

Designing for Trust: A Multi-Factor Investigation of Optometrists' Perspectives on AI-Based Glaucoma
Screening Systems

by

Ali Karim

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Applied Science

in

Systems Design Engineering

Waterloo, Ontario, Canada, 2025

© Ali Karim 2025

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This study builds upon the doctoral thesis of Faisal Ghaffar, a member of the Advanced Interface Design Lab (AIDL). The mock interface design, interview scripting, and data collection were his work in collaboration with other students. A portion of the transcripts from his study was incorporated into this thesis. For further details regarding the students who contributed to his thesis, please refer to his dissertation.

I would like to acknowledge the contributions of Wendy Xu, who served as a Research Assistant in the Advanced Interface Design Lab (AIDL). She collaborated as an independent coder, coding all transcripts and supporting the thematic analysis in a collegial and constructive manner.

Finally, I am grateful to Rachel Woima, a PhD student in AIDL, for her thoughtful review and feedback on the first chapter of this thesis.

Abstract

Introduction: Although glaucoma screening AI models show strong performance, their integration into clinical practice remains limited. Clinicians often face barriers rooted in technological acceptance, with trust emerging as a key determinant of adoption. Prior research has emphasized explainability, but a broader exploration of factors affecting trust is needed. This study investigates multiple factors shaping trust in AI and translates them into design requirements for next-generation glaucoma screening clinical decision support systems (CDSS).

Method: In a previous study, two real-world glaucoma patient cases, each comprising three visits at different times, were presented under both unimodal conditions (fundus images only) and multimodal conditions (fundus images, optical coherence tomography, visual fields, and medical history) through a mock interface simulating an AI-based glaucoma screening support system. During these simulated visits, nineteen licensed optometrists interacted with the system and participated in follow-up interviews, where they were asked whether they trusted the system and to explain their reasoning. The objective of this thesis is to identify the factors influencing optometrists' trust in an AI-powered glaucoma screening tool and to propose design recommendations that can enhance trust in future iterations. The interview data were analyzed using Braun and Clarke's thematic analysis approach.

Result and Discussion: The emerging themes indicate that trust in the AI system is shaped by multiple factors: **(1)** alignment with clinicians' expectations of AI's role: flagging tool vs. consultant; **(2)** completeness of information; **(3)** communications of performance metrics: accuracy, sensitivity, confidence scores, perceived consistency and perceived quality of training data **(4)** clinical relevance of outputs (trends, actionable recommendations, differential diagnosis); **(5)** transparency in risk factor weighting, exclusions, and considered variables; **(6)** decision alignment between optometrists and the AI, assessed across decision inputs, identified risk factors, their relative importance, recommended actions, and the gradient of concordance in final decisions; **(7)** optimized the AI for cautious screening to captures all potential cases; **(8)** interface usability supporting timely decisions; **(9)** users' self-perceived expertise, occasionally leading to overreliance; **(10)** onboarding and training that highlighted the system's features and limitations; and **(11)** increasing familiarity over time, which helped calibrate trust. Based on these findings, 17 design principles were proposed to guide the development of the next iteration of a trust-supportive interface for glaucoma screening decision support systems.

Acknowledgements

My deepest gratitude goes to my supervisor, whose support, trust, and encouragement made this journey possible. Her guidance has been instrumental not only in shaping this thesis but also in fostering my growth as a researcher and as a person.

I would also like to thank Faisal Ghaffar for generously sharing the data from his study, without which this work would not have been possible.

My sincere gratitude also goes to Dr. Siby Samuel and Dr. Samantha Meyer for serving on my thesis committee and generously dedicating their time to reviewing my work.

This thesis also holds a personal dimension for me; it serves as a diary that captures the ups and downs of life and the memories intertwined with each chapter. I want to thank all who stood by me whether their presence was fleeting or lasting, your presence, in any measure, meant more than you know.

Most importantly, I would like to express my sincere gratitude to my family, whose love and faith in me have been my greatest source of motivation, even from miles away.

Finally, I acknowledge the use of generative artificial intelligence (AI) tools in preparing this thesis, which assisted with proofreading, improving grammar and flow, and creating graphs.

Dedication

To Mom, Dad and my beloved sister and brother, whose unconditional love and constant encouragement made this journey possible.

Table of Contents

Author's Declaration	ii
Statement of Contributions.....	iii
Abstract	iv
Acknowledgements	v
Dedication	vi
List of Figures	xii
List of Tables.....	xiii
List of Abbreviations.....	xiv
Chapter 1 Introduction and Literature Review.....	1
1.1 Overview	1
1.2 Background	1
1.2.1 Clinical Decision Support System.....	1
1.2.2 Benefits of Using CDSS in Practice	2
1.2.3 Pitfalls of Using CDSS in Practice	4
1.2.4 CDSS in Ophthalmology and Glaucoma Detection	6
1.2.5 Glaucoma.....	7
1.2.6 Glaucoma Detection and Related AI-Based CDSS	8
1.3 Problem Statement and Motivation	8
1.4 Literature Review	9
1.4.1 Understanding Trust in CDSS	9
1.4.2 Factors Influencing Trust in CDSS	10
1.4.3 Glaucoma-specific Trust Factors.....	17
1.4.4 Identified Gaps	19

1.4.5 Research Objectives	19
1.4.6 Research Questions	19
1.5 Thesis Outline.....	19
Chapter 2 Methodology.....	21
2.1 Overview	21
2.2 Study Procedure	21
2.3 Mock Interface Design Description.....	24
2.4 Ethical Considerations.....	28
2.5 Participants	29
2.6 Data Collection Methodology	29
2.7 Data Analysis Methodology	30
2.7.1 Thematic Analysis	30
2.7.2 Decision Ladder Framework	32
Chapter 3 Results.....	34
3.1 Overview	34
3.2 Participant Demographics	34
3.3 Qualitative results.....	37
3.3.1 Expected AI Usage	38
3.3.2 More Data.....	41
3.3.3 Decision Alignment.....	43
3.3.4 Explainability and Transparency	47
3.3.5 Model Performance	51
3.3.6 Familiarity Through Continued Use.....	54
3.3.7 Clinical Relevance.....	55

3.3.8 Decision Philosophy	57
3.3.9 Self-Perceived Expertise	57
3.3.10 Training	58
3.3.11 Usability	58
Chapter 4 Discussion.....	59
4.1 Overview	59
4.2 Interpreting Key Factors Influencing Trust (RQ1).....	59
4.2.1 Expected AI Usage (Role Framing & Workflow Fit)	59
4.2.2 More Data (Relative Information Completeness).....	60
4.2.3 Decision Alignment (and the Concordance Gradient).....	61
4.2.4 Explainability & Transparency.....	61
4.2.5 Model Performance	62
4.2.6 Familiarity Through Continued Use.....	63
4.2.7 Clinical Relevance.....	63
4.2.8 Decision Philosophy (Screening vs Diagnostic)	64
4.2.9 Self-Perceived Expertise and Overreliance Risk.....	64
4.2.10 Training	65
4.2.11 Usability	65
4.3 How Performance Evaluations Translate to Trust (RQ2)	65
4.3.1 Decision Ladder Analysis.....	66
4.4 Design Principles (RQ3).....	69
4.4.1 AI diagnosis Validation and Override Mechanisms	69
4.4.2 Cautious Flagging Strategy	69
4.4.3 Comprehensive Review of Test Results.....	69

4.4.4 Integration of Multimodal Data into Diagnosis.....	69
4.4.5 Per-Visit Presentation of Model Performance.....	70
4.4.6 Presentation of Model Performance during Onboarding.....	70
4.4.7 Binary Diagnostic Output.....	71
4.4.8 Guideline-Based Recommendations for Next Course of Action.....	71
4.4.9 Graphical Visualization of Quantifiable Test Results	71
4.4.10 Differential Diagnosis with Comparative Explanations	71
4.4.11 Model Output Transparency: Risk Factors, Data Gaps, and Actions.....	72
4.4.12 Sensitivity-Oriented Design for Screening Systems.....	72
4.4.13 Communicating AI Limitations and Safeguards	72
4.4.14 Structured Onboarding and Educational Support.....	73
4.4.15 Parallel Trial Use in Low-Pressure Settings as a Trust-Building Strategy.....	73
4.4.16 Concordance Gradient Computation and Caution Cues.....	73
4.4.17 Structured and Concise Information Presentation	73
4.5 Strengths and Limitations.....	74
4.5.1 Study Strengths and Key Contributions:	74
4.5.2 Limitations.....	75
4.6 Future Work	76
References	78
Appendix A Participants’ Demographic Related to Experience	91
Appendix B Information Letter 1	92
Appendix C Information Letter 2	94
Appendix D Study Brochure	96
Appendix E Demographic Questionnaire.....	97

Appendix F Feedback Letter	98
Appendix G Consent Form.....	99
Appendix H Semi-structured Interview Script	100

List of Figures

Figure 1- Study Design.....	23
Figure 2-Unimodal System, Fundus Tab.....	25
Figure 3- Unimodal System, Other Tabs Except Fundus	26
Figure 4- Multimodal System, Visual Field Tab.....	27
Figure 5- Multimodal vs Unimodal Cross Comparison over Tabs.....	28
Figure 6- Decision Ladder Template.....	33
Figure 7- Participants' Ethnicity Distribution	35
Figure 8- Participants' Age Distribution	35
Figure 9- Participants' Gender Distribution	36
Figure 10- - Participants' Distribution of Average Number of Glaucoma Patients Seen per Week	37
Figure 11- Participants' Distribution of Average Number of Patients Seen per Week	37
Figure 12- Participants' Distribution of Years of Experience	37

List of Tables

Table 1. Trust Factor Emerging Themes Derived from Literature Review.....	12
Table 2- Data features collected for glaucoma diagnosis.....	22
Table 3- AI Diagnostic Changes	24
Table 4- Participants' Demographic Distribution	35
Table 5-Participants' Clinical Experience Metrics	36
Table 6-Emerging Themes Overview-ordered by frequency	38
Table 7-Expected AI Usage Approach-Emerging Theme.....	39
Table 8- Variation in Trust Toward Unimodal AI Systems	42
Table 9- Decision Alignment-Emerging Theme	43
Table 10- Diagnostic Categories	44
Table 11- Degrees of Disagreement	44
Table 12-Explainability & Transparency- Emerging Theme	47
Table 13- Model Performance -Emerging Theme.....	51
Table 14- Clinical Relevance -Emerging Theme	55

List of Abbreviations

- AI – Artificial Intelligence
- AI-CDSS – Artificial Intelligence-based Clinical Decision Support System
- ANN – Artificial Neural Network
- AIDL – Advanced Interface Design Lab
- CDSS – Clinical Decision Support System
- CNN – Convolutional Neural Network
- EHR – Electronic Health Record
- GPA – Glaucoma Progression Analysis
- ICU – Intensive Care Unit
- ML – Machine Learning
- MSO – Microsoft Office (software version label, appears in methodology section)
- OCT – Optical Coherence Tomography
- OD / OS – Oculus Dexter (Right Eye) / Oculus Sinister (Left Eye)
- PHR – Personal Health Record
- REB – Research Ethics Board
- RNFL – Retinal Nerve Fiber Layer
- RQ – Research Question
- SHAP – SHapley Additive exPlanations
- SP-LIME – Submodular Pick for Local Interpretable Model-Agnostic Explanations
- TAM – Technology Acceptance Model
- UTAUT – Unified Theory of Acceptance and Use of Technology
- UW – University of Waterloo
- FG – Faisal Ghaffar (researcher)
- AK – Ali Karim (researcher)
- WX – Wendy Xu (coder / researcher)
- YZI – Yousuf Zia Islam (interviewer)

Chapter 1

Introduction and Literature Review

1.1 Overview

Chapter 1 of this thesis reviews the existing literature on trust drivers in AI-based Clinical Decision Support Systems (AI-CDSS), with a particular focus on glaucoma screening, and highlights the research gap that forms the basis of this thesis project. Section 1.2 introduces clinical decision-support systems (CDSS), distinguishes rule-based from AI-driven approaches, and reviews both their benefits and real-world pitfalls before narrowing to glaucoma as a pressing public-health concern and a target for AI screening. Section 1.2 then frames the central problem: despite impressive model performance, glaucoma AI-CDSS remain sparsely adopted, largely because clinicians do not yet trust them. Section 1.4 surveys the literature to unpack how “trust” is defined, catalogues ten recurring trust determinants, and highlights factors unique to glaucoma tools. This review reveals that existing studies focus heavily on explainability, seldom involve Canadian optometrists, and overlook other critical trust drivers, which leads into Section 1.4.4’s statement of the evidence gap. The chapter closes by articulating the study’s objectives (Section 1.4.5) and three guiding research questions (Section 1.4.6), providing a clear springboard for the methods that follow.

1.2 Background

1.2.1 Clinical Decision Support System

A clinical decision support system (CDSS) is an electronic tool designed to aid clinical decision-making by analyzing individual patient characteristics and generating tailored recommendations for clinicians to consider [1].

CDSS are typically categorized into knowledge-based and non-knowledge-based types. Knowledge-based CDSS rely on predefined rules (such as IF-THEN statements), which use various forms of evidence, like literature, clinical practice, or patient-specific data, to generate actions or outputs based on retrieved information [2], [3]. In contrast, non-knowledge-based systems also use data but make decisions through artificial intelligence (AI), machine learning (ML), or pattern recognition, rather than explicit medical rules [2].

Artificial intelligence (AI) is broadly described as the capability of machines or computers to exhibit intelligent behavior, that is, to perform tasks that typically require human cognitive abilities

[4]. Machine learning (ML) is a subset within the broader field of artificial intelligence (AI) [5]. While numerous definitions of machine learning exist in the literature, Arthur Samuel, one of the field's pioneers [5], offered a succinct explanation, describing it as the process by which computers acquire the ability to perform specific tasks without being explicitly programmed for them [6].

1.2.2 Benefits of Using CDSS in Practice

Overall, CDSS offer several advantages, including enhanced efficiency and quality of healthcare delivery, improved access to medical information, strengthened communication, and potential cost reductions [7].

According to Sutton et al [8], In practice, CDSS provide benefits across multiple domains, including enhanced patient safety, more effective clinical management, cost containment, streamlined administrative processes, diagnostic support, and patient-oriented decision support, which will be elaborated in the following sections.

1.2.2.1 Patient Safety

CDSS can help reduce clinical errors and their consequences. An example of this error is drug–drug interactions, which is a common error that could be prevented by CDSS used in the medication prescribing process, where it checks the dosing, duplication of therapies, and drug–drug interactions [9], [10].

Reminder systems for medical events, like blood glucose measurement in the ICU, are among numerous examples where CDSS helps prevent missed tests, which leads to a decrease in hypoglycemia events, thereby improving patient safety [11].

1.2.2.2 Clinical Management

Studies have shown that CDSS can enhance adherence to clinical guidelines [12]. This is important, as traditional guidelines and care pathways are often challenging to apply in real-world settings due to low clinician compliance [13]. The expectation that practitioners will consistently read, absorb, and follow new guidelines has proven unrealistic [14].

1.2.2.3 Cost Containment

CDSS can help healthcare systems reduce costs by shortening inpatient stays, recommending lower-cost medications that are covered by insurance [15], minimizing unnecessary repeat testing, and improving test ordering practice, without increasing patient mortality or hospital stay duration [16].

1.2.2.4 Administrative Automation

CDSS provides support for clinical and diagnostic coding, ordering of procedures and tests, and patient triage. CDSS can directly improve quality of clinical documentation as documentation accuracy is important because it can directly aid clinical protocols [14].

1.2.2.5 Diagnostics Support

Knowledge-based imaging CDSS are commonly applied during the image ordering process, supporting radiologists by recommending the most suitable tests, offering reminders about best practice guidelines, and flagging potential issues such as contraindications to contrast agents [17].

As the volume of medical imaging data increases, interpreting it remains a largely manual and time-intensive task. To overcome this challenge, healthcare professionals require tools that can streamline the extraction, visualization, and interpretation of such data [18]. AI technologies, particularly those leveraging advanced techniques like deep learning for pixel recognition and image classification, are demonstrating the ability to reveal insights beyond human capability [19], [20]. For instance, Google's team developed a deep convolutional neural network (CNN) to identify diabetic retinopathy, achieving performance comparable to that of US board-certified ophthalmologists [21]. They also created a CNN to detect arrhythmias on electrocardiograms, which surpassed the average cardiologist in accuracy, based on F1 score and sensitivity at matched specificity across all rhythm categories [22].

1.2.2.6 Patient-facing Decision Support

With the introduction of Personal Health Records (PHRs), Clinical Decision Support Systems (CDSS) features are now being integrated much like in Electronic Health Records (EHRs), but with patients as primary users and managers of their own data. This shift marks significant progress toward patient-centered care. CDSS-enabled PHRs are particularly well-suited for supporting shared decision-making between patients and providers, as they help eliminate information gaps that might otherwise hinder patient involvement in their healthcare decisions [23].

1.2.3 Pitfalls of Using CDSS in Practice

Although machine learning has been explored extensively in healthcare through numerous proof-of-concept studies, and has demonstrated many benefits, as described in the previous section [18], its translation into clinical practice remains minimal, with real-world integration into clinical workflows still uncommon [24], [25].

Sutton et al [14] identify several pitfalls of using CDSS in practice, including fragmented workflows, alert fatigue and inappropriate alerts, impact on user skills, dependence on computer literacy, challenges with system and content maintenance, operational issues due to poor data quality and incorrect content, lack of transportability and interoperability, and financial challenges.

The pitfalls discussed below apply to all CDSSs, regardless of the domain they support, whether in eye care or other clinical contexts. Although trust-related factors, which also serve as barriers to adoption, will be discussed in Section 1.4.2, several general pitfalls related to trust are outlined here. These include irrelevant or non-significant alerts that may cause clinicians to lose confidence in the system's recommendations, as well as over-trusting the CDSS, which can reduce clinicians' skill levels and lead to over-dependence on the system.

1.2.3.1 Fragmented Workflows

CDSS can interfere with clinicians' workflow, particularly when systems are designed to require documentation or information retrieval outside the provider's usual workspace [14]. Workflow disruption also occurs when CDSS design fails to account for human cognitive processes [14]. Such inefficiencies can increase mental workload, extend task duration, and reduce the amount of time clinicians spend directly interacting with patients [14].

1.2.3.2 Alert Fatigue and Inappropriate Alerts

Studies indicate that up to 95% of CDSS alerts are of little significance, and physicians frequently question or distrust them [26]. Other times ignore them. When clinicians are exposed to an excessive number of irrelevant alerts, they may experience alert fatigue [27].

For example, an alert discouraging the use of broad-spectrum antibiotics such as vancomycin may be inappropriate in an ICU setting [26], where patients frequently present with severe or resistant infections requiring immediate intervention. In such high-risk environments, clinicians often need to initiate broad-spectrum therapy promptly to prevent clinical deterioration. However, in lower-risk

settings, such an alert can be beneficial in minimizing unnecessary use and reducing the risk of antibiotic resistance.

1.2.3.3 Impact on User Skills

Clinicians may become overly dependent on or place too trust in a CDSS for carrying out certain tasks [28]. This is like relying on a calculator for extended periods and consequently losing proficiency in mental arithmetic. Such dependency can be problematic, as it reduces the user's autonomy and leaves them less capable of performing the task independently in settings where the CDSS is unavailable [14].

