness to AI-generated Insights

Our incorporation of the bias mitigation strategies seemed to spark reflections in the ophthalmologists as to how they view their own expert opinion compared to the output of the AI system. As stated by P3: "*When you're an expert, you have many years of experience but you also think you do it better than others. [...] And then it is up to me [as an expert] how much doubt I allow in myself - whether I accept other information that contradicts me*", commenting on how some experts are more open to input than others due to their belief that they outperform automated solutions. Most of the ophthalmologists (P2, P3, P4, P5, P6) brought up how they think the way the AI system presents its output in the debiasing workflow made it seem like a precautionary measure that steps in to fact-check their answer, which pertains specifically to the strategy of *consider the opposite*. This clashes with P3's belief that experts think they do it better than others - because why would someone who is the best at something need to be fact-checked by someone less competent? As P3 commented: "*At the moment, [the AI system] suggests 'this could be a hemorrhage' and then I go onto the image and see if that is correct - so in that way, I fact-check the program. In this way [the debiasing workflow], the program is the one fact-checking me*".

Some of the ophthalmologists (P2, P3, P6) found the prospect of being fact-checked by the AI system appealing and a good way to take advantage of its capabilities. In addition, P6 elaborated that he would use the AI system's output in the debiasing workflow to go "*back to see if what I found is just nonsense*" and that "*even though we might agree, there could still be anomalies that you would notice after [being shown the result of the AI system] that you have missed*". On the contrary, other participants (P4, P5) were more skeptical of the debiasing workflow due to their preexisting distrust in the AI system as a result of it mistaking harmless retinal pigments for DR-related abnormalities. However, had the AI system been more capable, meaning that it could detect more types of abnormalities and at a higher precision, they would be more open to using the debiasing workflow. Moreover, P5 stated that "*if the AI [system] was very skilled and functioning, then I could easily imagine it as some sort of mentor that would hit people over the hand and say, 'Let's go through 20 images, is anything there or not?'*". He elaborated: "*It does not work if it's the janitor that comes in and lectures the executive. That is kind of what is happening here [in the debiasing workflow]. It has to be another executive, capacity-wise, someone who is capable and who knows what can be done, that comes in and lectures the executive*", alluding to the fact that he does not recognize the AI system as an equal collaborator, or as someone with the capacity to correct him in his diagnosis. This point of view is also reflected by P5 in the following comment: "*[...] it is very rare that [the AI system] shows me something that will make me change my mind*". Contrarily, as is evident in the following statement, P3 thinks more highly of the AI system: "*I think this [decision justification] works well because there is a disagreement between two systems, where I believe in myself, but I also believe in the system, and then it is important to document why I disagree*".

6.5 The Current Role of Bias in the Ophthalmologists' Daily Practice

The ophthalmologists recalled bias as being a lesser part of their medical education, and it is not an aspect that any of them put a lot of thought into in their daily practice. When asked about the role that bias plays to the them, P1, P5, and P6 were under the impression that bias does not affect their work much. Others (P2, P3, P4) thought it played a bigger and more unavoidable role in their work. P3 explained how she, in some cases, sees bias as playing in her favor: "*If I want to be able to use 15 minutes looking at the patient, then I need to have a preliminary answer. That means that I want to have bias because it helps me with my own decision-making*". P2 explained how they work on the basis of a *working diagnosis*, which is the most probable diagnosis that the doctors believe to be the cause of a patient's symptoms, guiding the choice of diagnostic tests to confirm or deny the working diagnosis. As new information is discovered, the doctors adjust by choosing among other potential causes, or *differential diagnoses*, before ultimately arriving at a final diagnosis. To that, P3 also explained: "*[...] you have to create a working diagnosis: 'What do I think it could be?', 'How can I treat it?', 'What if it's wrong?', 'What kind of mistakes might I produce from the treatment?'*". P2 furthermore stated: "*You receive a diagnosis that you work from, and at this point, it is important to 'reset' and instead create your own diagnosis, as it [the symptoms] could be caused by something else. We have to provide evidence for the diagnoses we make*". Thus, in an attempt to avoid premature closure of cases, the ophthalmologists use these different categories of diagnoses as a strategy to ensure that different possible diagnoses are considered.

7 DISCUSSION

In the following section, we discuss our findings and relate them to research within the fields of clinical decision support, cognitive reasoning, and bias in decision-making. First, we elaborate on the degree to which our findings can be transferred back to each bias mitigation strategy, after which we explain how the doctors' self-perception influenced their views on using decision support. Then, we discuss the trade-off between efficiency and accuracy, as well as why the ophthalmologists, in some cases, saw their own cognitive bias as an advantage. Lastly, we consider the ambiguity of advances within AI support, and elaborate on why these may need to accompany parallel increased resources in the public health care system.