1.2.3.4 Dependence on Computer Literacy

Insufficient technological skills can obstruct users from effectively engaging with a CDSS [29]. The extent of this challenge depends on the system's design, as some CDSSs are overly complex, demand high user proficiency, and require extended training to use effectively [30].

1.2.3.5 Challenges with System and Content Maintenance

Keeping a CDSS up to date is important, but this part of its life cycle is often overlooked. This involves taking care of the technical parts of the CDSS. Another challenge is keeping the knowledge base and its rules up to date, as medical practices and clinical guidelines change quickly. Even the most advanced healthcare institutions struggle to keep their systems current as medical knowledge constantly evolves [31].

1.2.3.6 Operational Issues Due to Poor Data Quality and Incorrect Content

The CDSS knowledge base depends on a high-quality clinical data repository created by its users. . However, users may adopt workarounds that compromise data integrity, for example, by entering incomplete, generic, or inaccurate information [31]. In such cases, the accuracy of decision support is affected, as it is directly tied to the quality of the underlying data. CDSSs also rely on data from constantly changing systems; for instance, certain modules may prompt clinicians to place orders even when hospital resources are insufficient.

1.2.3.7 Lack of Transportability and Interoperability

Many CDSS remain as cumbersome standalone tools or are embedded within systems that struggle to exchange information effectively with others [14]. Additionally, concerns or perceived risks surrounding the transfer of sensitive patient data further hinder seamless integration [14].

1.2.3.8 Financial Challenges

Up to 74% of CDSS users reported that maintaining financial sustainability continues to be a challenge [32]. The initial expenses associated with implementing and integrating new systems is considerable. Additionally, ongoing costs often persist, as continuous staff training and regular system updates are needed to ensure alignment with evolving clinical knowledge and practices [14].

1.2.4 CDSS in Ophthalmology and Glaucoma Detection

The use of artificial intelligence (AI) in the diagnosis and management of eye diseases has grown increasingly popular, as studies emphasize its potential to enhance personalized care and improve health outcomes [33]. Especially in the case of glaucoma diagnosis, AI in optometry involves sophisticated algorithms that process complex inputs like fundus images and relevant patient information. These systems leverage image processing, computer vision, deep learning, and transfer learning techniques to identify and diagnose glaucoma [34].

AI algorithms have been developed for analyzing images related to various retinal conditions, such as diabetic retinopathy (damage to the retina caused by diabetes), age-related macular degeneration (degeneration of the central part of the retina), retinopathy of prematurity (abnormal blood vessel development in the retina of premature infants), retinal vascular occlusions (blockage of blood vessels in the retina), and retinal detachment (separation of the retina from the underlying tissue). Additionally, their use has been explored in glaucoma, keratoconus (thinning and bulging of the cornea), cataracts (clouding of the eye's lens), refractive errors, intraocular lens power estimation (calculating the correct lens strength for cataract surgery), and in planning strabismus surgeries (procedures to correct misaligned eyes) [35], [36]. AI shows significant promise as a screening tool, particularly with the approval of systems like IDx-DR and EyeArt by the U.S. Food and Drug Administration for diabetic retinopathy screening, and the i-ROP Deep Learning system for assessing retinopathy of prematurity. These tools assist in identifying cases that require specialist referral [33].

AI-based diagnostic models in ophthalmology have demonstrated high performance across multiple conditions, with some systems achieving accuracy comparable to, or even surpassing, that of experienced ophthalmologists, particularly in tasks such as diabetic retinopathy detection and grading [37].

1.2.5 Glaucoma

Glaucoma is a group of progressive optic neuropathies characterized by the degeneration of retinal ganglion cells and the thinning of the retinal nerve fiber layer, which together result in structural changes to the optic nerve head [38]. Glaucomatous diseases are broadly classified into open-angle and angle-closure types [39]. Among Glaucoma's subtypes, Open-angle glaucoma accounts for the majority of glaucoma cases in the United States, exceeding 80% [40].

Open-angle glaucoma typically progresses without symptoms until its later stages, highlighting the importance of early diagnosis and intervention [39]. It presents as a slowly progressing optic neuropathy, typically beginning with peripheral visual field loss that advances toward the central field in a characteristic pattern [41]. If left untreated, this gradual decline in vision can ultimately lead to irreversible blindness [42].

As the leading cause of irreversible blindness globally, glaucoma has a profound impact on individuals' quality of life [43]. It ranks second only to cataracts as a cause of blindness worldwide [44]. In the United States, the economic burden of glaucoma alone is estimated at \$2.9 billion annually [45]. By 2022, approximately 4.22 million people in the U.S. were living with glaucoma [46] and nearly half were unaware of their condition [47].

Glaucoma is closely linked to aging, with older adults being disproportionately affected [43]. As a result, the global number of individuals aged 40 to 80 with glaucoma is projected to rise significantly, from 76 million in 2020 to 111.8 million by 2040 [48].

Given its high prevalence, asymptomatic progressive nature, and substantial economic cost, glaucoma represents a major public health concern; these challenges have led to increasing attention from healthcare planners and policymakers, who emphasize the importance of early detection and treatment as critical steps in preventing irreversible vision loss [49]. Identifying the disease in its early stages remains essential for halting its progression and preserving vision [50].

1.2.6 Glaucoma Detection and Related AI-Based CDSS

According to the UpToDate website, (an evidence-based, peer-reviewed clinical resource) an article titled “Open-Angle Glaucoma: Epidemiology, Clinical Presentation, and Diagnosis” [51] states that conventional methods for diagnosing glaucoma include fundus examination, visual field testing, intraocular pressure measurement, pachymetry, and optical coherence tomography (OCT).

Glaucoma is a slowly progressing eye disease that is often difficult to detect in its early stages due to the lack of clear or noticeable symptoms [52]. The absence of definitive tests or signs contributes to variability in clinical decision-making, which can be influenced by the optometrist’s level of expertise, subjective interpretation of patient data, and the equipment used [53], [54]. These challenges make diagnosis particularly difficult and highlight the potential value of AI-powered clinical decision support systems (AI-CDSS) in standardizing the diagnostic process and providing a more objective analysis of clinical data [34].

A major benefit of glaucoma screening algorithms is their consistently high predictive performance, particularly in detecting moderate to severe glaucoma [55], [56]. As examples of studies which measures. Niwas et al. assessed a fully automated model for classifying angle-closure glaucoma using anterior segment OCT scans, demonstrating an accuracy of 89.2% [57]. Li and colleagues assessed a deep learning algorithm applied to fundus photographs, which achieved strong performance in identifying referable glaucomatous optic neuropathy, with a sensitivity of 95.6% and specificity of 92% [58]. Andersson et al. demonstrated that a trained artificial neural network (ANN) achieved 93% sensitivity and 91% specificity in assessing glaucoma visual fields, performing on par with, or better than, clinicians [59].

1.3 Problem Statement and Motivation

Despite high performance of glaucoma screening models as an AI-driven CDSS, their translation into clinical practice remains minimal, with real-world integration into clinical workflows still uncommon [24], [25]. Sutton et al. [14] describe pitfalls that clinicians experience in the day-to-day use of CDSS (see Section 1.2.3). However, before reaching that stage, clinicians often encounter a different set of challenges in the adoption pathway that are rooted in technology acceptance [60].

In the context of CDSS, technology acceptance goes beyond simply using the system, it involves accepting and acting on its recommendations [61]. Trust is therefore a key determinant of whether

clinicians perceive AI-driven interventions as acceptable [62], and it has been repeatedly cited as a major factor in the uptake of data-driven tools [25].

A systematic review by AlQudah et al. [63] found that across 18 studies, trust was consistently identified as an external factor influencing technology acceptance in healthcare. While the Technology Acceptance Model (TAM) [64] and the Unified Theory of Acceptance and Use of Technology (UTAUT) [65] remain the dominant frameworks for explaining acceptance, neither explicitly incorporates trust. The review recommended extending these models by integrating trust and other external variables to achieve a more comprehensive understanding of user acceptance.

A cross-sectional study that surveyed Australian clinicians working in hospital and primary care settings found that 63% of respondents believed a lack of trust and confidence in the system's content among end users is a barrier to adoption [66]. Concerns related to trust are commonly mentioned in the literature, but they are rarely explored in detail [25]. Factors which contribute toward building trust are crucial to explore from clinicians perspective for the widespread adoption of AI [67].

1.4 Literature Review

1.4.1 Understanding Trust in CDSS

Trust is a multifaceted concept shaped by both emotional and rational evaluations. At its core, trust involves a willingness to accept one's dependence on another, recognizing that some level of vulnerability is inherent in such a relationship [68]

Building on this idea, O'Neill emphasizes that trust is not granted blindly; it is established through a combination of empirical and normative assessments [69]. We evaluate whether claims align with observable reality and whether actions adhere to recognized standards such as honesty, reliability, and competence.

Mayer et al [70], further elaborate on this by identifying three foundational dimensions that influence our willingness to trust: ability, benevolence, and integrity. Ability refers to the skills and expertise that allow a party to perform effectively within a specific context. Benevolence captures the extent to which the trustee is perceived to have genuine goodwill toward the trustor, independent of personal gain. Integrity reflects the perception that the trustee upholds a set of principles or values that the trustor finds acceptable.

Together, these perspectives show that trust is not a singular or static judgment, but a dynamic process grounded in evidence, shared norms, and personal beliefs about the other party's intentions and capabilities.

1.4.2 Factors Influencing Trust in CDSS

To better understand how trust in AI-based clinical decision support systems (AI-CDSS) is conceptualized among healthcare professionals, while reviewing the literature, a systematic review was found that matched the intended broad literature review goal, which is finding themes of trust in CDSS in general. A systematic review by Tun and colleagues [71], though still in production, is available as an accepted manuscript. The study aimed to identify the constraints and facilitators of trust in AI-CDSS from the perspective of healthcare.

The authors of above mentioned review study employed a structured search strategy using English keywords such as “trust,” “acceptance,” and “perception,” combined with terms like “artificial intelligence,” “AI,” “decision support systems,” “clinical decision support,” and “AI-based decision support,” alongside descriptors of healthcare roles such as “clinicians,” “nurses,” “medical professionals,” and “healthcare providers.” This comprehensive approach enabled the review to capture nuanced perspectives on trust across diverse professional and clinical contexts. The review synthesizes empirical studies published between January 2020 and November 30, 2024, a period selected to reflect the post-COVID-19 surge in healthcare digitization and the increased investment in AI technologies, including the emergence of generative models like ChatGPT in clinical contexts.

Tun et al. [71], Included studies were retrieved from PubMed, Scopus, and Google Scholar. The study populations represented a wide spectrum of roles, including physicians, nurses, nurse practitioners, general practitioners, ICU clinicians, pharmacists, ophthalmologists, oncologists, behavioral health specialists, interdisciplinary teams, and AI practitioners. A total of 27 articles met the inclusion criteria, focusing on healthcare workers’ perceptions, experiences, and levels of trust and acceptance in AI-CDSS across clinical environments, including primary care and hospital settings, with the majority of studies conducted in hospital contexts. Articles that were not peer-reviewed, including editorials, opinion pieces, and other forms of non-research literature, were excluded.

The included studies used a range of methodologies, encompassing qualitative, quantitative, and mixed methods studies, and quantitative cross-sectional surveys. In terms of geographic distribution,

the majority of studies took place in the United States (12 studies), followed by Europe (7), with a few conducted across multiple countries (3), in the United Kingdom (2), and individually in China, Singapore, and Egypt (1 study each). The sample sizes of the studies included varied widely, from small focus groups to large cohorts of over 1,000 participants.

A peer-reviewed preprint version of a systematic review was consulted to identify studies examining factors that influence trust in AI-based clinical decision support systems (AI-CDSS). The table in the systematic review paper lists the studies included along with the trust-related factors they reported. However, the synthesis and thematic descriptions in the discussion section lack proper citations linking the themes and relevant codes to corresponding studies. “Consequently, the table provided in the appendix of the aforementioned systematic review was used in this thesis as a reference point for identifying relevant papers. These papers were then independently reviewed by the author to extract trust factors, which were grouped and cited with their corresponding sources. The grouping process follows the fourth step of thematic analysis by Braun and Clarke [72], which involves categorizing trust factors identified in the papers into meaningful, logically coherent groups to form emerging concepts. Although this approach is technically designed for codes derived from interviews, I applied the same methodology to categorize trust factors that relate to one another. The process was conducted by Ali Karim in an interactive manner, which included regrouping, merging, and splitting categories until all trust factors within each group fell under the same conceptual umbrella.

Several themes emerged from the literature review regarding healthcare workers’ trust in AI-based Clinical Decision Support Systems (AI-CDSS), including (1) perceived ai model quality, (2) explainability, transparency and Interpretability of AI systems, (3) clinical accountability, (4) training and general familiarity with AI, (5) evidence-based and peer-endorsed validation, (6) alignment with clinical judgment, (7) experience-driven trust development, (8) miscellaneous factors. These themes, along with number of studies citing each theme, are summarized in the Table 1, and will be elaborated upon in the following paragraphs.

Table 1. Trust Factor Emerging Themes Derived from Literature Review

Emerging Theme	Number of Supporting Papers	References
Perceived ai model quality	12	[59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70]
Explainability, transparency and interpretability of ai systems	10	[59], [61], [62], [63], [66], [67], [71], [72], [73], [74]
Clinical accountability	9	[60], [64], [67], [72], [74], [75], [76], [77], [78]
Training and general familiarity with AI	7	[61], [67], [74], [79], [80], [81], [82]
Evidence-based and peer-endorsed validation	6	[61], [62], [65], [66], [81], [83]
Alignment with clinical judgment	4	[66], [84], [85], [86]
Experience-driven trust development	5	[65], [74], [76], [81], [87]
Miscellaneous factors	4	[61], [74], [78], [84]

1.4.2.1 Perceived AI model Quality

A total of 12 studies underscore the significance of AI model quality as a critical determinant of user trust across both healthcare and non-healthcare contexts. Trust in AI-CDSS is closely linked to perceptions of model quality. This begins with transparency regarding model inputs, including the type and source of training data, and whether the data used are relevant and representative of the clinical context [73], [74], [75], [76]. Furthermore, clearly communicating the criteria used for model selection, as well as the size and characteristics of the training cohort, is perceived as essential for ensuring the robustness of the model development process, thereby enhancing transparency and, ultimately, trust [77]. The credibility of the AI developer has also been cited as an important factor influencing clinicians' trust in the system [74]. These factors collectively contribute to perceptions of

model accuracy [77], [78], [79] and recommendation accuracy [79], [80], both of which are fundamental to establishing trust.

The accuracy of AI-CDSS is often associated with the system's ability to minimize false positives and false negatives [78], an essential indicator of its clinical reliability. Accuracy is frequently conveyed through confidence scores, typically expressed as percentage values. Low confidence scores (e.g., an AI recommendation with only 30% confidence) have been shown to significantly reduce trust in the system's outputs [81]. Similar patterns have been observed beyond the healthcare domain, where both the perceived accuracy and confidence of AI models are consistently identified as key determinants of user trust [82], [83], [84]. Additional concerns arise when AI systems fail to detect severe abnormalities beyond their intended scope or demonstrate performance limitations due to external factors such as poor image quality, both of which can substantially undermine clinicians' confidence in their diagnostic capabilities [78]. Addressing these issues is therefore essential for fostering trust in AI-CDSS.

1.4.2.2 Explainability, Transparency and Interpretability of AI Systems

Explainability, transparency, and interpretability significantly influence clinician trust and confidence in AI-based tools, as indicated by findings from 10 papers. Explainability encompasses clear documentation supporting algorithm validity, the logic behind the outputs, the meaning of risk scores for patients, and the rationale behind specific patient recommendations [85]. Transparency of AI systems involves understanding the data quality, data sources, underlying model logic, the population data used to develop the model, and clarity about the algorithm's intended optimization outcomes [73], [75], [76], [80], [86]. Interpretability primarily refers to the clarity and comprehensibility of AI-generated recommendations [73].

Clinicians' perception of AI credibility improves when explanations are example-based, clearly justifying the recommendations [77], [80]. Conversely, AI tools perceived as "black box" systems, characterized by opaque decision-making processes and insufficient explanation or justification, lead to distrust and diminished confidence in AI-generated diagnoses [80], [87], [88]. Additionally, the utility of explanations matters to clinicians, especially in distinguishing important from less important information [77].

Overall, clinician trust is closely linked to their understanding of AI system development, algorithm complexity, and the transparency of the model's operational logic [80], [89].

1.4.2.3 Clinical Accountability

Clinicians' sense of professional responsibility and their perception of risk, both play a critical role in shaping trust in AI-based clinical decision support systems (AI-CDSS), as supported by nine papers. One of the primary concerns is the possibility that the AI system may introduce medical errors, while the legal and ethical responsibility for patient outcomes remains with the clinician [90]. This perceived risk, particularly the fear of being held accountable for errors generated by the AI, can significantly reduce trust in the system [91]. As clinicians are ultimately responsible for making final decisions about patient care, they are unlikely to fully rely on AI-CDSS without the ability to independently assess its outputs [87]. Even when they do trust the AI, they still prefer to review and validate its recommendations before acting on them [74], [87]. Accordingly, the degree of perceived control over clinical decision-making emerges as a critical factor in building and maintaining trust [92].

Another related concern is the risk of over-reliance on AI, where clinicians may gradually become less engaged in active decision-making, leading to a decline in their clinical judgment and skills [89], [93]. Addressing this issue involves calibrating clinicians' expectations of what AI can and cannot do, a process shown to positively influence trust when managed appropriately [91]. Providing targeted training and onboarding can help clinicians develop a realistic understanding of the system's role, thereby fostering more balanced and informed use, rather than treating them as unquestionable outputs.

One study pointed out errors in the AI system, raising concerns about its reliability and suggesting that clinicians should not overlook green-labeled images without careful review [78]. This highlights the importance of thoroughly assessing all images, including those marked as low risk, to minimize errors and reinforce clinician accountability. Another study also emphasized the need to ensure that AI-generated recommendations are clinically appropriate and make sense within the context of patient care to maintain patient safety and support clinician accountability [86].

1.4.2.4 Training and General Familiarity with AI

Seven studies underscore the importance of general familiarity with AI, along with initial training and onboarding, as key factors influencing trust in both healthcare and non-healthcare domains. Trust in AI-based clinical decision support systems (CDSS) is shaped by individual perceptions, which vary significantly among clinicians [75]. A key contributing factor is the level of AI literacy, which differs

from person to person. Prior research in non-healthcare domains has highlighted AI literacy as a determinant of trust in intelligent systems [94], [95], [96]. Notably, clinicians with greater familiarity with machine learning (ML) models have been found to be less likely to follow ML-generated recommendations compared to those with lower familiarity [96], suggesting a more critical or cautious stance toward system outputs.

Effective onboarding and training are instrumental in fostering trust in CDSS by improving clinicians' understanding of system features and functionalities and by enhancing the perceived transparency of the AI model [81], [87]. Given the time-constrained nature of clinical environments, clinicians often prefer to calibrate their trust in AI systems at an early stage, rather than reassessing it continuously during patient care [96]. This preference underscores the importance of initial training and trust formation outside of direct clinical interactions [89].

1.4.2.5 Evidence-Based and Peer-Endorsed Validation

Validation and Endorsement of AI-CDSS is a key theme influencing clinicians' trust, supported by evidence from six unique papers. In the context of clinical decision support systems (CDSS), the validity of any intervention, particularly AI-powered CDSS, is highly valued and plays a critical role in shaping clinicians' trust. Clinicians consider several forms of validation when determining whether an intervention is trustworthy, with randomized controlled trials being the most highly regarded [75], [79], [96]. External validation is also valued due to its unbiased nature [97]. Additionally, pilot studies of AI-based CDSS that compare outcomes with usual care are considered another means of establishing preliminary validation [76].

Professional endorsement serves as a strong foundation for trust, especially when the AI-CDSS is recommended by medical societies [75], [96]. or expert clinicians in the field [89], [96]. Longitudinal prospective validation is also cited as a trust-building factor [80] because it confirms an AI-CDSS's performance using new patient data under real-world clinical conditions, providing stronger evidence of reliability than retrospective methods.

To develop a trustworthy and actionable AI-powered CDSS, the clinical recommendations it generates must be not only accurate but also clinically sound. Validation of these recommendations is therefore essential. Clinicians typically rely on two key sources: guidelines issued by professional societies and peer-reviewed journal articles [75]. Guidelines are particularly important because they reflect expert consensus, are methodologically rigorous, and serve as standardized references for

practice. However, since they evolve slowly, scientific literature, being more current, is often preferred in fast-moving clinical contexts [75].

1.4.2.6 Alignment with Clinical Judgment

This theme is evidenced by four studies, two conducted in healthcare settings and two in non-healthcare domains. Clinicians reported that when AI recommendations align with their own clinical assessments, it contributes to building trust [98]. Conversely, disagreement with AI recommendations tends to reduce trust [80]. This concept is even more rigorously supported in studies beyond healthcare, where the level of alignment or disagreement between individuals' own judgments and AI recommendations has been shown to influence their trust levels [99], [100].

1.4.2.7 Experience-Driven Trust Development

This theme, supported by five unique papers, highlights the role of previous general AI system usage as a trust-building factor, both within healthcare [96] and beyond healthcare contexts [101].

Familiarity with basic AI concepts enhances users' understanding of the model, which contributes to increased trust. Trust is developed over time, both positively and negatively, through direct interaction with the system across different clinical scenarios and patient cases [79], [89], where clinicians can observe the quality of outputs and recommendations. This process is closely tied to the perceived performance of the system: when the AI aligns with clinicians' expectations, it fosters trust [91]; however, perceived risk, such as the potential for errors or unpredictable behavior, can diminish clinicians' confidence in the system [91].

1.4.2.8 Miscellaneous Factors

Finally, there are several important but less frequently reported factors that contribute to trust in AI-based CDSS. One such factor is patient complexity. Clinicians tend to trust the system more when the patient presents with a common clinical status [98], as these cases are perceived to be within the AI system's capabilities. This reflects the broader concern about the system's performance when dealing with more complex, less routine scenarios.

Interface customization also plays a role in building trust. Allowing clinicians to ask questions and provide design input aligns with participatory design principles, which are widely regarded as effective for creating trustworthy systems [89]. Similarly, adjusting alert sensitivity to match user needs enhances perceived control and system relevance, further strengthening trust [89].

Interestingly, the sponsoring company behind the AI system is also seen as a trust factor, primarily due to the company's reputation, an organizational attribute is associated with system trust [75].

Finally, some clinicians express concerns that AI could contribute to the dehumanization of healthcare by diminishing patient-clinician empathy and human warmth [93]. This trust barrier may be addressed through a human-in-the-loop approach, ensuring that clinicians retain final decision-making authority and reinforcing the human role in patient care.

1.4.3 Glaucoma-specific Trust Factors

The literature identifies several key factors that influence clinicians' trust in glaucoma detection CDSS. These include: the system's reliability across different levels of disease severity, the degree of explainability and transparency, achieved through techniques ranging from visual explanations to feature attribution methods; the system's demonstrated consistency over time; and its ability to align with the multimodal nature of clinicians' diagnostic workflows. These factors are explored in detail in the following paragraphs.

Evidence indicates that clinicians' trust in glaucoma CDSS is not uniformly distributed across clinical scenarios involving varying disease severities. Clinicians tend to place greater trust in these tools when managing early-stage or moderate glaucoma, while their trust diminishes in advanced cases. Chen et al. reported that as the severity of glaucoma increases, the mean Likert trust level in the AI system decreases [102]; Van den Brandt et al. report that, in clinical settings, GLANCE, a glaucoma detection decision support system, is considered more trustworthy when used in mild to moderate cases of glaucoma[103]. This trend may reflect clinicians' increased caution when dealing with complex or high-stakes cases, where any system limitations are more consequential.

Explainable AI (XAI) techniques have been increasingly used to interpret the outputs of “black-box” deep learning models in glaucoma detection, with promising effects on clinician trust. Multiple approaches have been employed to increase transparency, focusing both on image interpretation and on explaining the contribution of clinical risk factors to the final diagnosis. Heatmap-based visual explanation for analyzing color fundus images has been shown to enhance trust among medical professionals [104]. Pixel Density Analysis (PDA) has been applied to fundus images to interpret image features, revealing that the optic nerve region appears more prominently in glaucoma-positive cases [105]. Visual-based explanation tools, such as bounding boxes highlighting areas of concern on

optical coherence tomography (OCT) images, further support clinicians in calibrating their trust in the system and making decisions with greater confidence [103].

Regarding transparency in contributing clinical risk factors, Submodular Pick for Local Interpretable Model-Agnostic Explanations (SP-LIME) has been employed to identify key clinical features influencing model predictions, thereby providing more interpretable and trustworthy explanations for glaucoma diagnoses [105]. In addition, SHapley Additive exPlanations (SHAP) have been used to highlight the top contributing features and quantify the weight of each factor in the final clinical decision, offering a scoring-based explanation of the model's output [106].

While Explainable AI (XAI) is generally thought to enhance trust in clinical decision support systems, some evidence suggests that not all explanation types are equally effective. One study found that the inclusion of explanations, specifically, Feature Importance Explanations (which present the top three most influential features based on SHAP values) and Scoring-based Explanations (which provide rules for score calculation based on feature distributions), did not significantly improve clinician trust in the system [106].

Visualizing the alignment between actual and predicted values over time can positively influence clinicians' trust by reinforcing the perceived reliability of the system. For example, displaying a trendline of the actual mean deviation from visual field measurements, as an indicator of glaucomatous damage, alongside the predicted trajectory, enables clinicians to evaluate whether the AI's predictions align with the observed progression of glaucoma over time. This visual consistency serves as a trust cue, helping clinicians gauge the system's performance over time and make more confident decisions [103].

Optometrists are more likely to trust a decision support system when it incorporates data modalities that align with their typical diagnostic workflow [107]. Trust is enhanced when the system reflects the multifaceted nature of clinical reasoning by integrating multiple sources of information, such as fundus images, optical coherence tomography (OCT) scans, medical history, and visual field results, rather than relying on a single input. This multimodal approach mirrors the way optometrists assess patients in practice, thereby increasing the system's perceived relevance and credibility [107].

1.4.4 Identified Gaps

Despite a growing body of work on clinician trust in CDSS, glaucoma-specific evidence remains thin. Only a handful of studies have examined trust in AI-driven glaucoma screening, even fewer have used Canadian optometrists, and the majority concentrate on explainability alone, ignoring other established determinants of trust. This narrow and geographically limited evidence base hampers the design of tools that Canadian optometrists will actually adopt. A multi-factor investigation that goes beyond explainability is therefore required to pinpoint the full set of trust drivers and translate them into design requirements for the next generation of glaucoma screening CDSS.

1.4.5 Research Objectives

The objective of this thesis is to identify the factors influencing optometrists' trust in an AI-powered glaucoma screening tool, conceptualized as an AI-based clinical decision support system (AI-CDSS), to propose design recommendations that enhance trust in future iterations of the tool.

1.4.6 Research Questions

To achieve the study's overall objective, the following Research Questions (RQ) were posed:

RQ1) What key factors influence optometrists' trust in AI-powered clinical decision-support systems (CDSS) for glaucoma screening?

RQ2) How do optometrists evaluate the performance of glaucoma-focused AI CDSS, and how do those evaluations translate into varying levels of clinical trust?

RQ3) What design principles should guide the development of glaucoma AI CDSS to foster and sustain optometrists' trust?

1.5 Thesis Outline

This thesis is organized into four chapters, followed by references and appendices. Chapter 2 describes the methodological approach, including an overview of the study procedure, the design of the mock interface, participant recruitment, data collection and analysis methods, and ethical considerations. Chapter 3 presents the results, beginning with participant demographics, and continuing with the qualitative findings, which are organized into themes and subthemes. Chapter 4 provides the discussion, interpreting the key factors influencing trust in AI-based glaucoma screening systems, examining how performance evaluations translate into trust, and outlining design principles

to support trustworthy CDSS design. It also highlights the study's strengths and limitations and concludes with directions for future work. The thesis ends with a list of references and appendices containing supporting materials, including demographic information, participant letters, the study brochure, questionnaires, feedback and consent forms, and the interview script.

Chapter 2

Methodology

2.1 Overview

This chapter begins by outlining the study procedure, followed by a description of the mock interface used to simulate patient–optometrist interactions. Sections 2.5 and 2.6 detail participant recruitment, and data collection. These are the procedures from Ghaffar’s study and are included here for context. The procedures for the thematic analysis that I did begins in Section 2.7.

2.2 Study Procedure

This study is a subset of a larger PhD research project designed and conducted by Faisal Ghafar [108]. While the broader study included multiple components, such as various questionnaires, this thesis focuses exclusively on the qualitative analysis of optometrists’ verbal responses regarding factors influencing their trust during AI-assisted simulated patient evaluations, a dimension not covered in the original thesis.

Each participant followed a structured, repeatable workflow. First, participants provided informed consent, as outlined in Appendix G. They then completed a demographic questionnaire, mentioned in Appendix E, which collected information on gender, age, ethnicity, highest level of education, and any specific glaucoma related certificates. Also, during each interview participants were asked about their years of practice as an optometrist, average number of patients seen per week, average number of glaucoma patients screened per week. This information helped contextualize the results. Next, participants interacted with two identical patient cases, each typically with three visits at different times, and were used to simulate the patient-optometrist interactions.

Data for this study were obtained from the School of Optometry and Vision Science at the University of Waterloo, Canada. Patient records from the institution’s database were reviewed, with selection restricted to those cases in which patients had provided explicit consent for their information to be used in research. Each patient’s record offered a longitudinal view, capturing data across multiple clinical visits. The selection of data attributes from patient records was guided by evidence and practices reported in prior studies of a similar nature [109], [110]. The data collection captures a broad range of patient characteristics, incorporating both key ocular health metrics and relevant

external factors. This comprehensive approach ensures that optometrists can conduct a thorough evaluation of glaucoma risk and progression. The patient features included in this study are outlined in Table 2. Out of the initial pool of patients, cases were chosen based on criteria aimed at ensuring they were appropriate for both unimodal and multimodal approaches, they were drawn from the suspect diagnostic range, an area where AI could provide added value, commonly marked by clinical uncertainty, though no explicit difficulty grading was applied. Ultimately, two patient records, referred to as Case A and Case B, were chosen for optometrists to review during in-depth interviews.

Table 2- Data Features Collected for Glaucoma Diagnosis

Category	Factors
Demographic factors	Age, gender, and race
Historical data	Ocular history, medical history, medication history, social history, and family history
Internal attributes	Habitual spectacle prescription, unaided visual acuity, aided acuity, pupils, anterior segment examination, Goldman applanation tonometry, central corneal thickness, and gonioscopy
Images	Fundus image and Optical Coherence Tomography (OCT) images
Reports	Posterior segment examination, visual field, and OCT

Patient cases were presented in both unimodal and multimodal versions (the unimodal and multimodal systems are discussed later in Section 2.3). The order of exposure was as follows:

- **Unimodal interface:** Case A (Visit 1, Visit 2), Case B (Visit 1, Visit 2, and Visit 3)
- **Multimodal interface:** Case A (Visit 1, Visit 2, Visit 3), and Case B (Visit 1, Visit 2, and Visit 3)

At each visit, participants first reviewed the patient data (e.g., patient history, fundus photos, OCT, and visual field results) along with the AI output (diagnostic label, suspicious level, confidence score, risk factors, current diagnostic findings, and conclusion and recommended actions). Participants were prompted to make a final diagnostic judgment while verbalizing their thought process using a think-

aloud protocol [111]. This method captured their reasoning and made them engaged with the data. After each visit, participants completed a brief follow-up questionnaire to assess their understanding of the case. The interview script containing these items is provided in Appendix H. Figure 1 summarizes the study design as described above.

It is worth noting that all cases in both unimodal and multimodal conditions included three visits, except for Unimodal Case A, which had two visits. This is because the domain expert who assisted in co-designing the system and refining the data suggested removing one visit due to ambiguous data.

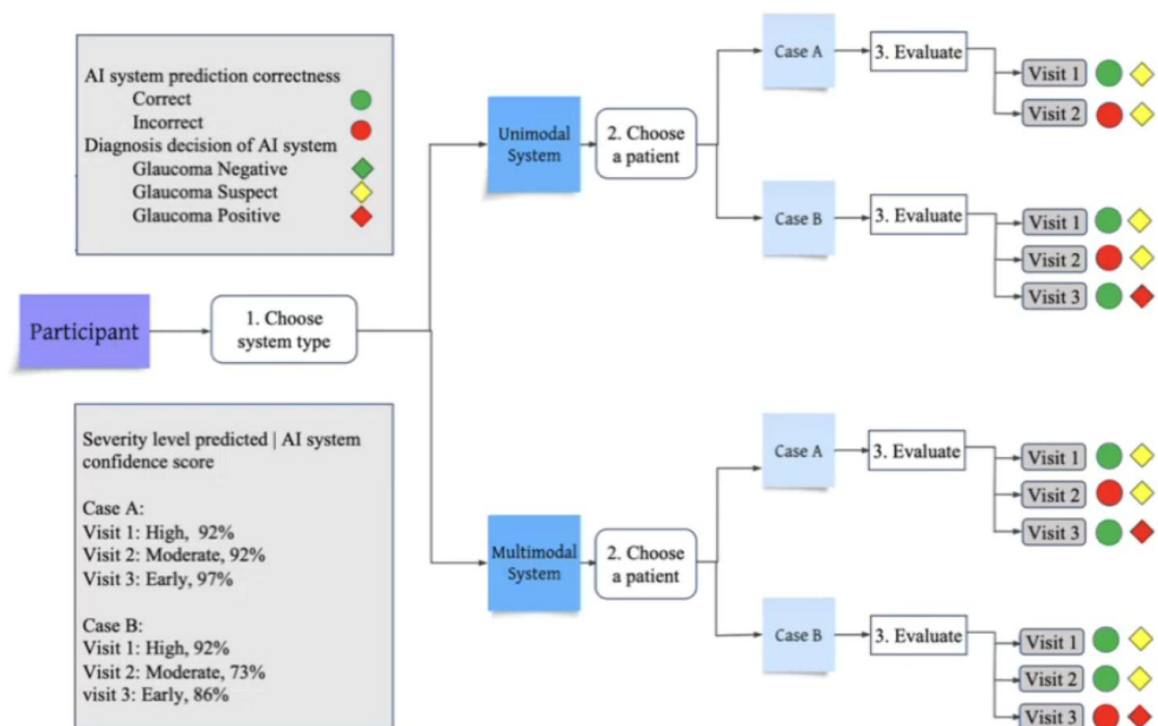


Figure 1- Study Design

This study focused on optometrists' responses to the follow-up question: **“Would you trust this AI system to be used in a clinical environment? Explain.”**, asked at the end of each visit by interviewer. Additionally, spontaneous mentions of trust during the debriefing sessions, which provided broader reflections on their experience with the AI system, were also included in the analysis.

Out of the 11 patient visits presented in the study, 4 cases were deliberately assigned an “incorrect” final diagnosis, noted in Table 3. This adjustment was intentional and served a methodological purpose: it was designed to introduce diagnostic uncertainty and observe how optometrists’ reason through ambiguous or conflicting information.

Table 3- AI Diagnostic Changes

Patient Case-Visit Number	Correct	Incorrect but presented on interface
Case A, multimodal and unimodal, visit 2	Glaucoma positive	Glaucoma suspect
Case B, multimodal, visit 3 Case B, unimodal, visit 2	No Glaucoma	Glaucoma positive

2.3 Mock Interface Design Description

Patient information is categorized into four sections: “General”, “Fundus”, “OCT”, and “Visual Field”, as indicated by label c in Figure 2. Each section is accessible through its own tab. The general section covers patient demographics, ocular and medical history, medication use, and social and family history. It also includes ocular details such as pupil testing, extraocular muscle movements, and anterior segment details for the right eye (OD) and the left eye (OS). The fundus tab presents fundus photographs for both eyes. The OCT tab provides retinal nerve fiber layer (RNFL) single-exam and trend reports for each eye. The visual field tab contains single field analysis as well as glaucoma progression analysis (GPA) with baseline GPA follow-up data.

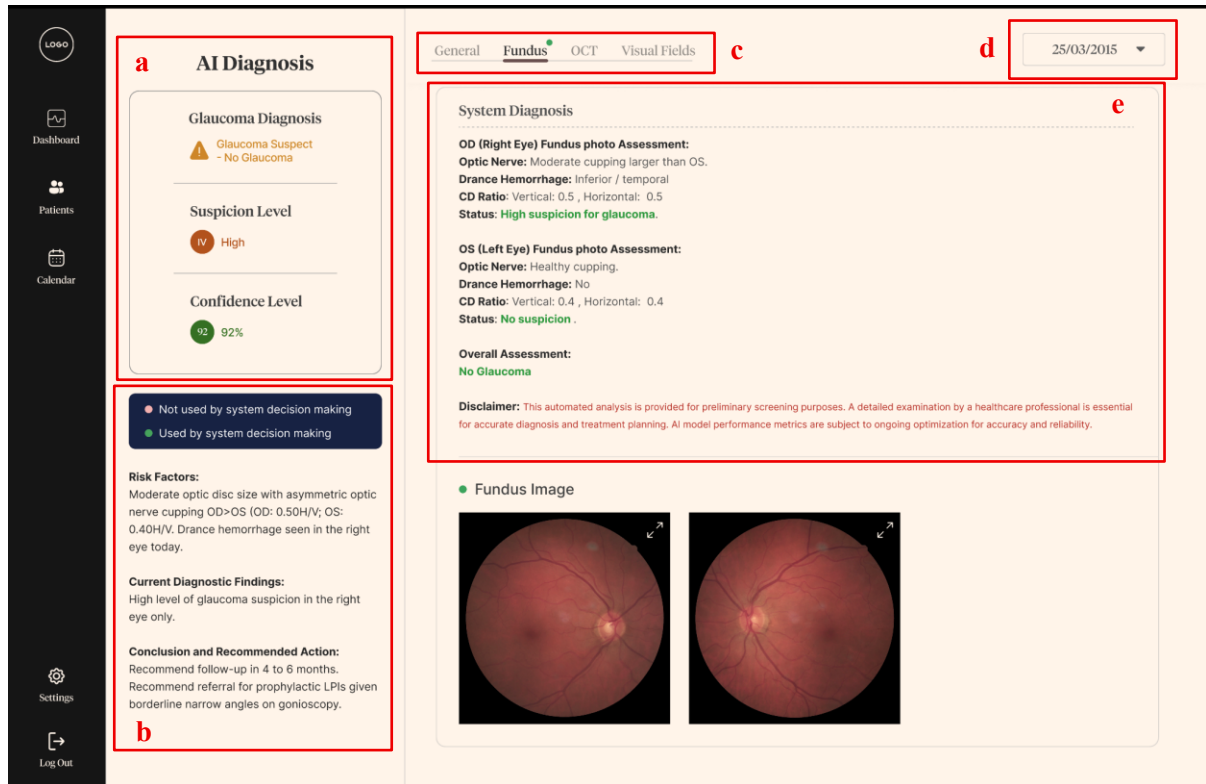


Figure 2-Unimodal System, Fundus Tab

The AI system output is structured into two components: diagnosis and explanation. The diagnosis component, labelled (a) in Figure 2, presents the key decisions generated by the AI system. These include the primary diagnosis (glaucoma positive, glaucoma suspect, or no glaucoma), a suspicion rating that further specifies the diagnosis (high, medium, or low for suspects; early, mild, moderate, or severe for glaucoma positive), and the system's confidence level in its decision.