Our findings reveal that the ophthalmologists currently using the AI-generated color labels to gain an immediate overview were not in favor of the strategy *hear the story first*, while others were indifferent due to its resemblance to how they currently conduct screenings. All ophthalmologists expressed efficiency concerns when it came to using the *decision justification* strategy, however one ophthalmologist brought up how they believed sharing justifications with each other could potentially result in an increased diagnostic accuracy. Lastly, the ophthalmologists thought that *consider the opposite* would be the most helpful strategy to avoid missing important details, although it made them feel fact-checked by the AI system, a notion that divided the doctors in light of their differing perceptions of the AI system's capabilities.

7.1 Translating Bias Mitigation Strategies into Design

In this study, we implemented the three bias mitigation strategies of *hear the story first*, *decision justification*, and *consider the opposite* into the DR screening workflow of ophthalmologists employed at a Danish hospital. Specifically, we embodied them in an interactive prototype, which means that we went through a process of interpreting the written definitions of the strategies to translate them into a UI design. To our knowledge, no previous studies have focused on making this transition for these three bias mitigation strategies, in spite of how two of them are recurring in the bias mitigation literature [2, 16, 20, 26]. Thus, we are open to the fact that these bias mitigation strategies may be expressed in UI designs by other means, and that the ophthalmologists' statements regarding our implementation of the strategies may not be directly transferable back to the bias mitigation strategies as conveyed in literature. However, if we are to begin employing bias mitigation strategies into decision support systems, we need to begin investigating and suggesting concrete ways to do this in practice. Doing so enables us to discover clinicians' perceptions towards using bias mitigation strategies as a part of their CDSSs, which further allows for making adjustments to increase the chances of these CDSSs being adopted by clinicians into their work practice.

7.2 Self-perception Influencing Ophthalmologists' View on Decision Support

Both the AI system and the three bias mitigation strategies function as a form of decision support that assists the ophthalmologists in assessing patient cases and reaching diagnoses. Our findings show that the adoption of such decision support into the workflows of doctors is impacted to some extent by how the individual doctor views their role, skills, and fallibility in relation to the given task. As highlighted by Mussweiler et al. [26] as well as Lighthall and Vazquez-Guillamet [22], decision-makers can become overconfident, which can lead to overestimating one's experience and to underestimating uncertainty, which we tried to counteract using the strategy of *consider the opposite*. Interestingly, this strategy played the largest role in making doctors feel fact-checked, and underlined some doctors' overconfidence in themselves as they were questioning the usefulness of the strategy. However, we wonder whether some of this (over)confidence was warranted due to the AI system currently only being able to detect microaneurysms and hemorrhages. In our evaluations, this overconfidence was, for instance, reflected in the statement made by one participant regarding how you, as an expert, believe that you do it better than non-experts, and that the degree to which you accept and are open to information that contradicts you is up to the individual expert. This indicates that there is more to being an expert than just the years of experience, namely the way you view yourself in relation to others and the degree to which you are open to being wrong.

This becomes particularly interesting when considering the AI system as an entity that presents its 'opinion' on the patient data, potentially leading to doctors comparing themselves to the AI system as they would with other doctors. This is also reflected in the statement made by one of the ophthalmologists in which they compare themselves to an executive and the AI system to a janitor, saying that janitors should not try to lecture executives, indicating that they see themselves as being above the AI system and therefore should not accept its contribution. The doctors' comparison of themselves to the AI system may have been provoked by how the debiasing workflow branches off into four flows based on a comparison between their assessments, as viewed in Figure 3, as well as our use of *consider the opposite* where the AI system challenges their position. Specifically, as mentioned, this made several of the participating ophthalmologists state that they felt fact-checked by the AI system, as opposed to them fact-checking the AI system in their current workflow. Even though some doctors were skeptical towards the prospect of being fact-checked, we believe that confronting the doctors with alternative diagnoses, a concept that resembles their use of differential diagnoses, is necessary in shifting doctors from System 1 thinking into System 2 thinking.

However, we believe that steps could be taken towards creating a UI design where the strategy of *consider the opposite* makes doctors feel less fact-checked than was the case for our implementation of it. For instance, in the debiasing workflow, doctors are forced to consider the AI system's output before they can submit their diagnosis, creating a checkpoint to go through before doing so. Alternatively, we see potential in exploring more subtle or optional ways to introduce alternative diagnoses into the workflow, similar to what was done in the study by Tao et al. [33] where an AI-supported CDSS recommended 10 probable diagnoses, which ultimately had a positive effect on diagnostic accuracy.

Part of being open to the fact that you could be wrong includes accepting that you are influenced by different types of cognitive biases. As stated earlier, humans are predisposed to a wide variety of biases, and doctors are no exception [6]. Therefore, in order to adopt bias mitigation strategies into their screening workflow, we believe doctors must accept the fact that they are influenced by cognitive biases. Otherwise, it will naturally be difficult for them to see the point in spending additional time mitigating bias with the extra steps we have introduced in the debiasing workflow.