In the interface, diagnoses and suspicion levels were determined using the majority decision from Faisal Ghaffar's earlier study [112] while confidence scores were derived from the proportion of participants who selected that majority diagnosis in the same study.

The explanation component, labeled (b) in Figure 2, outlines the rationale behind the AI system's decision (including risk factors and diagnostic findings), along with a conclusion and recommended next steps. These explanations were manually designed to mirror the reasoning process of each system. Their content differed by modality: unimodal systems referred to a single input source (fundus image), while multimodal systems incorporated all available input data (general information,

fundus imaging, OCT, and visual field). Each explanation included the decision factors, diagnostic findings, and the conclusion with recommended actions.

The initial explanations were developed from the qualitative findings of the earlier Faisal Ghaffar's study [112], which examined how participants interpreted and reasoned with identical patient data. These qualitative insights informed the creation of explanation content designed to reflect realistic clinical reasoning. The explanations were subsequently refined through iterative feedback from a practicing optometrist to ensure accuracy, relevance, and consistency across cases.

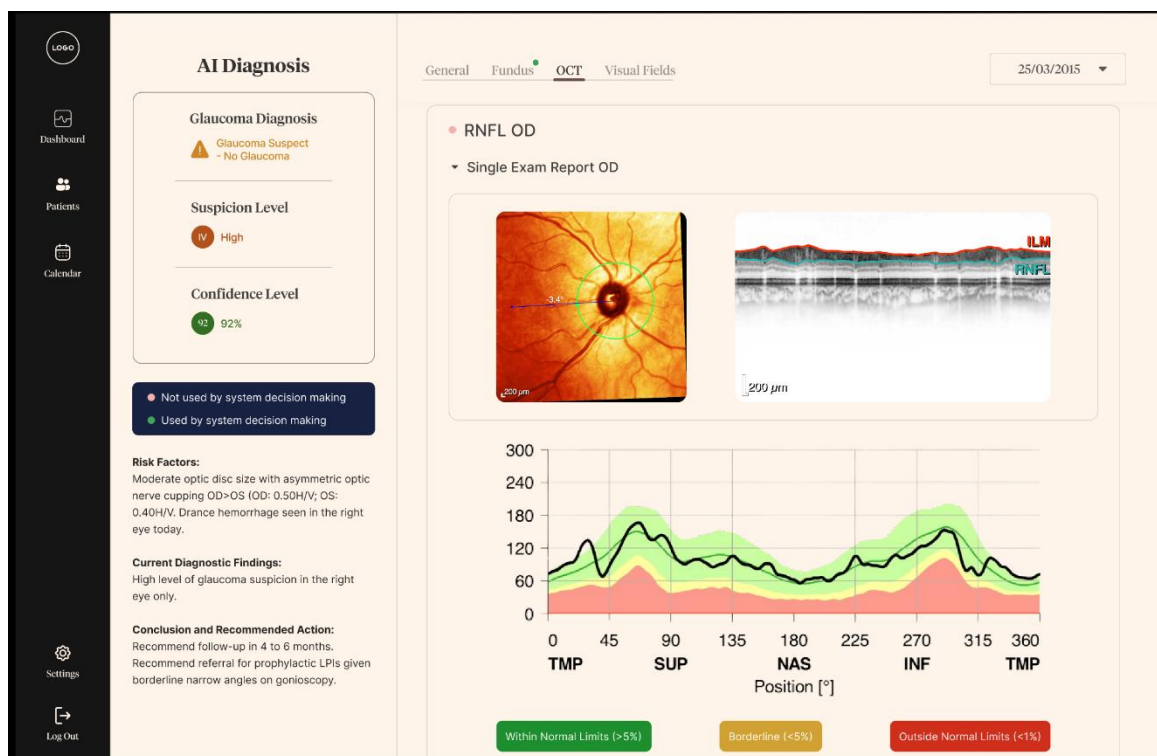


Figure 3- Unimodal System, Other Tabs Except Fundus

The unimodal and multimodal AI systems were designed with notable differences, which were explained to participants to ensure they could clearly recognize the system modality they were using. In addition to the verbal explanation, a visual aid was provided on the interface: a green dot with a legend indicated that the corresponding tests were used by the AI system for diagnosis, while sections marked with a red dot were not utilized by the system.

In the unimodal system, decisions relied exclusively on fundus images. AI-generated explanations were provided for both the eyes within the fundus imaging tab, and a summary of diagnostic

decisions was displayed in the AI diagnosis panel. However, the explanations were limited to information derived solely from fundus imagery.

The multimodal AI system followed a multi-agent design, integrating separate AI models for fundus images, OCT, and visual field data. Each model produced its own interpretation and explanation within its respective tab, shown as local explanations for fundus photo in a red box labeled (e) in Figure 2, when selected by the user. These tab-specific explanations, together with general patient information, contributed to the local explanation displayed in the AI diagnosis panel. Notably, the decision factor visualization was a feature unique to the multimodal system, labeled (a) in Figure 4.

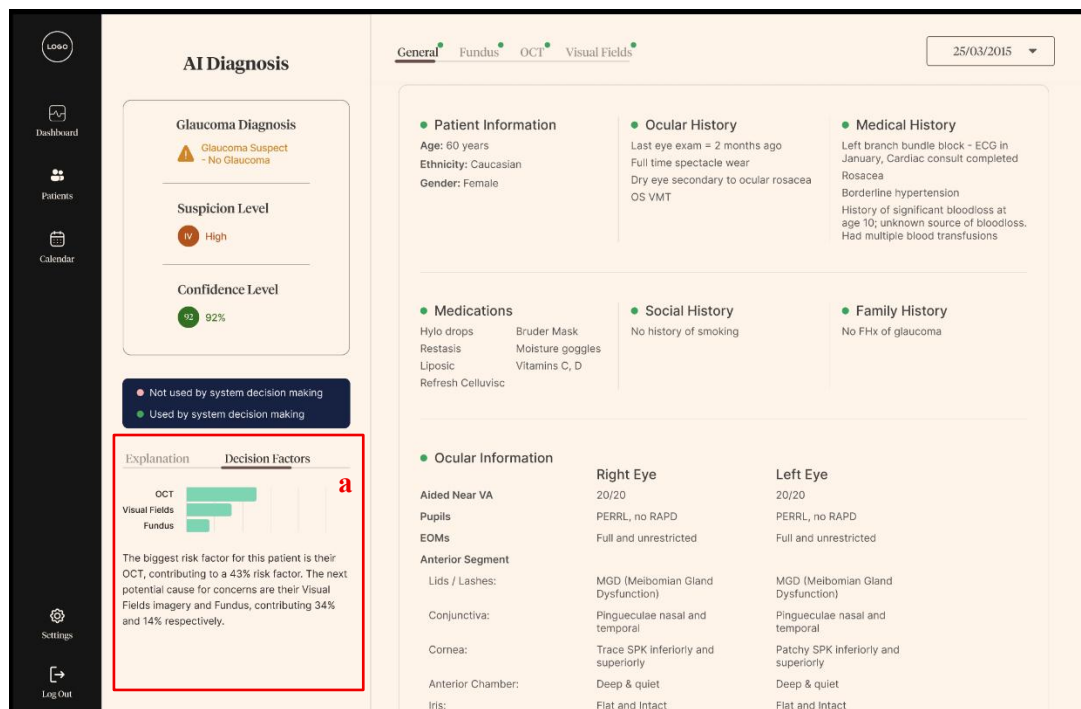


Figure 4- Multimodal System, Visual Field Tab

For easier cross comparison of unimodal and multimodal systems, Figure 5 is provided. This photo taken from Faisal Ghaffar's thesis [108].

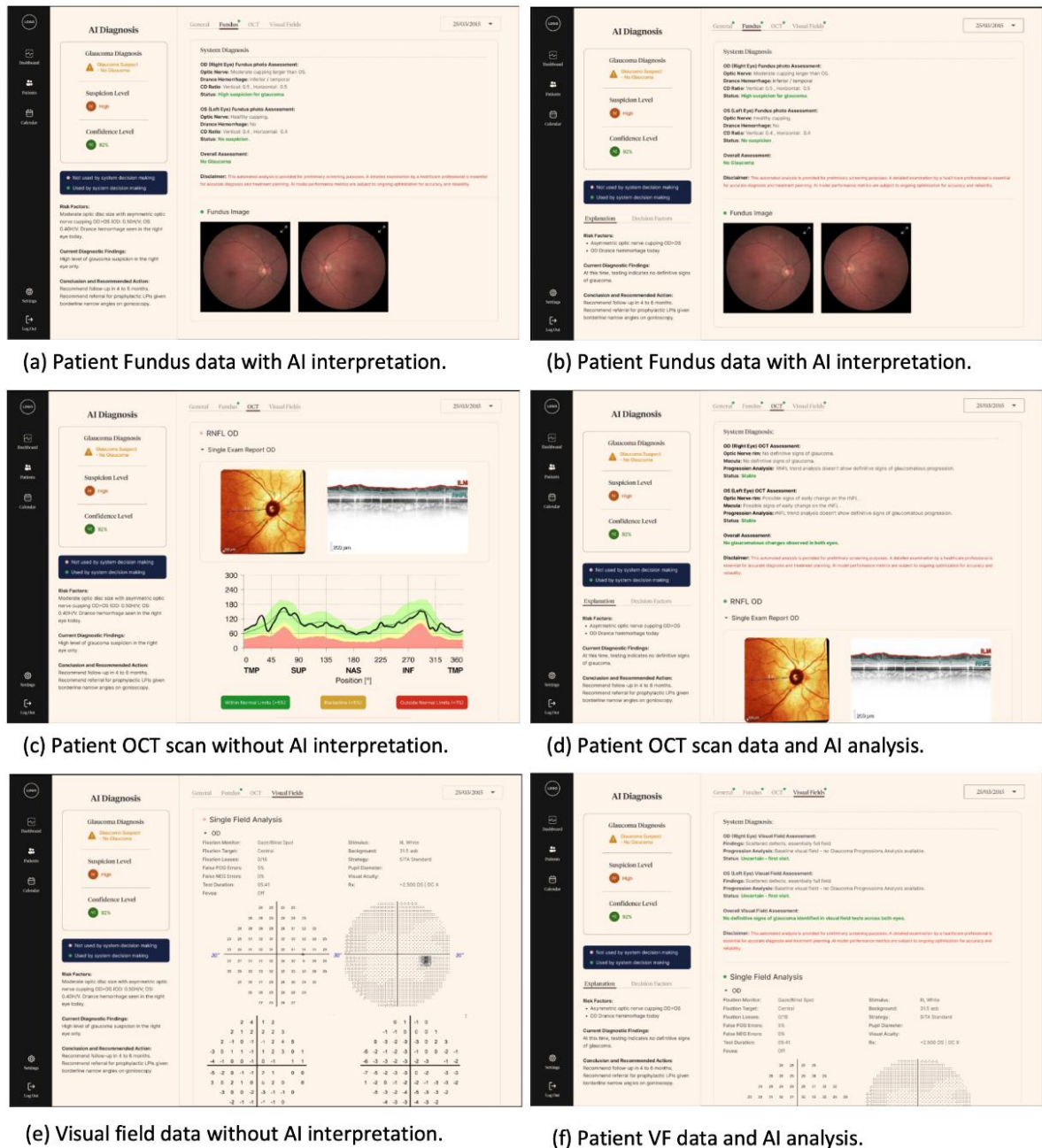


Figure 5- Multimodal vs Unimodal Cross Comparison over Tabs

2.4 Ethical Considerations

Ethics approval for this study was obtained from the University of Waterloo Research Ethics Board (REB) under application number #45248, titled “Human decision variation with accuracy and data

modalities of AI system.” Informed consent was obtained from all participants prior to their involvement in the study, and participants were assured of confidentiality, voluntary participation, and the right to withdraw at any time.

Informed consent was obtained from all participants prior to their participation. Consent was documented both through a written signature on the consent form and through verbal confirmation at the beginning of the session. A copy of the consent form is provided in Appendix G.

2.5 Participants

A total of 20 licensed optometrists participated in the study; all actively engaged in clinical practice with prior experience in glaucoma screening and diagnosis. Eligibility criteria specified that participants had to be licensed optometrists practicing in Ontario and between 18 and 64 years of age. No exclusion criteria were applied.

2.6 Data Collection Methodology

The participant pool consisted of optometrists from the School of Optometry and Vision Science at the University of Waterloo, as well as other optometry clinics across Ontario. These optometrists were contacted via their personal email addresses by Dr. Nadine Furtado. The recruitment email scripts are provided in Appendix B (for optometrists known to Dr. Furtado) and Appendix C (for those not known to her). In addition, optometrists were informed about the study through brochures, included in Appendix D.

The study begins with a brief introduction by the research team, outlining the investigators’ backgrounds and the purpose of the project. Participants are thanked for their time and provided with an overview of the study and its objectives. They were informed that the session will be recorded, and the structure of the study is explained. Participants are reminded that their involvement is voluntary, that they may decline to answer any questions, and that they may withdraw at any time, in which case their data will be erased. Also, participants are also asked to complete a demographic questionnaire explained in Participants section.

The interviews were conducted as part of Faisal Ghaffar’s (FG) main study and were carried out by FG and Yousuf Zia Islam (YZI), with the majority led by FG. Microsoft Teams was used to record

and transcribe interviews. A semi-structured interview format was followed, concluding with a short debrief session. The interview script is provided in Appendix H. Each interview lasted approximately two hours. As part of the feedback and appreciation process, participants were thanked immediately following the session and subsequently received a follow-up email, the script for which is included in Appendix F. As compensation for their time, participants received an honorarium of \$150 upon completing the study, regardless of the exact time spent. If a participant chose to leave the study before completing the session, a pro-rated amount of \$75 per hour was provided.

2.7 Data Analysis Methodology

2.7.1 Thematic Analysis

The interviews were analyzed using inductive thematic analysis, guided by Braun and Clarke's widely adopted six-phase framework [72]. This approach was chosen because it allows for a systematic yet flexible process of identifying, analyzing, and reporting patterns within qualitative data. The analysis was conducted iteratively, with multiple coders involved to enhance rigor, reduce individual biases, and ensure that the findings reflected participants' perspectives rather than researchers' assumptions. The following sections outline each stage of the analysis process in detail.

- 1) **Familiarizing with the data:** The transcripts were read by researchers, Ali Karim and Wendy Xu, who became familiar with domain-specific terminology, interview structure, and potential sections of interest, while taking notes to gain a high-level overview of the data.
- 2) **Generating initial codes:** Microsoft Word and Microsoft Excel for Microsoft 365 MSO (Version 2507) were used to manage the data. Word was used to code the transcripts, while Excel was used to track codes across transcripts and to organize them into groups. To ensure rigor, multiple coders were involved in the thematic analysis process. Different areas of expertise were incorporated, and multiple perspectives were brought in to highlight nuances that might have been missed by a single coder.

All transcripts were coded independently by two researchers with different backgrounds, Ali Karim (AK) and Wendy Xu (WX), in a flexible, iterative, exploratory, and inductive manner. Coding decisions were compared in multiple in-person sessions, during which discrepancies were identified and discussed. When disagreements persisted, they were presented to Faisal Ghaffar (FG) in a virtual meeting on Microsoft Teams, attended by both AK and WX.

Through this triangulation process, codes were applied consistently, and consensus was achieved, thereby strengthening inter-coder reliability, reducing individual biases, and enriching data interpretation. The complementary expertise of the research team further enhanced the analysis. An outsider's perspective was provided by WX, who was not initially familiar with trust and glaucoma screening terminology, which helped make the codes clearer and more accessible for readers from different backgrounds. Domain-specific expertise was contributed by FG, who had PhD-level experience and direct engagement with optometrists during study design and interviews, ensuring that field-specific terminology was applied appropriately while remaining understandable. Qualitative analysis expertise was contributed by AK, who was familiar with trust-related concepts through extensive literature review.

- 3) **Searching for themes:** Codes were collated into potential categories that centered on a concept, with all relevant data gathered under each. Broader patterns, similarities, and connections between codes were examined to determine whether themes encompassed multiple sub-categories (sub-themes) formed around those central concepts. After the codes were consolidated, themes and sub-themes were independently developed by AK and WX,
- 4) **Reviewing themes:** Themes independently developed by researchers (AK and WX) were compared and refined between researcher's emerging themes through multiple online discussions on Microsoft Teams, during which each coder explained the reasoning behind theme development and definitions. The themes were then reviewed to ensure they were sufficiently distinct from one another in terms of assigned codes. This process sometimes involved merging two or more themes into one, splitting a theme into multiple themes, or adjusting the codes assigned to each category. Within each theme, codes remained centered on the same topic, while across themes, each was kept unique and sufficiently distinct.
Defining and naming themes: Naming each theme was carried out by AK, while the process of providing definitions and clarifying the overall story conveyed by the themes was undertaken jointly by AK and WX.
- 5) **Producing the report:** A detailed description of each theme was provided by AK in this thesis, including illustrative quotes and an explanation of their broader implications.

This analytic process not only structured the data but also transformed it into meaningful themes that captured participants' perspectives on trust and glaucoma screening. The emerging themes

generated through this process provide the foundation for the findings presented in the next chapter, where they are discussed in detail and interpreted in relation to the research question.

2.7.2 Decision Ladder Framework

This study employed decision ladder analysis based on the cognitive decision-making framework originally developed by Rasmussen [113] and later improved by Vicente [114]. This approach was used to provide a structured, visual depiction of the reasoning steps and knowledge states optometrists move through when evaluating the performance alignment between the AI system and their own judgment, a process that ultimately help shaping their trust in the system.

A schematic of the decision ladder is presented in Figure 6. The framework includes two categories of nodes: rectangular nodes indicate the cognitive or information-processing activities performed by the decision-maker, while circular nodes denote the resulting knowledge states. The process begins when an external alert in the environment triggers the decision-maker to react to a situation. The left side of the ladder represents the stage of situational assessment, where individuals collect information and identify possible courses of action in response to the initial cue. Once the most suitable response is determined through knowledge-based reasoning, the right side of the ladder illustrates the planning and execution steps that follow.

Although the framework is depicted sequentially, experienced decision-makers often bypass certain steps to reach conclusions more efficiently. These cognitive shortcuts are depicted in two forms. A **leap** is a shortcut that connects directly between two knowledge states (circle to circle), imply that one knowledge state triggers another without any explicit intermediate processing. A **shunt** is a shortcut where an information-processing activity (box) leads to a knowledge state (circle), in other words: some processing is done, then a knowledge state is achieved, but it skips one or more of the intermediate steps. The structure presented in Figure 6 is adapted from Rasmussen 1974 [113].

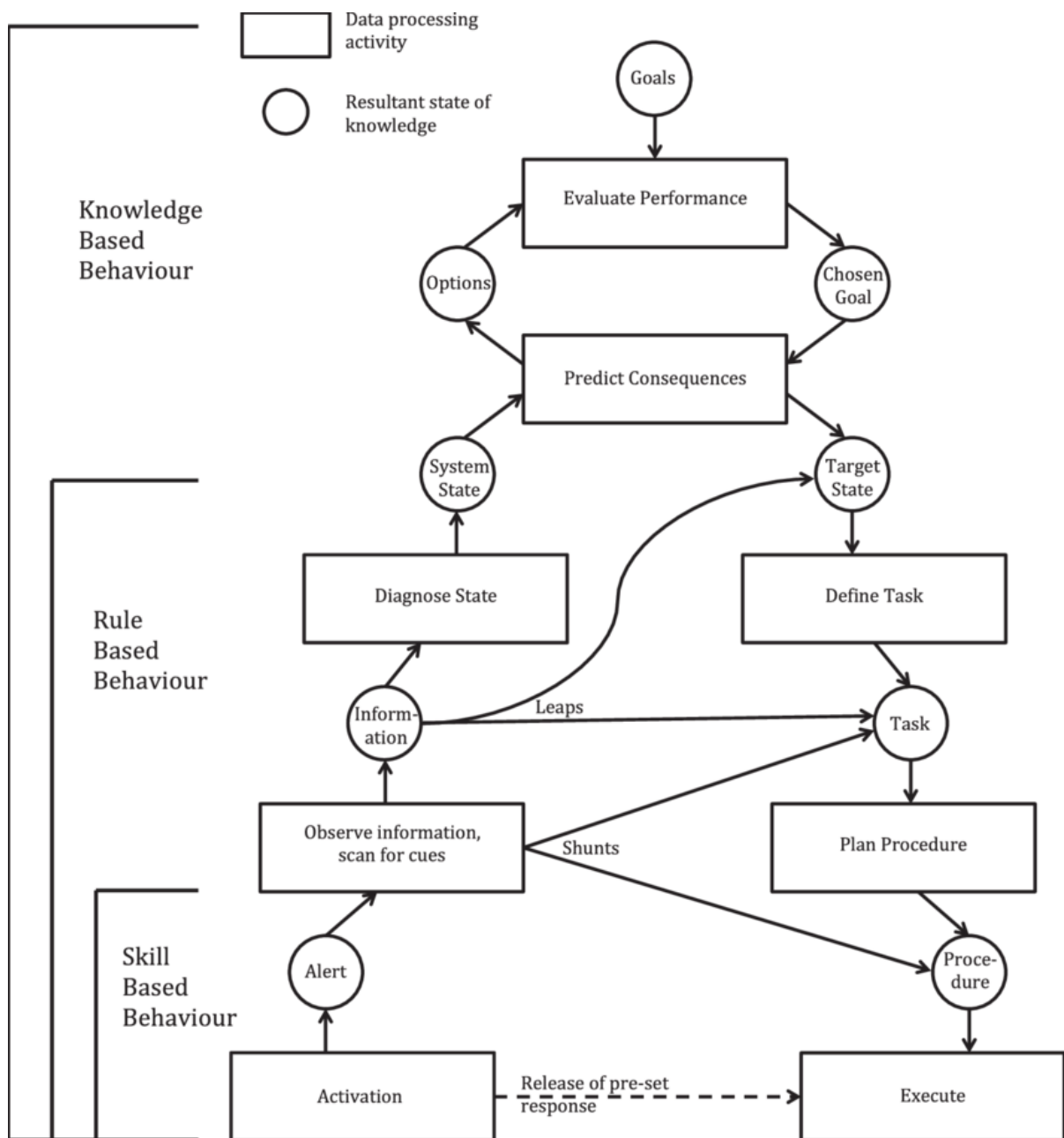


Figure 6- Decision Ladder Template

Chapter 3

Results

3.1 Overview

This chapter presents the results of the study, specific to the thematic analysis of trust. It starts with a summary of the participants' demographic characteristics, including their gender identity and clinical experience. Following this summary, the chapter reports findings from the qualitative analysis of participant interviews. These findings are organized into thematic categories that reflect the participants' trust forming factors in glaucoma screening clinical decision support system. The themes include expected AI usage, more data requirements, model performance, clinical relevance, explainability and transparency, decision alignment, decision philosophy, usability, self-perceived expertise, training, and familiarity through continued use. Each section is supported by illustrative quotes to capture participants' perspectives in their own words.

3.2 Participant Demographics

The study included 20 participants, one of whom (P4) was excluded from analysis totally due to missing information. The final analysis of this study included 19 participants. As noted in Table 4, of the 20 participants, 50% (n=10) identified as female, 40% (n=8) as male, and 10% (n=2) as trans or non-binary (Figure 9). Most participants reported Asian backgrounds including South, East, West, including Iranian and Middle Eastern, 50% (n=10), followed by Caucasian/European, 45% (n=9), with 4% (n=1) identifying as “Other” (Figure 7). The group also spanned a broad age range: 45% (n=9) were between 20–30 years, 25% (n=5) between 30–40, 15% (n=3) between 40–50, 5% (n=1) between 50–60, and 10% (n=2) chose not to disclose their age (Figure 8Figure 8– Participants' Age Distribution).

As mentioned in Table 5 participants had an average of 6.85 ± 8.24 years of clinical experience as practicing optometrists, with a range from 0.5 to 25 years (Figure 12). On average, participants saw 59.1 ± 25.4 patients per week ranging from 8 – 100 (Figure 11), including 8.18 ± 6.27 glaucoma patients, with weekly glaucoma screenings ranging from 0 to 20 patients (Figure 10). This participant pool represented a broad spectrum of clinical experience, making them well-positioned to reflect on diagnostic workflows with the integration of AI into glaucoma screening practices.

Appendix A provides full demographic information for each participant.

Table 4- Participants' Demographic Distribution

Gender Identity	Percentage (n)
Male	40% (n = 8)
Female	50% (n = 10)
Non-Binary	10% (n = 2)
Ethnicity	Percentage (n)
Asian (South, East, West, Iranian, Middle Eastern)	50% (10)
European/Caucasian	45% (9)
Other	4% (1)
Age	Percentage (n)
20-30	45% (9)
30-40	25% (5)
40-50	15% (3)
50-60	5% (1)
N/A	10% (2)

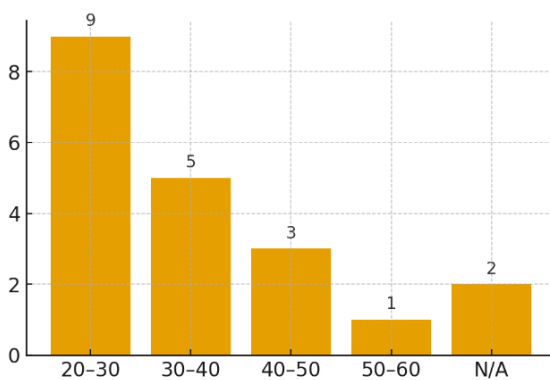


Figure 8- Participants' Age Distribution

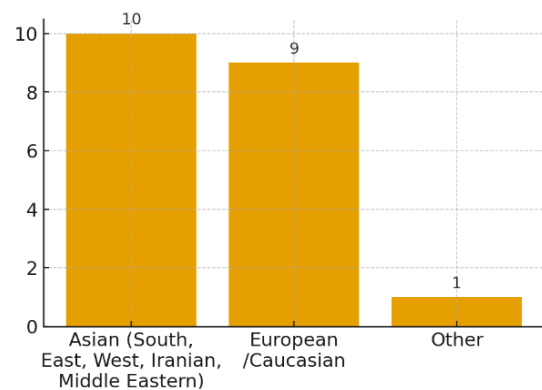


Figure 7- Participants' Ethnicity Distribution

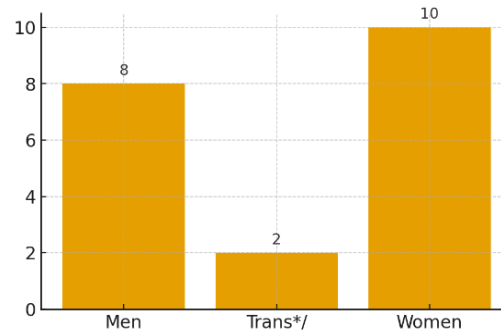


Figure 9- Participants' Gender Distribution

Table 5-Participants' Clinical Experience Metrics

Metrics	Mean ($\pm$ SD)	Range
Years of practicing as an optometrist	6.85 ($\pm$ 8.24) years	0.5 – 25 years
Average number of patients seen per week	59.1 ($\pm$ 25.4) patients/week	8 – 100 patients/week
Average glaucoma patients screened per week	8.18 ($\pm$ 6.27) glaucoma patients/week	0 – 20 glaucoma patients/week

The graphs below illustrate the distribution of participants' years of experience, average number of patients seen per week, and average number of glaucoma patients screened per week. Although the sample covers a range of clinicians, it is skewed toward women, those with fewer years of experience, and those with lighter glaucoma exposure, while the average number of patients seen per week is relatively normally distributed. This imbalance should be taken into account when interpreting the findings, as the results may disproportionately represent the perspectives of earlier-career optometrists. The skewness values shown above each histogram are derived from Fisher's moment coefficient of skewness.

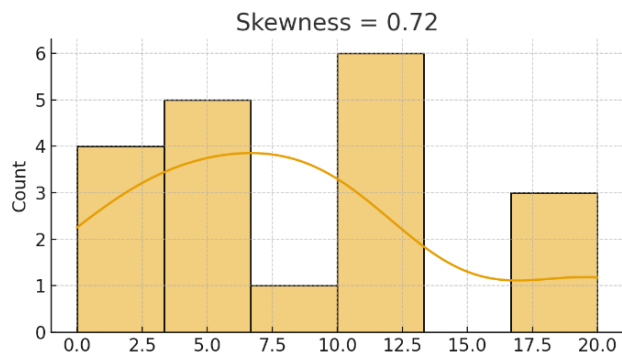


Figure 10- - Participants' Distribution of Average Number of Glaucoma Patients Seen per Week

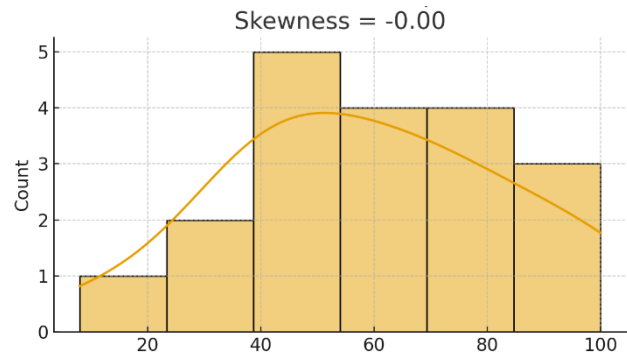


Figure 11- Participants' Distribution of Average Number of Patients Seen per Week

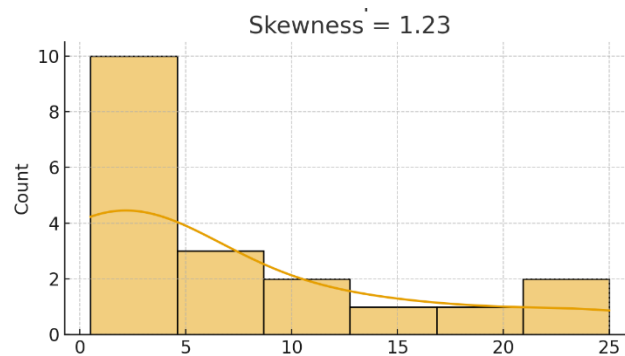


Figure 12- Participants' Distribution of Years of Experience

3.3 Qualitative results

Thematic analysis of the interview data revealed a set of recurring themes that reflect key factors influencing clinicians' trust in AI-based diagnostic systems. These themes capture a range of considerations, from practical system-level concerns to deeper philosophical and experiential influences on clinical decision-making. Table 6 presents the emerging themes along with the number of supporting codes associated with each. The emerging themes are expected AI usage, more data, decision alignment, explainability and transparency, model performance, familiarity through continued use, clinical relevance, decision philosophy, self-perceived expertise, training, and

usability. Together, these themes provide a comprehensive view of the trust-building landscape and inform design considerations for future AI-enabled clinical tools.

Table 6-Emerging Themes Overview-ordered by frequency

Emerging Themes	Number of Codes
Expected AI Usage	62
More Data	39
Decision Alignment	34
Explainability and Transparency	23
Model Performance	17
Familiarity Through Continued Use	8
Clinical Relevance	6
Decision Philosophy	5
Self-Perceived Expertise	4
Training	3
Usability	2

The order of emerging themes presented here is based on the number of codes associated with each theme. This approach, known as frequency analysis, reflects the relative importance of each emerging theme.

3.3.1 Expected AI Usage

This theme highlights that the way an AI system is introduced to clinicians must align with their expectations and clinical workflow. A key finding from this study is that clinicians consistently view AI as a supplementary tool rather than independent decision-maker. This supplementary role appears in two main ways. First, as a “test result flagging tool”, the AI flags suspicious test results that may indicate glaucoma, early in the diagnostic process. Second, as a “second opinion”, the AI provides its

own diagnostic conclusion and reasoning after clinicians have formed their initial judgment, allowing them to review or reconsider their decision. Table 7 shows the frequency of these two approaches.

Table 7-Expected AI Usage Approach-Emerging Theme

AI Usage Approach	Number of Supporting Codes
Test result flagging tool	21
Second opinion	18

3.3.1.1 Test Result Flagging Tool

The first approach involves using the AI system as a flagging or screening tool. In this case, the optometrist uses the system during the diagnostic stage, where the AI flags suspicious test results. This allows the optometrist to focus on validating those flagged suspicious test results more carefully, rather than reviewing all test results, thereby saving time.

“Like as a final diagnosis? No [I don’t trust it], but I would trust it to flag things [suspicious test results].”

(Quote from P15)

Although they trust the flagging/screening feature, they want to validate the short-listed suspicious test result, especially in cases of low AI-reported confidence score (73%).

“Everything's going to have errors, every practitioner has errors. It's reasonable to have errors. I feel like I would still be very confident in the system itself. As long as there is an opportunity to review each thing [suspicious test results] as well.”

(Quote from P10)

Optometrists noted that flagging/screening suspicious test results will lead to them saving time and cognitive resources.

“It is helpful that it flags things [suspicious test results] because if you're seeing a lot of patients and, you may not have time to go through all the different trends that we just went over in depth.”

(Quote from P14)

“It would kind of hone my cognitive resources and focus them as opposed to having them spread everywhere.”

(Quote from P10)

Paying attention only to suspicious test results reflects a level of trust in the AI system and a belief in its reasonably accurate performance.

“Personally, I would still want to just review the results myself just because I like to review them, but if I did become familiar with the multimodal system and trust it and [thinking that it] was fairly accurate, then I could see myself, not necessarily having to go through all the visual field, OCT results.”

(Quote from P11).

3.3.1.2 Second Opinion

Optometrists in this group insist on making their own diagnostic judgment first, without looking into AI diagnosis and corresponding explanation, based on the raw test data and the complete patient record, before consulting the AI's output. The tendency to review test results is particularly evident with unimodal systems, where the AI does not consider enough of the relevant test results.

The importance of combining AI's output with the optometrists' clinical interpretation is highlighted in the following quotes.

“So, I think you need to integrate it with what you're seeing as well, because it is still a machine”

“I like the fact that we can marry just our own viewing of the information as well as what their [AI system] opinion is”

(Quotes from P20)

Some optometrists are skeptical of the AI system and believe test results should still be evaluated independently by the optometrist.

“You know, I partially would [trust it], but what I also don't trust, I think it's important for doctors to still interpret the data themselves.”

(Quote from P12)

“I would trust to be using a clinical environment as long as it was being reviewed by a practitioner.”

(Quote from P10)

This skepticism is often the case with unimodal systems, where the AI does not consider enough of the test results. Optometrists want to evaluate the results that the AI system has overlooked.

“Yes [I trust unimodal], as long as the practitioner makes sure to use other [test results], like OCT, visual field as well.”

(Quote from P14)

3.3.2 More Data

This theme does not include sub-themes. It is generally recognized that AI systems incorporating multiple data sources (multimodal systems) are perceived as more trustworthy than those relying on a single data source (unimodal systems). In this study, trust in unimodal systems varied, from distrust to hesitant or conditional trust as shown in Table 8. However, regardless of these differences, optometrists consistently expressed greater trust in multimodal systems.

Table 8- Variation in Trust Toward Unimodal AI Systems

Variation in Trust Toward Unimodal AI Systems	Number of Supporting Codes
Conditional trust in unimodal	1
Hesitant trust in unimodal	4
Trust unimodal < trust multimodal	27
Untrustworthy unimodal	4

Trust in unimodal systems appears to be context/optometrist dependent, with some users showing hesitation or conditional trust, while others view unimodal systems as outright untrustworthy.

P20 finds the unimodal system trustworthy despite knowing it does not incorporate all available information and still uses it as a second opinion.

“I would trust it to be used in a clinical environment, knowing that it's not incorporating all the information that I have. It may give you that extra opinion.”

(Quote from P20)

P19 and P7 express hesitant trust in the AI system, recognizing that it does not account for all the information they consider in their clinical practice.

“Not too confidently [trusting it], again, doesn't give us more information. It doesn't take OCT and Visual field analysis plus the other general health factors and the gonioscopy into play.”

(Quote from P19)

P1, P6, and P12 find the unimodal system untrustworthy for the same reason: it does not take enough information into consideration.

No, [I don't trust unimodal]. Because it's just the photo and I need more.

(Quote from P1)

3.3.3 Decision Alignment

Decision alignment refers to the degree of alignment between the AI's output and the optometrists' clinical judgment. This theme is characterized by the presence of agreement, the extent of that agreement, and the specific aspects it pertains to. It is defined by seven sub-themes, the first five, representing different dimensions of agreement: agreement with the AI screening output, alignment with the decision input, alignment on identified risk factors, agreement with the AI's recommendation, and agreement with the AI's reasoning. Sixth and seventh sub-themes capture other agreement related topics. These sub-themes are summarized in Table 9.

Table 9- Decision Alignment-Emerging Theme

Theme	Sub-theme	Number of Codes
Decision Alignment	Agreement with AI screening output	10
	Decision Input alignment	9
	Identified risk factor alignment	4
	Recommendation agreement	1
	Agreement with AI reasoning	6
	Cumulative Impact of Agreement and Disagreement	2
	Agreement Overrides Performance Metrics	2
Total		34

3.3.3.1 Agreement with AI Screening Output

One of the key dimensions of agreement, and the most frequently reported by optometrists, is agreement with the AI's final diagnosis. In general, when optometrists agreed with the AI's screening output, their trust in the system increased. Trust could also build gradually if the AI's screening output aligned with the optometrist's judgment in most cases. The extent of agreement also matters which led this thesis to introduce the concept of a concordance gradient, a scale reflecting the degree of difference between the AI's screening output and the optometrists. To contextualize this, Table 10 outlines the various categories of glaucoma diagnosis used in this study. For example, if the AI labels

a case as "glaucoma suspect" and the optometrist labels it as "glaucoma medium," this would represent one degree of disagreement. However, if the AI classifies a case as "glaucoma positive, early stage" while the optometrist identifies it as "glaucoma suspect, medium," this would represent two degrees of disagreement, and vice versa.