7.3 Prioritizing Efficiency over Accuracy

It quickly became clear to us that efficiency was of high priority to the department [3], and therefore also the ophthalmologists employed there. One participant referred to efficiency as being 'king', and indicated that they often have to neglect other matters to accommodate the overarching objective to be efficient. We believe that these other matters potentially include the accuracy of diagnoses, which can turn out to have severe consequences for patients. As explained by one of our participants, the number of tests performed on each patient has gone up, increasing the amount of test data to be interpreted by doctors and thereby the need to be more efficient. In addition, a decision was made in Denmark to screen all diabetes patients regularly to keep an eye on progressing DR symptoms in some patients, which has increased the pressure on the public health care system.

If the debiasing workflow was to be implemented and replaced with the way screenings are currently conducted, doctors would likely experience a drop in efficiency overall given the additional tasks that the bias mitigation strategies introduce. Our participants all agreed that this lowered efficiency is not desirable despite the potential reduction of bias that the use of bias mitigation strategies would induce. In other words, even though these strategies could potentially increase their overall diagnostic accuracy, efficiency takes precedence. Furthermore, as emphasized by Lighthall and Vazquez-Guillamet [22], the decisions made by high-risk professionals are often based on their subconscious ability to recognize patterns, which ultimately allows them to be more efficient. As this act of relying on one's subconscious draws parallels with the intuitive way of approaching decisions in System 1 thinking [6], which can ultimately oversimplify and compromise a decision, this prioritization of efficiency over accuracy seems questionable.

Our findings indicate that all three bias mitigation strategies are perceived by the ophthalmologists to be unrealistic in the context of screening for DR due to a lacking focus on efficiency. Therefore, based on the significance of efficiency in this context, we argue that the HCI research community should accept the reality in which medical practitioners work and the challenges they face surrounding efficiency. In extension, we believe that this study underlines the potential that lies in finding ways to implement bias mitigation strategies into AI-supported CDSSs while also considering the demand for efficiency.

7.4 Utilizing Bias as an Advantage

Throughout the course of this study, we have thought of bias as something that should be removed and that is undesired in the decision-making process. However, during the evaluation with one participant, it became apparent to us that doctors can have a more nuanced view of what bias may be. As explained by that participant, the ophthalmologists are also biased in the sense that they know of a given patient's health history, which could point them in wrong directions as to a diagnosis. However, being aware of factors such as which diseases a patient currently has or which treatments they have undergone can also help doctors find health issues that have arisen due to that history. Therefore, we wonder where the line is drawn between harmful and helpful bias. In our implementation of *hear the story first*, we removed the AI-generated color labels from the fundus image application in an attempt to eliminate the potential anchoring bias they could introduce. However, this caused some of the ophthalmologists to complain about how this change means that they would lose the instantaneous overview that the color labels provide. One could suspect that the doctors' positive view on their cognitive bias could be due to the fact that it allows them to use System 1 thinking, which Croskerry [6] characterizes as more low-effort than System 2 thinking. Thus, in long screening sessions, System 1 thinking may be a more comfortable and therefore more desirable state of mind to the doctors.

With this study, we attempt to move the ophthalmologists from System 1 thinking to System 2 thinking. However, using System 2 thinking for long periods of time can be tiring, as it requires a high mental effort, which can ultimately have a negative effect on diagnostic accuracy. Therefore, it seems that there is a need to shift between the two systems. One participant presented an interesting possible solution to this challenge, saying that the debiasing workflow could be used periodically. For instance, if used at different intervals of a given day, it could facilitate System 2 thinking temporarily, and otherwise allow the ophthalmologist to recuperate when not in use. While a solution as such would still impose the possibility of biased decision-making, as the ophthalmologist would use System 1 thinking while recuperating, it introduces a means to establish a balance between the benefits of System 2 thinking and the mental exhaustion that follows.

7.5 The Ambiguity of Advances in AI Support

When discussing the implementation of bias mitigation strategies with the ophthalmologists, several of them mentioned how they would be more likely to use the debiasing workflow if the AI system was more capable than it currently is. Specifically, they would want the AI system to be able to detect more kinds of abnormalities and for it to be less error-prone, which currently results in many false positive detections. An AI system capable of finding every possible retinal abnormality would lead to a higher screening efficiency, and could open up to screening even more people at an even higher pace than currently possible. This was also highlighted by Kapoor et al. as one of the clear advantages of pairing the image-centered nature of ophthalmology with the pattern recognition abilities of AI technology [17].

Though screening more people at a higher pace than currently possible may seem desirable, doing so would also mean that the number of patients in need of care would increase without the necessary resources in the public health care system to tackle the additional workload. In addition, as underlined by Chandakkar et al. [4], the involvement of ophthalmologists is still required despite technical advances in DR detection systems, emphasizing how technical advances do not obsolete the need for human involvement. Thus, we view the future advances within AI-supported CDSSs as a double-edged sword, and we wonder whether the use of AI to detect abnormalities should go hand in hand with the resources available to deal with what is being found. Otherwise, we could find ourselves in a position where we are unable to help patients in need of care.