Table 10- Screening Output Categories

Glaucoma negative	Glaucoma suspect			Glaucoma positive		
-	Low	Medium	High	Early	Mild	Advanced

Table 11 illustrates how the degree of disagreement between the AI screening output and the optometrist's judgment correlates with reported trust levels. When there was only one degree of disagreement, four instances were recorded, two were associated with trust in the AI system, while the remaining two indicated lower trust levels. In cases with two degrees of disagreement, three instances occurred, leading to a decline in trust overall: two as untrustworthy, one case was marked as less trusted. A single instance of three degrees of disagreement was reported, which resulted in hesitant trust.

Table 11- Degrees of Disagreement

Degree of Disagreement	Number of instances	Trust Levels
One degree of disagreement	4	2 trustworthy
		1 less trusted
		1 hesitant trust
Two degrees of disagreement	3	1 less trusted
		2 untrustworthy
Three degrees of disagreement	1	1 hesitant trust

3.3.3.2 Decision Input alignment

When AI considers the same inputs, it contributes to building trust in the system. Also, multimodal systems are considered more trustworthy in comparison with unimodal because their decision inputs more closely align with those of optometrists.

“[I trust it] because it [AI system] looks at the same factors that are important for diagnosis as a clinician.”

(Quote from P9)

“I trust the multimodal system more [than unimodal] just because that is aligned better with how my clinical practice is.”

(Quote from P8)

3.3.3.3 Identified Risk Factor Alignment

The difference between alignment on identified risk factors and decision input alignment lies in the level of interpretation. The former refers to whether the AI identifies the same risk factors as the clinician, while the latter concerns how the AI interprets those inputs as risk factors or not, based on their contribution to the final screening output. It is important to note that not all decision inputs are risk factors, decision inputs may include normal findings as well. Therefore, alignment on risk factors is a specific subset of decision input alignment, focused only on clinically significant indicators.

“I think so [trusting AI system] ... all the risk factors are there; it just doesn't comment on the rim tissue which is super important.”

(Quote from P1)

“It does look like it is picking up on cup-to-disc-ratios in the way [as a risk factor] that I would agree.”

(Quote from P2)

3.3.3.4 Recommendation Agreement

Within the domain of decision alignment, agreement with the AI's recommended follow-up time serves as an example of trust building factor.

3.3.3.5 Agreement with AI Reasoning

This sub-theme concerns the level of agreement between the AI's assessment of how much a given risk factor contributes to the final screening output (as represented by feature importance percentages in explainable AI models) and the optometrist's own perception of that factor's significance.

Optometrists tend to compare the AI model's explanations with their own clinical judgment to validate the AI's reasoning and build trust, as reported by P15. When optometrists disagree with the AI's reasoning, it tends to lower their trust in the system and increases their need to independently review the patient's information.

"I think for the most part of this case, yes [I trust it], but, in terms of the weights of the risk factors for glaucoma, it's a little bit different than how I would weigh it. So, I would have to review it."

(Quote from P18)

On the other hand, presenting AI-weighted risk factors contributing to the final diagnosis may help resolve disagreements between humans and AI by enabling granular validation of each risk factor's effect.

"If I'm disagreeing, but I see that it is heavily weighted on one thing, but I just disagree on another thing, then maybe I agree with everything else that AI said. Because it is heavily weighted on one factor which I don't agree with how it interpreted, I can kind of go through the other things to see if I agree with the other ones."

(Quote from P5)

It is worth noting that in the absence of feature importance explanations, one optometrist (P18) assumed that the AI followed a decision-making process equivalent to their own.

3.3.3.6 Cumulative Impact of Agreement and Disagreement

The number of disagreements and agreements with the AI output influenced trust differently across participants. For example, P7 reported that a single instance of disagreement would not impact on their trust if the AI had aligned with their judgment in previous cases. In contrast, P1 noted that a later disagreement could diminish the trust that had been built during earlier interactions.

3.3.3.7 Agreement Overrides Performance Metrics

Although AI-reported model performance metrics, such as accuracy and AI confidence level, are considered important (e.g., P20 notes their relevance when experts disagree with the AI), trust is more strongly influenced by the level of agreement with the AI's output than by the system's reported confidence levels, as highlighted by P1.

3.3.4 Explainability and Transparency

Explainability and transparency emerged as a theme. It reflects the importance of enabling users to understand how the AI system works and makes decisions. However, transparency is expected at varying forms and in different styles, which influences optometrists' trust. These variations are captured in the sub-themes outlined in Table 12.

Table 12-Explainability & Transparency- Emerging Theme

Theme	Sub-theme	Number of Codes
Explainability and transparency	Risk Factor transparency	11
	Explanation style	8
	Perception of explanation output	2
	Recommendation transparency	1
	Not agreeing with AI reasoning	1
Total		23

3.3.4.1 Risk Factor Transparency

Risk factor transparency emerged as the most prominent sub-theme. It was considered more important than other forms of transparency, such as final diagnosis transparency and recommendation

transparency. This sub-theme involves flagging AI-identified risk factors or highlighting suspicious test results for optometrists to review, enabling faster validation of the AI output and saving time by avoiding the need to examine all test results, if AI is trusted.

“...with multimodal system, I can look at the summaries for all of them and then if it's flagging something as being a diagnosis, then I can go into it and analyze it a little bit quicker and kind of use it more as a screening tool. That way, as opposed to unimodal, going through and analyzing everything with its raw data. It would hone my cognitive resources and focus them as opposed to having them spread everywhere.”

(Quotes from P10)

“If it doesn't give me that breakdown, then I have to look at every single thing, and then it doesn't really save me that much time.”

(Quote from P5)

Interestingly, one participant did not necessarily trust the final AI's screening output but still found the AI helpful in flagging suspicious factors that contributed to the screening output .

“I wouldn't trust it necessarily to make the final diagnosis, but I think it could be helpful in terms of pointing out to features of the nerve that are suspicious for glaucoma, to kind of help, and give me a second eye almost on things”

(Quote from P8)

3.3.4.2 Explanation Style

Explanation style is another emerging sub-theme that affects trust. It encompasses how explanations should be delivered to the optometrist to build their trust in the AI system. Communicating missing input data helps optometrists calibrate their trust, as illustrated by two participants. P11 noted that if the missing data (fundus photo) is not as critical as the information already considered by the AI system (OCT), their trust in the output is maintained.

“Even though the fundus images aren't there, I would still trust the system because it is analyzing from the OCT result”

(Quote from P11)

P7 emphasized that being aware of missing input data, such as fundus images, not only helps them understand why the AI's output might be limited but also reassures them that the system is making decisions based on available, reliable information rather than potentially low-quality inputs.

“Do I agree with the glaucoma diagnosis? No. Do I trust the system? Yes, to somewhat, because I think it's making it based on the data available [fundus photo is not available]”

“I trust it more if there were fundus images, definitely I would, but I also understand the app for various reasons, it's very possible, like just being clinician, they tried to get fundus photos at this visit and they weren't accurate. And you don't want to feed the system garbage data.”

(Quotes from P7)

An optometrist emphasized explaining how each factor was interpreted in forming the diagnosis would lead to trusting the AI system more confidently.

“I would say if it showed me how they interpret that information, I'll be more confident using them”

(Quote from P19)

One approach to explaining the AI model output is by presenting AI-weighted risk factors that contribute to the final screening output. This method can help resolve disagreements between the AI and the optometrist by enabling granular validation of each factor's influence.

“The other thing is like, if I'm disagreeing, but I see that it is heavily weighted on one thing, but I just disagree on another thing. Then, maybe I agree with everything else that AI said because it is heavily weighted on one factor which I don't agree with how it interpreted that, I can kind of go through the other things to see if I agree with the other one.”

(Quote from P5)

Optometrists expressed a preference for granular yet concise interpretations of individual test results. They valued one-line AI-generated insights for each specific test, as these were quick to review and did not require much time.

“The multimodal one was nice because for every single stage, there's just a quick one-line answer where it tells you what the interpretation the AI is getting from it.”

(Quote from P5)

3.3.4.3 Perception of Explanation Output

The optometrists emphasized that trust increases when the AI not only displays all contributing risk factors, beyond just the top-ranked ones. This reflects a preference for full transparency in the diagnostic rationale, enabling experts to independently assess the relevance of each factor, regardless of the feature importance rankings generated by explainable AI models. A participant assumed that the AI had overlooked key patient information, such as medical or family history, because it wasn't visibly presented in the output. In reality, the system had considered this data but only displayed the top contributing factors, which don't include medical history. This led to confusion and reflected how limited explanations can make users feel like important context is missing, potentially affecting their trust in the system. It underscores the importance of making explanation outputs more transparent and complete to avoid misunderstandings.

3.3.4.4 Recommendation Transparency

An optometrist expressed greater trust in the system when the AI provided reasoning behind its recommendations, indicating that transparent explanations of why a specific recommendation is made can contribute to building trust.

3.3.4.5 Not agreeing with AI Reasoning

Although presenting AI-weighted risk factors that contribute to the final diagnosis would help optometrists build trust, it can act as a double-edged sword: if the AI's feature importance weights do not align with the optometrist's clinical judgment, particularly when a feature is perceived as undervalued, it may lead to decreased trust in the system. It is important to note that, as mentioned in

the relevant section, the feature-importance weights presented here are local, meaning they vary from one visit to another. Consequently, their influence on trust is also dynamic, as local explanations evolve with each patient’s data and visit over time.

3.3.5 Model Performance

Perceived performance of the AI model emerged as another key theme influencing participants’ trust. This theme encompasses five sub-themes: confidence score, perceived sensitivity, perceived accuracy, perceived consistency and, perceived data input quality. Table 13 summarizes the prominence of each sub-theme based on the frequency of reports in interviews.

Table 13- Model Performance -Emerging Theme

Theme	Sub-theme	Number of Codes
Model Performance	Confidence score	8
	Perceived sensitivity	5
	Perceived accuracy	2
	Perceived consistency	1
	Perceived input quality	1
Total		17

3.3.5.1 Confidence Score

The confidence score was ranked first within the model performance theme. Confidence score is a numerical value produced by a model that reflects its estimated certainty in a given prediction, typically derived from the model’s output probabilities. Unlike accuracy, which is an evaluation metric calculated after comparing predictions against ground truth across many cases.

To provide some context, the confidence scores across different visits for two patient cases were 73%, 86%, 92%, and 97%. All were color-coded in green and reported alongside the final AI screening output for each visit. Participants expressed a range of sentiments toward the 73% confidence score, spanning from perceiving the AI system as untrustworthy (reported by one optometrist) to exhibiting hesitant or conditional trust (reported by two optometrists).

An optometrist (P1) found the AI system untrustworthy when it did not indicate its confidence, by reporting a low score of 73%. Other optometrists (P18 and P20) stated that a drop in the confidence score from 92% to 73% led them to have conditional trust in the AI system, if it was used as a assistive tool requiring optometrist validation, either by themselves if they were comfortable with glaucoma diagnosis (P18) or by another optometrist (P20). To add context, P20 had limited knowledge of OCT and limited experience with glaucoma diagnosis, which led to over-trusting the AI system. However, in cases with lower confidence scores (73%), they preferred to have another optometrist validate the AI's diagnosis and interpretation of the OCT test results.

“I think because its confidence level is lower, mine is lower and having what looks like a change in the OCT, I'm blindly trusting that it's accurate. I would want somebody who knows OCT to comment on the accuracy of [AI system] reading an OCT before I bought into the system.”

(Quote from P20)

Our study also has limited instances of confidence scores, as the scope was focused on identifying factors that affect trust, rather than examining how different confidence levels influence optometrists' trust. However, as an interesting finding, P1 stated that they would only trust the AI system if the confidence score were above 90%.

Another optometrist (P14) perceived the 73% confidence score as a sign that the AI system is not as adaptive as it should be to different cases, which raises concerns about the overall accuracy of the AI system. This connects to the next sub-theme under the model performance theme.

Additionally, the level of agreement with the AI diagnosis was identified as a trust-building factor (discussed later in Section 3.3.3). The relationship between the communicated AI confidence score and decision alignment is particularly interesting. P1 stated that their trust was driven more by agreement with the diagnosis than by the AI's confidence score:

“[I don't trust AI system] Not with that confidence [score, 73%]. I mean the output [the AI diagnosis] is correct. So, I suppose yes [I trust AI system].”

(Quote from P1)

3.3.5.2 Perceived Sensitivity

Among the sub-themes under model performance, perceived sensitivity was the second most frequently reported. Participants referred to sensitivity in terms of whether an AI system could outperform humans in detecting negative changes in test results. In machine learning and statistics, sensitivity is formally defined as the proportion of actual positives that the system correctly identifies. While participants did not use this technical terminology, their descriptions align with the broader notion of an AI's ability to reliably detect subtle or adverse changes that humans might overlook.

P20 believed the AI system was trustworthy as it picks up subtle differences that a human optometrist may have overlooked, and it was reported three times in different patient visits by P20. P11 also agrees with them that identifying small clinical changes and identifying the most diagnostic changes would lead to trust.

P12 challenged the AI system by stating that the AI system should be able to diagnose non-obvious patient cases where there could be a high level of uncertainty among optometrists:

“But if it's obvious to the AI, it should be obvious to a licensed and properly trained eye doctor to make the diagnosis ... It's the non-obvious stuff that I think like these cases where it gets a bit trickier.”

(Quote from P12)

3.3.5.3 Perceived Accuracy

Two codes referred to accuracy, making it the third most frequently reported sub-theme under model performance. In general, accuracy as a model performance metric is perceived as a factor in building trust. The term accuracy was mentioned by participants in this study and is assumed to refer to its standard definition in machine learning and statistics, namely the proportion of correct predictions (true positives and true negatives) among the total number of predictions.

P20 believed that accuracy is important when optometrists disagree with the AI. It is projected that, in case of disagreement with the AI diagnosis, high accuracy would make them reconsider their diagnosis with more caution.

Also, optometrists will cross-check the AI system's diagnosis with their judgment over time to form a perceived accuracy of the AI system, as noted from P11. However, the accuracy can also be communicated by the AI model.

3.3.5.4 Perceived Consistency

Perceived consistency emerged as the fourth most frequently mentioned sub-theme within model performance. For example, P2 highlighted the importance of the AI providing consistent analyses across cases with similar clinical manifestations.

3.3.5.5 Perceived Input Quality

The input data quality is the fourth most frequently discussed sub-theme under model performance. The quality of input data used to train the model is perceived as a trust building factor as noted by P15.

“Now this current system [unimodal] is based on fundus images and that's dependent on the technician or the person that's taking these photos. So, let's say, they're not trained well, or the photo is off axis; there's a little bit of haze, give or take, maybe there's a floater on top. That could skew the AI a little bit.”

(Quote from P15)

3.3.6 Familiarity Through Continued Use

Familiarity through continued use highlights how trust in the AI system is not always immediate but develops gradually through repeated use and familiarity. While some optometrists (P3, P15, and P17) expressed initial skepticism or hesitation, trust tended to increase as they interacted with the system more frequently and became familiar with its outputs and behavior. For several optometrists (P2 and P12), trust grew as they observed the AI making recommendations that aligned with their own clinical judgment. The process of gaining experience with the tool, especially when it consistently demonstrated accuracy and alignment with optometrists' judgment played a critical role in transforming early doubt into trust. This underscore repeated interaction and experiential learning as key trust-forming factors in clinical decision support systems.

“Just based on the number of years of practice, there's a level of confidence that I can have for myself, and I guess if it's used over time and it seems to agree [with me] more so then, it would increase my confidence.”

(Quote from P2)

“Personally, I would still want to just review the results myself just because I like to review them. But if I did become familiar with the multimodal system and trust it and that was fairly accurate, then I could see myself, not necessarily having to go through all the visual field, OCT results.”

(Quote from P11)

3.3.7 Clinical Relevance

Providing expected clinically relevant output from AI system is another emerging theme in forming trust in a glaucoma diagnosis AI system. It includes three sub-themes, which are actionability, trend analysis and, system output expectation.

Table 14- Clinical Relevance -Emerging Theme

Theme	Sub-themes	Number of codes
Clinical Relevance	Actionability	3
	Trend Analysis	2
	System Output Expectation	1
Total		6

3.3.7.1 Actionability

This sub-theme is the most frequently reported sub-theme within the clinical relevance theme.

Actionability is defined as the extent to which AI suggests practical action items that enhance patient care or improve health outcomes.

One optometrist (P6) stated that they find the AI system trustworthy if it provides recommendations. Additionally, (P15) emphasized that not only the presence of recommendations but also their perceived usefulness plays an important role in increasing trust.

Interestingly, an optometrist (P1) stated that a definitive diagnosis (either glaucoma diagnosed or not diagnosed) is preferred over a suspect diagnosis. This preference may stem from the fact that a definitive diagnosis allows the optometrist to act, whereas suspect cases usually do not require significant intervention, highlighting the importance of the diagnosis's actionability.

“Yeah, because the most important thing is the diagnosis. Anyone can call them a suspect, but the most important thing is, is it correct? When we finally say it's glaucoma or not.”

(Quote from P1)

3.3.7.2 Trend Analysis

Trend analysis sub-theme is the second most frequently reported sub-theme under the clinical relevance theme. Trends are considered more important than individual data points, as noted by an optometrist (P11), who stated that every patient has a baseline condition, and the AI should be able to compare the current condition with that baseline to draw conclusions. Additionally, not accurately reporting risk factor histories, such as previous hemorrhage, was identified by (P11) as a reason to perceive the AI as untrustworthy.

“The three-drance hemorrhage history should probably be in risk factors, and if it was, then that would probably equate to a trustworthy diagnosis for glaucoma.”

(Quote from P11)

3.3.7.3 System Output Expectation

This sub-theme highlights the expected output of diagnosis. One optometrist (P12) stated that a differential diagnosis, such as distinguishing between melanoma and a birthmark freckle, would be an interesting output for the AI system.

“So, if there's a retinal lesion on a fundus, and the AI can give me differentials of hey, is this a melanoma or is this something like just a birthmark freckle? I think that would be mentally helpful to a lot of doctors.”

(Quote from P12)

3.3.8 Decision Philosophy

Optometrists indicated that the glaucoma AI system, designed as a screening tool, should prioritize high sensitivity. Participants explicitly stated a preference for more false positives over false negatives, emphasizing caution in decision-making (P12, P16). Two related statements were noted:

- **Medico-legal risk:** It is more defensible to be cautious than to miss a diagnosis (P12).
- **Under-diagnosis and misdiagnosis** are primary reasons for a lack of trust in AI-powered clinical support screening tools (P12).

3.3.9 Self-Perceived Expertise

Self-perceived expertise captures how clinicians' level of expertise influences their trust in the AI system. An optometrist noted that less experienced clinicians might either second-guess themselves excessively or place too much confidence in the system, potentially overlooking important clinical judgment. This highlights the risk of disrupted clinical decision-making when the user lacks sufficient experience.

“I think...if they [users] are over reliant or —I don't want to say expert, but in a way, if they're not expert in their area— then this might open a can of worms where they suddenly will self-doubt and become overly cautious about something, or they become under-cautious about it.”