8 LIMITATIONS & FUTURE WORK

To make the workflow changes brought about by the bias mitigation strategies feel more realistic to the ophthalmologists during the evaluations, we decided to merge all three of them into one single design. Looking back, this approach made it more difficult to evaluate the bias mitigation strategies individually, and most of our findings therefore pertain to general perceptions that do not revolve around one strategy in particular. This is the case, as most comments made by the doctors naturally regarded their overall impression and experience of the design as one workflow and not separate parts. Thus, in the future, researchers could incorporate just one strategy into a workflow in order to ascertain the perception of doctors towards that specific bias mitigation strategy. Also, we believe that introducing one bias mitigation strategy at a time could heighten the chances of the doctors accepting the introduction of bias-mitigating measures into their workflow, as opposed to introducing too many changes all at once, possibly resulting in rejection of the proposed alterations of the workflow altogether.

While the immediate next steps would be to understand more about the perceptions of ophthalmologists towards the incorporation of single bias mitigation strategies into their workflow, later steps could be to conduct clinical trials in which the new CDSS is tested to understand doctors' use of it in practice. Eventually, it would be interesting to see whether the inclusion of bias mitigation strategies into the doctors' workflow in fact decreases bias and thereby increases the accuracy of diagnoses.

In this study, including the pre-study, we have worked with the process of screening for DR, which is perceived by the ophthalmologists to be a relatively straightforward, though tedious, task. One participant explained how she views the outcome of a DR screening to be the direct result of a predetermined form with specific symptoms that correspond to specific stages of DR. Since the form makes reaching a diagnosis a rigid process that does not allow for the doctors to make their own interpretation of the patient data, one could argue that the risk of the doctors using System 1 thinking is ruled out, which seems to be the belief of this participant. On the other hand, the participant described how more complicated diagnostic scenarios call for a more 'artistic' mindset, which seems to entail System 1 thinking. We believe that this may have been a deciding factor in why the ophthalmologists did not see the need for bias mitigation in this particular case, given their perceived lack of System 1 thinking in the DR screenings. Therefore, we see this as a central limitation to our work, even though the use of System 1 thinking in these DR screenings cannot be ruled out entirely. In fact, even though the act of entering the diagnosis into the patient journal is viewed by the participant as a rigid process that rules out any System 1 thinking, the preceding analysis of fundus images can still lead to anchoring bias by virtue of the AI system's color labels, which in the end influences the diagnosis. In future inquiry, it would be interesting to see how the bias mitigation strategies of *hear the story first*, *decision justification*, and *consider the opposite* are perceived by doctors in diagnostic workflows of more complicated and high-risk patients to see how those differ from the perceptions highlighted in this study.

9 CONCLUSION

In this study, we investigated the perceptions of ophthalmologists towards the incorporation of bias mitigation strategies into their AI-supported CDSS workflow in screenings for DR in diabetes patients. Specifically, we carried out evaluations of an interactive prototype that embodies the strategies of *hear the story first*, *decision justification*, and *consider the opposite* with ophthalmologists employed at a Danish hospital. With this work, we firstly contribute with insights into the existing workflow of ophthalmologists in the context of screening diabetes patients for DR, one of several possible complications of diabetes. Secondly, we make concrete UI design suggestions for the implementation of bias mitigation strategies into a real-world AI-supported CDSS. Lastly, we provide an overview and assessment of the perceptions of ophthalmologists towards the application of bias mitigation strategies into their DR screening workflow. In order for us to take advantage of the bias-mitigating properties held by bias mitigation strategies, we must explore ways to employ them into both new and existing CDSSs, as well as seek to understand how they are perceived by medical practitioners. Given that the most common cognitive cause of diagnostic error is the failure to continue considering alternative diagnoses, we hope that in contributing to this line of research we contribute to an increased diagnostic accuracy to positively affect patient outcomes.

ACKNOWLEDGMENTS

We are grateful to the ophthalmologists who participated in our study for taking the time to provide us with insights about their screening practice and for helping us evaluate the debiasing workflow. We also thank our liaison at The Department of Ophthalmology for their role in facilitating our efforts at the hospital. Finally, we are thankful for the guidance and support provided by our supervisor, Niels van Berkel, over the course of this study.

REFERENCES

[1] Michael David Abràmoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C Folk, and Meindert Niemeijer. 2016. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative ophthalmology & visual science* 57, 13 (2016), 5200–5206.

[2] Bradley J Adame. 2016. Training in the mitigation of anchoring bias: A test of the consider-the-opposite strategy. *Learning and Motivation* 53 (2016), 36–48.

[3] Anne K P Bach, J Christian Brok, and Trine M Nørgaard. 2021. Human-AI Collaboration in Ophthalmology: Investigating the Integration of AI-based Decision Support for the Detection of Diabetic Retinopathy. (2021), 16.

[4] Parag S Chandakkar, Ragav Venkatesan, Baoxin Li, and Helen K Li. 2012. A machine-learning approach to retrieving diabetic retinopathy images. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*. Association for Computing Machinery, 588–589.

[5] Victoria Clarke and Virginia Braun. 2014. Thematic analysis. In *Encyclopedia of critical psychology*. Springer, 1947–1952.

[6] Pat Croskerry. 2013. From mindless to mindful practice—cognitive bias and clinical decision making. *N Engl J Med* 368, 26 (2013), 2445–2448.

[7] Diabetesforeningen. 2022. *Diabetes i Danmark. (Danish) [Diabetes in Denmark]*. Diabetesforeningen. https://diabetes.dk/forskning/viden-om-diabetes/diabetes-i-danmark

[8] Rishab Gargeya and Theodore Leng. 2017. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology* 124, 7 (2017), 962–969.

[9] Joey F George, Kevin Duffy, and Manju Ahuja. 2000. Countering the anchoring and adjustment bias with decision support systems. *Decision Support Systems* 29, 2 (2000), 195–206.

[10] Mark L Graber, Nancy Franklin, and Ruthanna Gordon. 2005. Diagnostic error in internal medicine. *Archives of internal medicine* 165, 13 (2005), 1493–1499.

[11] Ninel Gregori. 2021. *Diabetic Retinopathy: Causes, Symptoms, Treatment.* American Academy of Ophthalmology. https://www.aao.org/eye-health/diseases/what-is-diabetic-retinopathy

[12] Jerome E Groopman and Michael Prichard. 2007. *How doctors think*. Vol. 82. Houghton Mifflin Boston.

[13] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* 316, 22 (2016), 2402–2410.

[14] Ryuji Hamamoto, Kruthi Suvarna, Masayoshi Yamada, Kazuma Kobayashi, Norio Shinkai, Mototaka Miyake, Masamichi Takahashi, Shunichi Jinnai, Ryo Shimoyama, Akira Sakai, et al. 2020. Application of artificial intelligence technology in oncology: Towards the establishment of precision medicine. *Cancers* 12, 12 (2020), 3532.

[15] National Eye Institute. 2022. *Diabetic Retinopathy*. National Eye Institute. https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/diabetic-retinopathy

[16] Ozan Isler, Onurcan Yilmaz, and Burak Dogruyol. 2020. Activating reflective thinking with decision justification and debiasing training. *Judgment and Decision Making* 15, 6 (2020), 926–938.

[17] Rahul Kapoor, Stephen P Walters, and Lama A Al-Aswad. 2018. The current state of artificial intelligence in ophthalmology. *Survey of ophthalmology* 64, 2 (2018), 233–240.

[18] Gary Klein. 2008. Naturalistic decision making. *Human factors* 50, 3 (2008), 456–460.

[19] Clemens Scott Kruse and Nolan Ehrbar. 2020. Effects of computerized decision support systems on practitioner performance and patient outcomes: systematic review. *JMIR medical informatics* 8, 8 (2020), e17283.

[20] Kathryn Ann Lambe, Gary O'Reilly, Brendan D Kelly, and Sarah Curristan. 2016. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ quality & safety* 25, 10 (2016), 808–820.

[21] Yu-Hao Lee, Norah E Dunbar, Claude H Miller, Brianna L Lane, Matthew L Jensen, Elena Bessarabova, Judee K Burgoon, Bradley J Adame, Joseph J Valacich, Elissa A Adame, et al. 2016. Training anchoring and representativeness bias mitigation through a digital game. *Simulation & Gaming* 47, 6 (2016), 751–779.

[22] Geoffrey K Lighthall and Cristina Vazquez-Guillamet. 2015. Understanding decision making in critical care. *Clinical medicine & research* 13, 3-4 (2015), 156–168.

[23] Wei Lu, Yan Tong, Yue Yu, Yiqiao Xing, Changzheng Chen, and Yin Shen. 2018. Applications of artificial intelligence in ophthalmology: general overview. *Journal of ophthalmology* 2018 (2018).

[24] Henrik Madsen. 2013. *Diabetisk øjensygdom (Retinopati). (Danish) [Diabetic eyedisease (Retinopathy)]*. Netdoktor. https://netdoktor.dk/sygdomme/fakta/diabetesoejne.htm

[25] MD. 2022. *James E. Lock, MD*. MD. https://www.md.com/doctor/james-e-lock-md

[26] Thomas Mussweiler, Fritz Strack, and Tim Pfeiffer. 2000. Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin* 26, 9 (2000), 1142–1150.

[27] Quang H Nguyen, Ramasamy Muthuraman, Laxman Singh, Gopa Sen, Anh Cuong Tran, Binh P Nguyen, and Matthew Chua. 2020. Diabetic retinopathy detection using deep learning. In *Proceedings of the 4th International Conference on Machine Learning and Soft Computing*. Association for Computing Machinery, 103–107.

[28] Vimla L Patel, José F Arocha, and Jiajie Zhang. 2005. Thinking and reasoning in medicine. *The Cambridge handbook of thinking and reasoning* 14 (2005), 727–750.