(Quote from P12)

An optometrist (P20) with limited experience in glaucoma diagnosis or interpreting OCT results reported relying more heavily on the AI's recommendations. While AI can support decision-making, it also raises concerns about overreliance, particularly if the system makes an error.

3.3.10 Training

Due to the limited information about how the AI system operates, as reported by P12, training on its use is seen as essential. Both P12 and P17 mentioned that training would likely increase trust. Maybe by improving understanding of the system's scope and functioning, which in turn could lead to greater AI transparency.

“So, I can trust AI, but without proper training I can't.”

(Quote from P12)

3.3.11 Usability

Optometrists found the AI more trustworthy when it presented information in a condensed, well-organized, and easy-to-read format. The clarity and neatness of the output were seen as enhancing usability, leading to trust.

Chapter 4

Discussion

4.1 Overview

This chapter opens with an overview of the study objectives and research questions, followed by responses to the research questions, a discussion of the study's strengths and limitations, and recommendations for future research.

The primary objective of this thesis is to examine the factors that shape trust in AI-powered glaucoma screening tools and to propose design recommendations that enhance optometrists' trust. Building such trust is essential for supporting the adoption of these tools in clinical practice and fostering their widespread integration into healthcare environments. Recalling from Chapter 2, the research questions were as follows:

RQ1) What key factors influence optometrists' trust in AI-powered clinical decision-support systems (CDSS) for glaucoma screening?

RQ2) How do optometrists evaluate the performance of glaucoma-focused AI CDSS, and how do those evaluations translate into varying levels of clinical trust?

RQ3) What design principles should guide the development of glaucoma AI CDSS to foster and sustain optometrists' trust?

4.2 Interpreting Key Factors Influencing Trust (RQ1)

To address RQ 1, I examined how optometrists envision the role of AI-powered glaucoma screening tools within their clinical workflow and what conditions shape their trust in such systems. Each following subsection corresponds to an emerging theme presented in detail in the results chapter and begins with a summary of the key findings, followed by an explanation of why the factor matters, and its potential interactions with other trust-related factors.

4.2.1 Expected AI Usage (Role Framing & Workflow Fit)

Optometrists consistently regard AI as a supplementary tool rather than an independent decision-maker. This role takes two equally important forms. **1)** as a flagging tool, drawing attention only to suspicious test results that may indicate glaucoma early in the screening process. **2)** AI serves as a

second opinion support tool consulted after optometrists have formed their own conclusions. Both forms appear with similar frequency, highlighting their equal significance in how optometrists integrate AI into clinical practice.

Using AI as a supplementary tool is important because, in the author's view, preserving final autonomy for optometrists encourages more responsible use of AI, ensuring clinicians do not simply defer to its outputs. Instead, they remain engaged and accountable for their joint judgments with AI. Supplementing human decision-making with AI also creates a complementary relationship: clinicians can mitigate AI's gaps, such as biases or limitations in handling complex patient cases, while AI can draw attention to aspects of diagnosis that humans may overlook. In this way, both sides reinforce each other.

Considering the two forms of AI use discussed earlier, the **flagging of suspicious test results** helps clinicians better understand AI's reasoning and directs their attention to areas requiring validation, making the diagnostic process more efficient. However, an exclusive reliance on AI-flagged results carries the risk that unflagged but clinically relevant findings may be overlooked. This underscores the need for cautious AI systems that prioritize over-flagging in cases of uncertainty, thereby minimizing the chance of missing suspicious results. The other form, in which **AI is consulted** after optometrists have reached their own decision, appears safer, as clinicians are not influenced by the AI's diagnosis and can instead cross-compare their judgment with the AI's output for additional validation.

The flagging of suspicious test results leading to a AI's final screening result can also be viewed as a form of explainable AI, as it not only directs clinicians' attention to specific areas but also makes the system's reasoning more transparent and interpretable in the screening process.

4.2.2 More Data (Relative Information Completeness)

It is widely acknowledged that AI systems drawing on multiple data sources (multimodal) tend to be perceived as more trustworthy than those based on a single source (unimodal). In this study, trust in unimodal systems ranged from outright distrust to cautious or conditional trust.

Optometrists don't diagnose based on a single test in real life; they cross-reference multiple modalities. A multimodal AI mirrors this workflow, making it feel more trustworthy because it behaves more like a human expert. A unimodal system, by contrast, feels incomplete or

oversimplified. Even if the actual model performance of unimodal and multimodal is comparable, Optometrists often perceive multimodal systems as more accurate, because combining data sources suggests thoroughness and reduces the chance of missing important cues.

It is worth noting that information completeness is inherently fluid across different clinical diagnostic and screening contexts, as it is shaped by the specific workflows and practices followed in each clinical setting.

4.2.3 Decision Alignment (and the Concordance Gradient)

Decision alignment refers to how closely the AI's output matches the optometrist's clinical judgment across five aspects: the final diagnosis, interpretation of decision inputs (normal/abnormal across tests), identified risk factors, their relative importance, and recommended actions. We also introduce a concordance gradient, the categorical distance between the AI's diagnosis and the optometrist's (see results chapter, Section 3.3.3.1; diagnostic categories in Table 10). Alignment serves as a practical validity cue, whereas misalignment raises concerns about calibration and safe use. To avoid redundancy, we discuss empirical results regarding alignment and the concordance gradient in upcoming Section 4.3 (RQ2).

4.2.4 Explainability & Transparency

Model transparency means showing how the AI reached its diagnosis: which risk factors contributed, how much each influenced the final estimate, and what inputs were excluded due to poor quality or unavailability. It also means providing granular interpretations of individual test results, even when a result isn't treated as a risk factor, so clinicians have full context. Finally, the system should explain the rationale for any recommendations and link them to the underlying evidence and patient-specific inputs. The quality of explanations which leads to model transparency, is dependent on how many of the above-mentioned points are included.

Explainability builds trust by making the model's reasoning inspectable. When clinicians can see which risk factors were considered and how each was weighed, they can compare the AI's logic with their own diagnostic reasoning. If they disagree with the AI's conclusion, they can audit each factor, review the evidence, question the assigned weight, and identify missing inputs, to understand exactly why the system reached that result. This ability to look into and compare reasoning increases trust and clarifies when to accept, adapt, or override the AI's diagnosis and recommendations.

As also noted in Section 4.2.3, the same principle applies here: when optometrist's assessments line up with the AI-identified risk factors and their relative weights, trust in the system increases. That agreement signals that the model is noticing the same clinical signals they are, which makes the output feel accurate and useful. When there's a mismatch, the system should make it easy to review each factor. This quick compare-and-review loop helps calibrate trust over time.

4.2.5 Model Performance

In the study, several factors influenced how optometrists perceived and responded to the AI system performance. Confidence scores varied visit by visit; for instance, the lowest score in the current study (73%) triggered mixed reactions ranging from distrust to conditional trust, where validation by the optometrist was deemed necessary. Perceived sensitivity referred to how effectively the AI could detect subtle changes in test results that a human optometrist might overlook, which was valued as a potential advantage of the system. Perceived accuracy developed gradually, as optometrists compared the AI's diagnoses with their own clinical judgments over time to assess how dependable it seemed. Perceived consistency emphasized the need for the AI to deliver consistent results across analyses. Finally, perceived input quality played a central role, as the AI's trustworthiness was strongly linked to the quality of the training data on which it was built.

Although accuracy and confidence scores are distinct performance metrics, they were often interpreted in similar ways by optometrists. Accuracy reflects the overall correctness of the model across many cases, while confidence scores indicate the model's certainty in a specific instance. Yet, both were used by clinicians as signals of trustworthiness. In situations where the AI's output conflicted with their own judgment, a higher reported accuracy or confidence score could prompt clinicians to revisit their initial conclusions more cautiously. Perceived consistency was closely associated with the expectation that identical inputs should always yield identical outputs. Without such stability, screening outcomes may fluctuate unpredictably, thereby undermining optometrist's trust.

Perceived sensitivity was equally important, reflecting the expectation that AI could detect subtle or early changes in test results more effectively than humans. Unlike human optometrists working under time pressure and fatigue, the AI could scan large volumes of signals, benchmark them against patient baselines and extensive reference sets, and do so consistently. This ability was seen as one of its key advantages in supporting clinical practice.

Finally, the perceived quality of the data that is used to train the system. Optometrists saw training data as a proxy for the AI's reliability and generalizability. If the dataset was biased or failed to represent diverse patient populations, outputs risked being unstable or misleading, particularly for underrepresented groups. In such cases, clinicians struggled to tell whether changes in results reflected a real shift in patient status or merely artifacts of poor data quality.

Notably, the link between reported confidence scores or accuracy and alignment of decisions revealed that optometrists placed greater weight on agreement with the diagnosis itself than on the AI's stated performance metrics as noted in Section 3.3.3.7.

4.2.6 Familiarity Through Continued Use

Trust in the AI system emerged over time, built through repeated use and growing familiarity rather than forming instantly. While some optometrists were initially skeptical, their trust increased as they saw the system's outputs consistently align with their own clinical assessments. This process highlights the importance of experiential learning and ongoing interaction in shaping trust.

Gradual trust-building is important because it fosters calibrated trust, allowing clinicians to discern when to rely on the AI and when to question its outputs. This balanced trust helps reduce the risks of over-reliance or misuse and supports the meaningful integration of AI into clinical decision-making.

Although multiple interactions appear important for building trust, the extent to which the number of interactions influences this process remains unclear in this study.

4.2.7 Clinical Relevance

Trust is linked to clinical relevance, which encompasses actionability, trend analysis, and system output expectations. Actionability refers to the extent to which AI provides definitive screening output and practical recommendations that improve patient care. Trend analysis involves comparing a patient's current condition with their baseline over time to draw meaningful conclusions. System output expectations relate to the AI's ability to differentiate between diagnoses when two conditions present with similar signs and symptoms.

The system's components and outputs should align with clinical practice to strengthen optometrists' trust. In clinical settings, data points alone add little value unless they include baseline information and are tracked over time. Regarding actionability, data-driven insights are only meaningful if the AI provides a clearer diagnosis and practical, well-supported recommendations for

next steps, along with an explanation of why those actions are suggested. Finally, a comprehensive screening tool that can differentiate between conditions with closely overlapping clinical manifestations will better reflect real patient–optometrist interactions and more closely resemble daily practice.

Explaining why a recommendation is suggested as the next action item directly relates to explainable AI, too, as discussed in the Section 4.2.4. Providing this rationale helps clinicians better understand and accept the recommendations, thereby increasing the system’s perceived actionability.

4.2.8 Decision Philosophy (Screening vs Diagnostic)

A fundamental principle in clinical practice is that a system’s priorities must align with its function. Screening systems are optimized for high sensitivity (recall), designed to cast a wide net and capture all potential cases. This minimizes false negatives, the most critical error in a screening context, while accepting a higher false-positive rate as a deliberate trade-off to ensure suspected cases proceed to further testing. In contrast, diagnostic systems prioritize specificity to confirm disease and avoid unnecessary interventions.

Our glaucoma screening system reflects this cautious philosophy, which emphasizes that false positives are more acceptable than false negatives. Their views also underscore that caution is more defensible from a medico-legal perspective and that under-diagnosis or misdiagnosis can undermine trust. By aligning the AI’s decision-making with the cautious-by-design role of screening tools, the system builds professional trust and demonstrates clinical responsibility.

In simpler terms, when the system is uncertain, it will classify the case as positive rather than risk overlooking a potential condition by classifying it as negative. This cautious approach aligns with clinical expectations for screening systems, which aim to identify all possible cases for further evaluation rather than provide definitive diagnoses.

Our discussion of sensitivity also aligns closely with the discussion of perceived sensitivity as a performance metric in Section 4.2.5.

4.2.9 Self-Perceived Expertise and Overreliance Risk

Less experienced clinicians may either doubt their own judgment excessively or rely too heavily on the system. For instance, in the current study, a participant with limited experience in glaucoma diagnosis described depending heavily on the AI’s output. This underscores the risk of flawed

decision-making, particularly when users with limited experience rely on AI outputs that may contain errors.

When the AI output carries high uncertainty, the risk of overreliance becomes more critical and should be taken seriously. While perceived model performance, communicated through performance metrics, can support trust, users must also remain aware of their own level of expertise. Decisions should be guided by both the model's performance indicators and the user's self-assessed competence, prompting a more cautious approach when either is lacking.

4.2.10 Training

Optometrists highlighted that adequate training is essential for fostering trust in the AI system. They observed that trust grows when users are well-informed about how to operate the tool, whereas limited understanding of its functionality was viewed as a barrier to trust.

Proper training ensures that clinicians understand both the capabilities and limitations of the AI system. Without knowledge, optometrists may either over-rely on the system without critical evaluation or dismiss it entirely out of uncertainty, both of which undermine safe and effective use. Training therefore helps calibrate trust, striking the right balance between reliance and skepticism.

4.2.11 Usability

Optometrists considered AI more trustworthy when information was delivered in a clear, structured, and concise manner. In a fast-paced clinical context it is important to be mindful of the importance of lowering the mental workload by providing to the point statement while maintaining the important information.

4.3 How Performance Evaluations Translate to Trust (RQ2)

This section addresses RQ2: "How do optometrists evaluate the performance of a glaucoma-focused AI CDSS, and how do those evaluations translate into varying levels of clinical trust?" It describes how clinicians assess the system and how those assessments map to trust. The core signal considered is decision alignment, the fit between the AI's outputs and optometrist' judgments.

Agreement is multifaceted: it refers to the degree of alignment between the AI's output and optometrists' clinical judgment, and it depends on whether agreement is present, how extensive it is within a single visit, how often it occurs across different patient visits, and which specific aspects it

concerns. In practice, several aspects of agreement matter. First, **agreement with the AI's final diagnosis** is central (the extent of this agreement is discussed below); when the system reaches the same clinical endpoint as optometrists, trust tends to rise. Second, **agreement on decision inputs**, for multimodal AI, whether normal/abnormal calls across tests match the optometrists' reading, signals that the model has taken in the big picture. Third, **agreement on risk factors** indicates the AI is attending to clinically significant abnormalities rather than noise. Fourth, **agreement on the importance (weighting) of risk factors** shows the AI is calibrated to place appropriate emphasis on what clinicians consider most consequential. Finally, **agreement on recommended actions** makes the system feel useful and safe to follow, because next steps align with clinicians' plans.

The extent of **agreement** also matters. Accordingly, this thesis introduces a **concordance gradient**, a scale for the categorical distance between the AI's screening output and the optometrist's screening output (Section 3.3.3.1; diagnostic categories in Table 10). Because optometrists may place the same case in different categories in real-world practice, the gradient should be interpreted as an approximation rather than an absolute measure. As summarized in Table 10, cell sizes for each disagreement level were small; accordingly, the observed patterns are suggestive rather than conclusive. For **one-degree** disagreements ($n = 4$), two cases were judged trustworthy and two were less trusted or hesitant. For **two-degree** disagreements ($n = 3$), two were untrustworthy and one was less trusted. For the single **three-degree** disagreement ($n = 1$), trust was hesitant. These observations tentatively suggest that trust declines as disagreement widens, but the sample is underpowered to establish a clear trend. Some disagreement is tolerable: in real-world settings, optometrists may assign the same case to different diagnostic categories. A targeted, adequately powered study focused on disagreement magnitude is needed to test this relationship more robustly.

Overall, **trust tracks agreement more than confidence scores**: when the AI's conclusions align with optometrists' judgments, clinicians tend to rely on the system; when they diverge, optometrists defer to their own judgment rather than the AI's self-reported confidence score. A disagreement over a single patient visit likely lowers trust, but the size of that decrease and its downstream effect on adoption cannot be determined from this study.

4.3.1 Decision Ladder Analysis

A decision ladder was applied to decision-making process in glaucoma screening with AI support, examined through the lens of performance alignment between the AI and optometrists. Decision

ladder methodology was explained in Section 2.7.2. The adapted ladder maps the sequence of cognitive activities involved when optometrists interact with AI-generated outputs, progressing from the initial alert activation to the execution of procedures.

The left side of the ladder represents information-driven reasoning, where optometrists interpret patient data, observe the AI's diagnostic output, and compare it with their own judgment. This process includes reviewing patient test results and medical history to form their diagnosis, as well as examining the AI's final diagnosis and its underlying reasoning, such as decision inputs, risk factors, confidence scores, and recommended actions. After gathering information, optometrists compare their own diagnostic judgment with the AI's output to identify agreement aspects (highlighted in red in the figure).

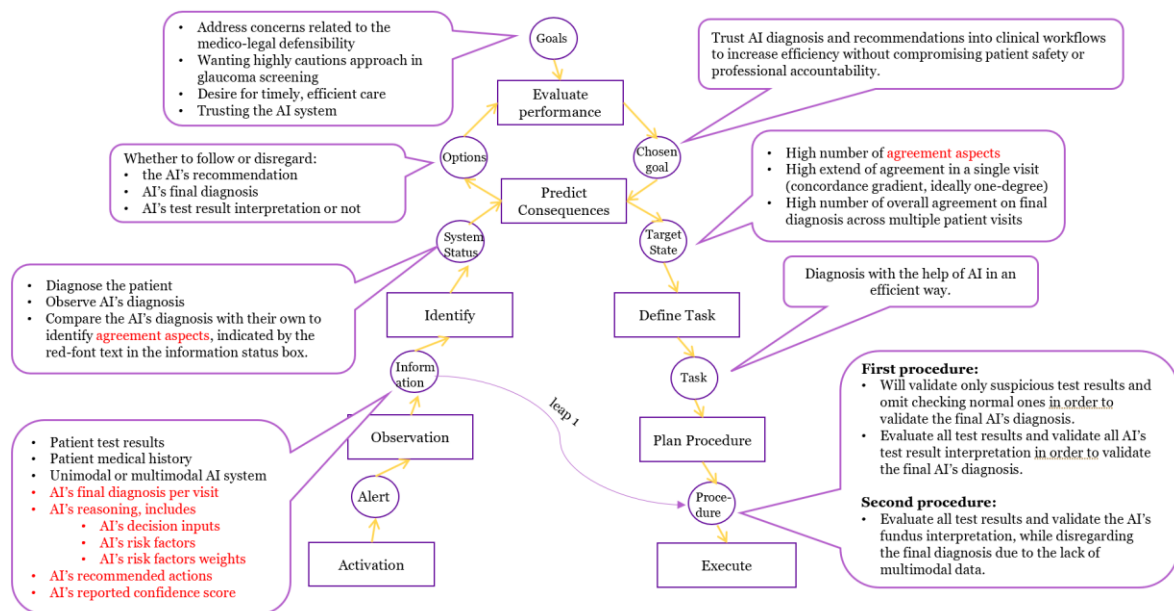


Figure 13- Decision Ladder, Analyzing Decision Making from the Lense of Performance Alignment

The upper part of the ladder depicts reflective evaluation, where clinicians assess whether using the AI system's outputs, including its final diagnosis and recommendations, to enhance diagnostic accuracy, workflow efficiency, and clinical defensibility.

The right side of the ladder represents the planning and execution phase, where optometrists define their course of action based on the chosen goal, trusting a safe and efficient AI tool. At this stage, clinicians want the AI system's reasoning and outputs to align with their own diagnostic practice, which serves as the foundation for building trust. Based on this alignment, they plan how to proceed with the co-diagnosis process in collaboration with the AI.

Two procedural pathways are illustrated.