[29] Daniel J Power, Dale Cyphert, and Roberta M Roth. 2019. Analytics, bias, and evidence: the quest for rational decision making. *Journal of Decision Systems* 28, 2 (2019), 120–137.

[30] Henk G Schmidt, Sílvia Mamede, Kees Van Den Berge, Tamara Van Gog, Jan LCM Van Saase, and Remy MJP Rikers. 2014. Exposure to media information about a disease can cause doctors to misdiagnose similar-looking clinical cases. *Academic Medicine* 89, 2 (2014), 285–291.

[31] Linda J Skitka, Kathleen L Mosier, and Mark Burdick. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies* 51, 5 (1999), 991–1006.

[32] Jacob Solomon. 2014. Customization bias in decision support systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. Association for Computing Machinery, 3065–3074.

[33] Liyuan Tao, Chen Zhang, Lin Zeng, Shengrong Zhu, Nan Li, Wei Li, Hua Zhang, Yiming Zhao, Siyan Zhan, and Hong Ji. 2020. Accuracy and effects of clinical decision support systems integrated with BMJ best practice–aided diagnosis: interrupted time series study. *JMIR medical informatics* 8, 1 (2020), e16912.

[34] Daniel Shu Wei Ting, Carol Yim-Lui Cheung, Gilbert Lim, Gavin Siew Wei Tan, Nguyen D Quang, Alfred Gan, Haslina Hamzah, Renata Garcia-Franco, Ian Yew San Yeo, Shu Yen Lee, et al. 2017. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *Jama* 318, 22 (2017), 2211–2223.

[35] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–18.

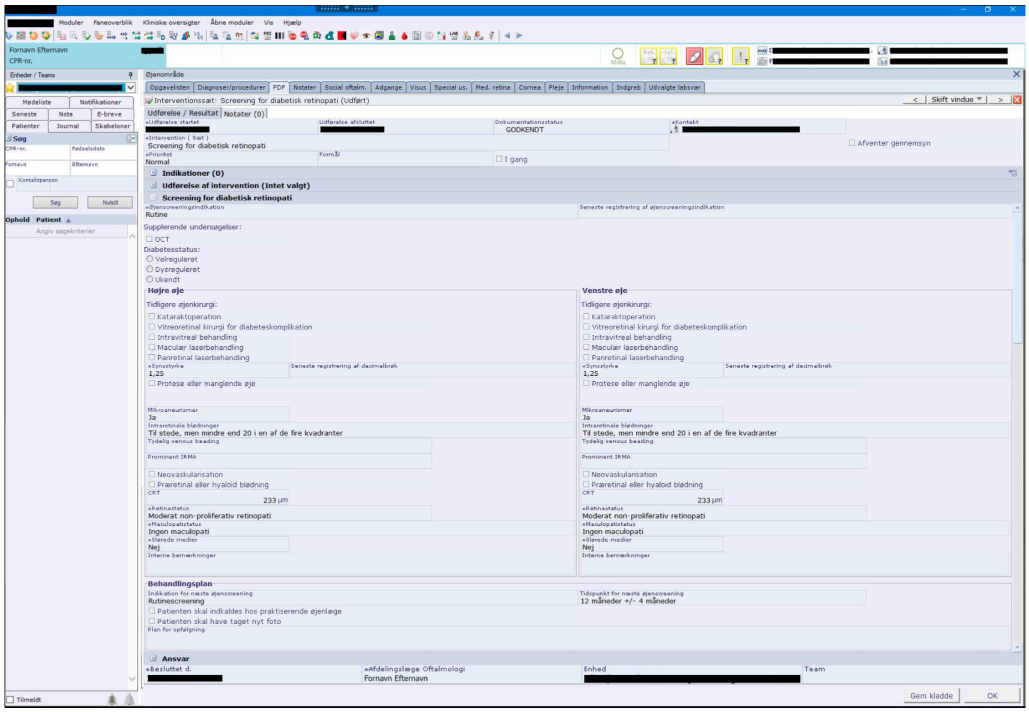
[36] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. Association for Computing Machinery, 1–15.

[37] Yunfeng Zhang, Rachel KE Bellamy, and Wendy A Kellogg. 2015. Designing information for remediating cognitive biases in decision-making. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. Association for Computing Machinery, 2211–2220.

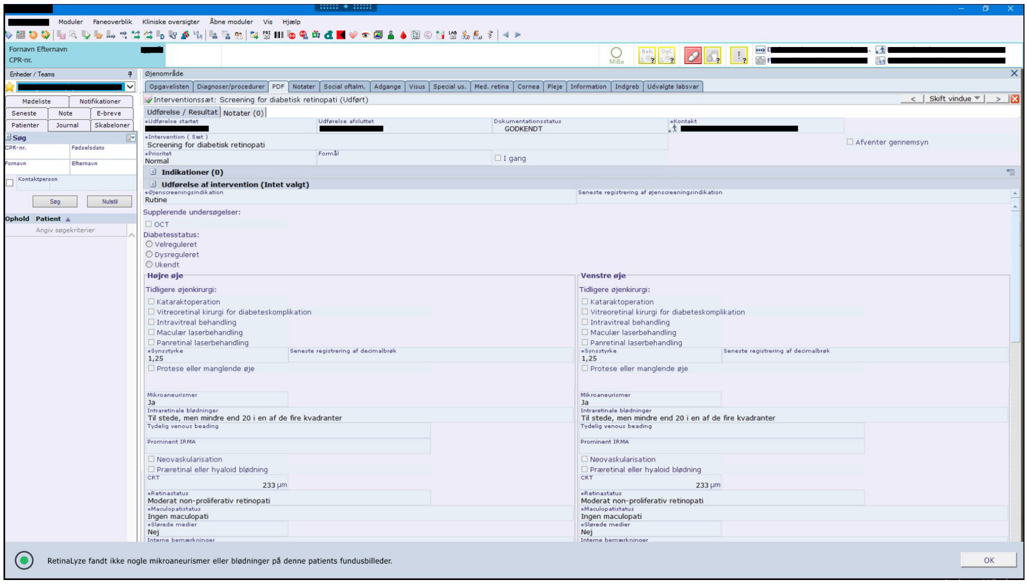
A THE DEBIASING WORKFLOW: UI DESIGN

The numbers referred to in this appendix correspond to the numbers in the flowchart depicted in Figure 3.

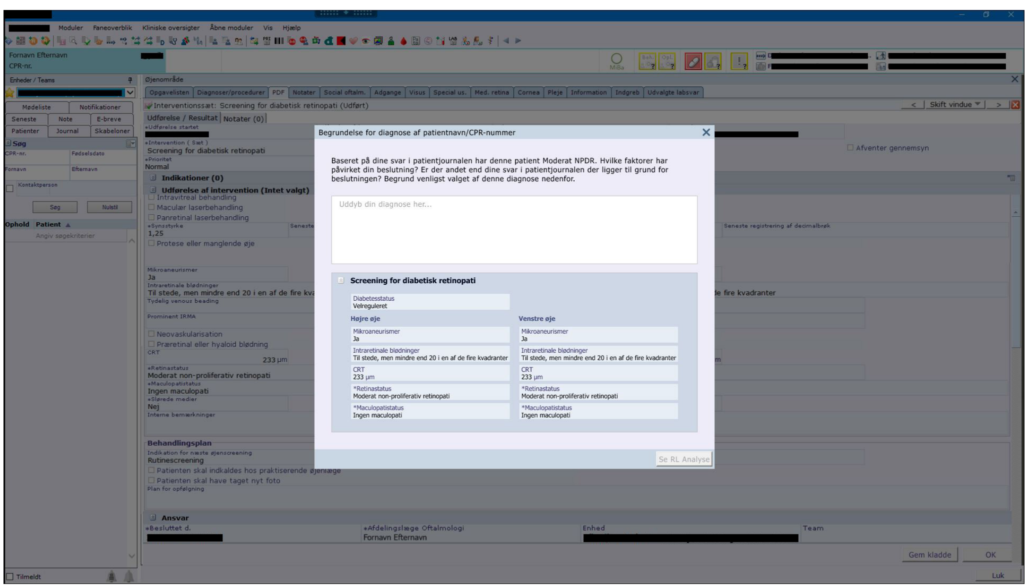
PATIENT JOURNAL 0.2



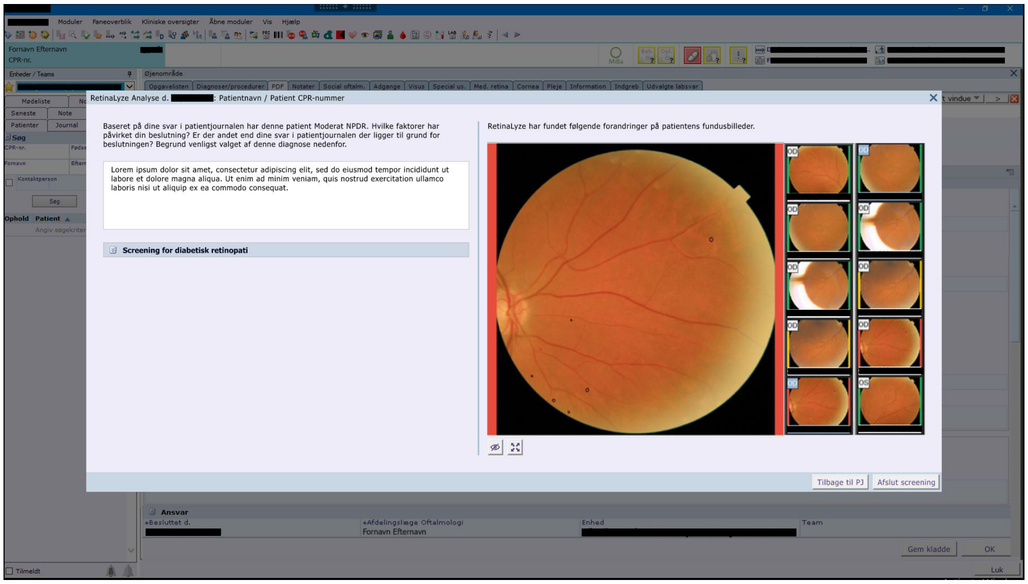
FLOW 1.1



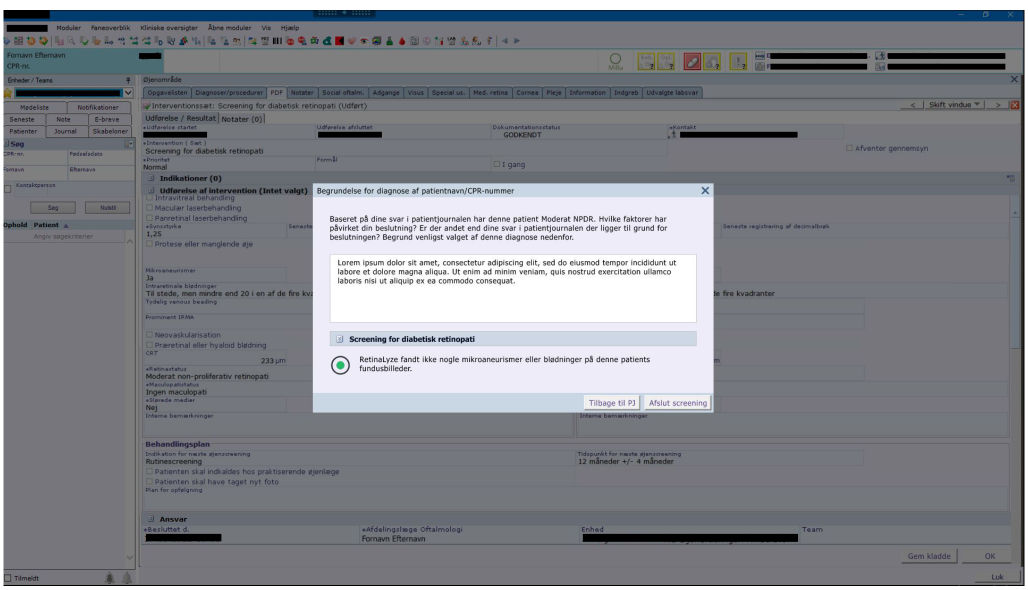
FLOW 2.1 & 3.1



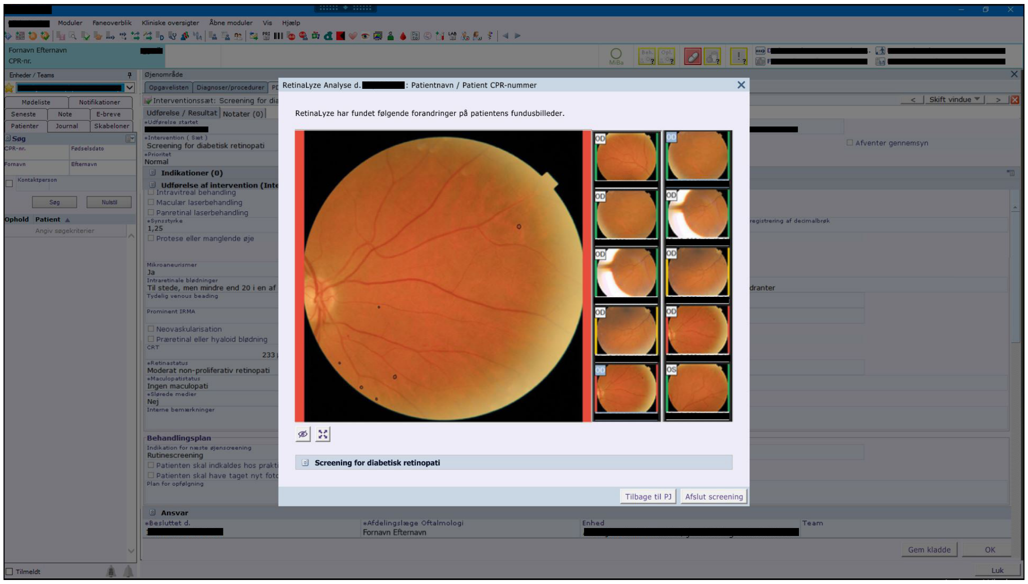
FLOW 2.2



FLOW 3.2



FLOW 4.1


RetinaLyze Analyse d. [redacted] : Patientnavn / Patient CPR-nummer
RetinaLyze har fundet følgende forandringer på patientens fundusbilleder.
Screening for diabetisk retinopati
Tilbage til PJ
Afslut screening