1. **Procedure one:** Optometrists may validate only AI-identified suspicious test results while omitting normal ones to confirm the AI's final diagnosis or validate all AI test result interpretations. This strategy is typically applied in multimodal AI systems and depends on the established level of trust between the clinician and the system.
2. **Procedure two:** Optometrists evaluate all test results and validate the AI's fundus interpretation while disregarding the system's final diagnosis when only unimodal fundus data are available.

Leap 1 occurs when the optometrist recognizes that the AI system is *unimodal*, focusing solely on fundus photographs. This recognition triggers the second procedure and allows the clinician to bypass several intermediate steps, as glaucoma screening in practice is inherently a multimodal process that extends beyond fundus imaging. Although, in the traditional Decision Ladder framework, such shortcuts are typically taken by experts with extensive experience, in this study the identified leap could be made by optometrists regardless of their level of expertise. Participants consistently agreed that a multimodal AI system would be more useful and trustworthy, as it better reflects real-world clinical reasoning and diagnostic completeness.

Overall, this model demonstrates that AI-assisted decision-making in glaucoma screening involves a dynamic interplay between the AI's screening output and the optometrist's interpretation of

patient information. It highlights that optometrists' trust is not static but evolves through ongoing comparisons of agreement between the AI's conclusions and their own screening judgments.

4.4 Design Principles (RQ3)

This section addresses RQ3: "What design principles should guide the development of glaucoma AI CDSS to foster and sustain optometrists' trust?" It translates the identified trust-forming factors into concrete design requirements for the next generation of CDSS in glaucoma screening and presents the following requirements.

4.4.1 AI diagnosis Validation and Override Mechanisms

The system shall be designed to reinforce accountability by ensuring that optometrists remain the ultimate authority in the diagnostic process. AI outputs should not be presented as final conclusions; rather, the system should require explicit optometrists review of the test results and other clinical information and confirmation before a diagnosis is accepted. In addition, the system shall provide clear mechanisms for optometrists to validate, override, or reject AI diagnosis and test result interpretation, thereby supporting professional judgment and maintaining clinical responsibility.

4.4.2 Cautious Flagging Strategy

To minimize the risk of missed suspicious test results in the context of glaucoma screening, the system should adopt a cautious flagging approach that prioritizes sensitivity over specificity when uncertainty arises.

4.4.3 Comprehensive Review of Test Results

The system should encourage clinicians to review all test results, whether flagged by AI or not. This design prevents over-reliance on the flagging mechanism and ensures that unflagged results, some of which may still contain clinically relevant information, are not disregarded.

4.4.4 Integration of Multimodal Data into Diagnosis

The system should integrate and present results from multiple diagnostic modalities (e.g., OCT, fundus imaging, visual fields, and patient history) rather than relying on a single source. When generating outputs, the system shall display how each modality contributed to the diagnosis, this integration should mirror established clinical practice, where clinicians cross-reference multiple tests

before making a judgment. Presenting multimodal evidence not only improves diagnostic robustness but also enhances clinician trust by demonstrating that the AI considers the broader clinical picture. If only unimodal data are available for a given case, the system should flag this limitation clearly so that clinicians understand the increased uncertainty and can adjust their reliance accordingly.

4.4.5 Per-Visit Presentation of Model Performance

The system should present a confidence score for each individual visit, expressed as a percentage (e.g., “73% confidence”). This score must be displayed alongside the AI’s output for that specific patient encounter. To avoid presenting it as a raw number alone, the confidence should also be categorized (e.g., “medium confidence”) and color coded, with green indicating high confidence, amber indicating medium confidence, and red indicating low confidence. The cut-off ranges for low, medium, and high categories should be determined empirically through user studies, since evenly distributed thresholds may not accurately reflect users’ perceptions of confidence levels.

4.4.6 Presentation of Model Performance during Onboarding

During onboarding, the system should provide clinicians with a clear overview of its key performance characteristics to support informed use. It should display an overall accuracy metric (e.g., “validated accuracy: 92% on test dataset”) so that optometrists understand the model’s general performance independent of individual cases. Consistency should also be emphasized by ensuring the system consistently returns the same output when identical inputs are provided; this property should be explicitly stated in the onboarding documentation (e.g., “The system produces stable results for repeated identical inputs”). Sensitivity rates must be disclosed as well (e.g., “sensitivity: 88% for detecting early glaucoma changes”), with an emphasis on the model’s ability to detect subtle or early changes more reliably than human optometrists working under time pressure. Communicating both the numerical sensitivity rate and its comparative advantage highlights the added clinical value of the system. Finally, transparency on input data quality is essential: onboarding should include a structured summary of the data sources used for model training (e.g., electronic health records, imaging datasets), the cohort coverage (e.g., demographics, geography, disease representation), and the curation methods applied (e.g., cleaning, labelling, handling of missing values). Known gaps or biases, such as the underrepresentation of specific populations or rare conditions, must also be disclosed to help clinicians judge the limits of generalizability and determine when greater caution is warranted in interpreting AI outputs.

4.4.7 Binary Diagnostic Output

The system shall provide only two possible outcomes, “No Glaucoma” or “Glaucoma Positive”, to reduce diagnostic uncertainty and align with clinicians’ preference for clear, actionable outputs. To prevent misclassification at the decision boundary, the operating threshold separating these two outcomes shall be configurable and calibrated to local clinical practice. A study is required to define a clearer, evidence-based cutoff line for distinguishing “No Glaucoma” from “Glaucoma Positive” cases.

4.4.8 Guideline-Based Recommendations for Next Course of Action

The system should provide recommendations for the next course of action, based on widely accepted guidelines to ensure greater acceptability by optometrists, and each recommendation should be referenced to its source to enhance credibility.

4.4.9 Graphical Visualization of Quantifiable Test Results

The system should use graphical representations to plot quantifiable test results. These graphs should display longitudinal data over time, enabling clinicians to quickly determine whether a patient’s results are stable, improving, or deteriorating. Visual indicators, such as threshold lines, should highlight values that exceed clinically significant cutoffs. This approach reduces cognitive workload, supports faster decision-making, and strengthens the connection between raw test data and diagnostic decisions.

4.4.10 Differential Diagnosis with Comparative Explanations

The system should be able to differentiate glaucoma from common look-alike conditions by generating a ranked list of possible diagnoses, ordered by their estimated probability. For each condition, including glaucoma, the system should present both supporting evidence that explains why the condition is plausible and counterevidence that explains why it may not apply. These “why/why not” explanations should be displayed side by side within the diagnosis panel, allowing clinicians to directly compare glaucoma with alternative conditions and better understand the reasoning behind the system’s output.

4.4.11 Model Output Transparency: Risk Factors, Data Gaps, and Actions

The system presents all AI-identified risk factors with normalized weights in the final diagnosis, not only the top-weighted factor, so users can verify that even the lowest-weighted factors were considered. It also surfaces any inputs the model excluded due to missing or low-quality data, labeling each item with the exclusion reason (e.g., missing value, low quality) and a suggested remediation (e.g., order the missing test, repeat the measurement, or request manual review). The final-diagnosis section includes the following disclaimer: “[Test Name] is [missing/low quality] and was excluded from the model.” For each recommended next step, the system provides concise reasoning and cites its source (e.g., a well-accepted clinical guideline) in a tooltip.

4.4.12 Sensitivity-Oriented Design for Screening Systems

The system shall disclose its validated sensitivity rate (e.g., “sensitivity: 88%”) as part of the onboarding module. Sensitivity shall be highlighted as a core performance metric, reflecting the screening tool’s primary objective of minimizing false negatives. The interface shall communicate that higher sensitivity has been intentionally prioritized, even at the expense of more false positives, to align with the clinical expectation that screening systems adopt a cautious approach.

4.4.13 Communicating AI Limitations and Safeguards

The interface should include a clear disclaimer stating that the AI may contain biases, may not generalize across all patient cohorts, and may overlook comorbidities that contribute to the broader context of glaucoma diagnosis. It should emphasize that AI is a support tool rather than a replacement for clinical judgment, and that clinician validation is required. When the system identifies the user as a novice or when the AI’s confidence score falls below a defined threshold, the system should prompt users to interpret the output with extra caution. Such prompts are essential, as this thesis has highlighted concerns about overreliance on AI diagnoses and recommendations, particularly among less experienced clinicians, who may accept system outputs uncritically or overlook contradictory clinical cues. Further controlled studies are required to determine the threshold that differentiates novice from expert users, as well as to define appropriate confidence score limits that would trigger cautionary prompts emphasizing the importance of critical interpretation rather than blind reliance on AI outputs.

4.4.14 Structured Onboarding and Educational Support

The system should incorporate structured onboarding training and ongoing educational support. Onboarding should familiarize users with the system's intended role, core functionalities, and limitations, while also clarifying how AI outputs are communicated. This includes guidance on interpreting outputs such as suspicious level, confidence scores, decision factors, and recommendations. Clear explanation of what is included, excluded, or simplified in the AI outputs is essential to ensure appropriate interpretation and prevent misinformed use. Educational support should extend beyond initial training, offering accessible resources and refreshers that reinforce appropriate use over time.

4.4.15 Parallel Trial Use in Low-Pressure Settings as a Trust-Building Strategy

To foster calibrated trust and ensure safe integration into clinical practice, the system should include a structured trial phase prior to full implementation. In this phase, clinicians would first make their own diagnostic decisions and then review the AI's output in parallel, allowing them to compare judgments without being biased by the system. The trial should occur in a low-pressure, non-urgent environment, where clinicians have sufficient time to reflect on and familiarize themselves with the system's behavior. By experiencing repeated interactions in this controlled setting, clinicians can build trust gradually, calibrate their reliance on the AI, and better understand both its strengths and limitations before depending on it in real-world, time-sensitive cases.

4.4.16 Concordance Gradient Computation and Caution Cues

The system should compute a non-negative integer distance between the AI's diagnostic category and the clinician's selected category. For larger degrees of disagreement (e.g., ≥ 2), the system should display a non-blocking caution cue and provide a short prompt encouraging the optometrist to review key risk factors.

4.4.17 Structured and Concise Information Presentation

The system should present information in a clear, structured, and concise format, particularly in relation to explanations, so that clinicians can quickly grasp key insights without being distracted or burdened by clutter or unnecessary detail.

4.5 Strengths and Limitations

4.5.1 Study Strengths and Key Contributions:

A central strength of this study is the use of **real-world patient cases** which inherently capture the complexity, variability, and noise that define clinical practice. Unlike synthetic cases, which are often simplified or idealized, authentic patient records capture the full spectrum of challenges clinicians actually face, such as comorbidities, incomplete documentation, atypical disease presentations, and measurement issues. These factors introduce uncertainty into decision-making, and evaluating the interface's trustworthiness under such conditions provides a more accurate assessment of its practical utility.

Another strength lies in our **direct focus on trust**. Prior work often investigates trust through proxies such as reliance and confidence [115], although these concepts are related, they differ in definition. Trust is generally defined as a willingness to accept dependence on another while acknowledging vulnerability [68], Confidence denotes the conviction that a given decision or statement is accurate, based on the evidence at hand [116]. Reliance can be understood as the behavioral manifestation of trust; we choose to rely on someone only when trust is already established, making trust a prerequisite for reliance [117]. In this study, however, we explicitly asked participants about trust itself, focusing on direct trust rather than relying on proxies. This approach strengthens the accuracy of our findings by more clearly capturing the factors involved in trust formation, which we highlight as a key strength of this work.

The study also distinguishes itself as a **multi-factor investigation** that extends beyond explainability, the dimension most emphasized in trust research. Instead, we examine the full set of trust drivers and translate them into design requirements for the next generation of glaucoma screening CDSS. This comprehensive approach strengthens the applicability of our results by connecting trust-related insights directly into design practice.

Another key strength of our study lies in its rigor methodology: every transcript was independently coded by two researchers and then cross-compared, with a third person resolving any disagreements. This consensus-driven process reduces individual coder bias, something not typically seen in qualitative research, making our methodology substantially more robust and credible.

Finally, the study reveals **unique trust drivers in the context of glaucoma CDSS**, some of which include:

- **Differential diagnosis of similar conditions:** The AI's capability to distinguish between potential diagnoses when two or more conditions share overlapping signs and symptoms with glaucoma.
- **Transparency on excluded inputs:** Presenting to the user which inputs were excluded due to poor quality or unavailability (e.g., missing or low-quality data) as part of the model's explanation strategy at the point when the AI provides its diagnosis.
- **Screening model design requirements:** Incorporating an over-cautious approach in screening models by prioritizing false positives over false negatives, consistent with clinical expectations for safety in early detection in screening tools.
- **Agreement concordance as a trust driver:** Introducing concordance between clinician judgment and AI output as a factor influencing trust. While not previously highlighted in glaucoma-related trust studies, it has been noted rarely in other CDSS use cases while discussed outside healthcare contexts.

4.5.2 Limitations

The data for this study was drawn from Faisal Ghaffar's PhD dissertation, which he independently designed and conducted. Accordingly, the author of the present thesis was not involved in the study's design, participant recruitment, or interview process.

Although the distribution of participant characteristics has been described in Section 3.2, it is important to note its implications as a limitation. The skew toward women, earlier-career clinicians, and those with lighter glaucoma exposure means that the findings may not fully generalize to the wider optometrist population, particularly more experienced practitioners.

All participants were recruited from the School of Optometry at the University of Waterloo, which may introduce limit representativeness of the wider optometrist population beyond University of Waterloo School of Optometry. Because of Waterloo's strong technological orientation, participants in this study may have had greater prior exposure to integrating digital tools into their clinical practice. Consequently, the findings may not be fully generalizable to optometrists in other regions, especially those working in rural or resource-constrained environments. In addition, the sample size is

small; however, given that the study is qualitative, the focus is on depth of insight rather than statistical generalizability.

A further limitation relates to the study design. The research was conducted using a mock interface rather than a fully operational AI-based clinical decision support system. Participants interacted with a simulated environment rather than a functioning model supported by real patient data, which may influence the way trust is perceived. The findings, therefore, may differ from those obtained in studies where clinicians engage with an actual working system in real-world practice. In addition, the study was cross-sectional in nature: participants engaged with the interface during a single two-hour interview. Longitudinal studies involving repeated interactions over time may yield different insights, as perceptions and trust in AI systems are likely to evolve with continued use.

4.6 Future Work

Several avenues for future research arise from the limitations identified in this study. First, expanding participant recruitment beyond the School of Optometry at the University of Waterloo would help mitigate selection bias and improve the representativeness of the findings. Including optometrists from diverse geographic, institutional, and practice settings would strengthen the generalizability of results. Increasing the sample size would also allow for more robust comparisons across subgroups, such as levels of experience or glaucoma exposure.

Second, subsequent studies should employ a fully functional AI-based clinical decision support system integrated with real patient data. Such an environment would provide participants with more authentic interactions and allow researchers to assess trust under conditions closer to clinical practice.

Third, adopting a longitudinal design would be valuable. Repeated interactions with an AI system over an extended period could capture how clinicians' perceptions and trust evolve with familiarity and continued use, offering insights that cannot be obtained from single-session studies.

Fourth, the literature review indicates that approximately 89% of publications measure trust in AI using questionnaires [118]. These are often self-developed instruments or single-item questions, many of which lack rigorous validation, making it difficult to compare results across studies [119]. At present, there is no validated trust questionnaire specifically tailored to the healthcare domain. A commonly used questionnaire, the Trust Between People and Automation (TPA) scale [120] is applied in the context of AI-CDSS but do not capture the full range of trust dimensions identified in

this thesis. Prior studies on trust in AI-based clinical decision support systems similarly underscore the need for a comprehensive, domain-specific questionnaire.

Fifth, a future study should investigate adaptive caution prompts in AI-based decision support systems. Specifically, when the system identifies the user as a novice or when the AI's confidence score falls below a defined threshold, the system should prompt users to interpret the output with extra caution. Controlled experiments will be needed to (1) establish objective criteria for distinguishing novice from expert users and (2) determine appropriate confidence score thresholds. These investigations will help ensure that caution prompts are both meaningful and calibrated to user expertise and system reliability.

Sixth, as noted in Section 1.4.2.4, general familiarity with AI leads to a lower likelihood of following AI-CDSS recommendations, which in turn affects adoption through its influence on trust. However, the question remains whether this familiarity is related to negative experiences with AI. If so, the extent to which such negative experiences influence trust in AI-CDSS recommendations remains uncertain. This factor was not examined in the current study, as it was beyond its scope. It is therefore recommended that future controlled experiments investigate how negative prior experiences with AI might affect the adoption of AI-based CDSS.

And finally, A targeted, adequately powered study focused on the magnitude of disagreement is needed to rigorously evaluate the relationship between the extent of disagreement and trust.

References

- [1] C. L. Cifra, J. W. Custer, and J. C. Fackler, “A Research Agenda for Diagnostic Excellence in Critical Care Medicine,” *Crit. Care Clin.*, vol. 38, no. 1, pp. 141–157, Jan. 2022, doi: 10.1016/j.ccc.2021.07.003.
- [2] E. S. Berner, Ed., *Clinical Decision Support Systems*. in Health Informatics. Cham: Springer International Publishing, 2016. doi: 10.1007/978-3-319-31913-1.
- [3] I. Sim *et al.*, “Clinical Decision Support Systems for the Practice of Evidence-based Medicine,” *J. Am. Med. Inform. Assoc. JAMIA*, vol. 8, no. 6, pp. 527–534, 2001, doi: 10.1136/jamia.2001.0080527.
- [4] R. O. Mason, “Ethical Issues in Artificial Intelligence,” in *Encyclopedia of Information Systems*, H. Bidgoli, Ed., New York: Elsevier, 2003, pp. 239–258. doi: 10.1016/B0-12-227240-4/00064-2.
- [5] İ. Yazici, I. Shaye, and J. Din, “A survey of applications of artificial intelligence and machine learning in future mobile networks-enabled systems,” *Eng. Sci. Technol. Int. J.*, vol. 44, p. 101455, Aug. 2023, doi: 10.1016/j.jestech.2023.101455.
- [6] A. L. Samuel, “Programming Computers to Play Games,” in *Advances in Computers*, vol. 1, F. L. Alt, Ed., Elsevier, 1960, pp. 165–192. doi: 10.1016/S0065-2458(08)60608-7.
- [7] P. A. Pawloski, G. A. Brooks, M. E. Nielsen, and B. A. Olson-Bullis, “A Systematic Review of Clinical Decision Support Systems for Clinical Oncology Practice,” Apr. 2019, doi: 10.6004/jnccn.2018.7104.
- [8] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, “An overview of clinical decision support systems: benefits, risks, and strategies for success,” *Npj Digit. Med.*, vol. 3, no. 1, p. 17, Feb. 2020, doi: 10.1038/s41746-020-0221-y.
- [9] P. J. Helmons, B. O. Suijkerbuijk, P. V. Nannan Panday, and J. G. Kosterink, “Drug-drug interaction checking assisted by clinical decision support: a return on investment analysis,” *J. Am. Med. Inform. Assoc.*, vol. 22, no. 4, pp. 764–772, July 2015, doi: 10.1093/jamia/ocu010.

- [10] P. Vonbach, A. Dubied, S. Krähenbühl, and J. H. Beer, “Prevalence of drug–drug interactions at hospital entry and during hospital stay of patients in internal medicine,” *Eur. J. Intern. Med.*, vol. 19, no. 6, pp. 413–420, Oct. 2008, doi: 10.1016/j.ejim.2007.12.002.
- [11] S. Eslami, N. F. de Keizer, D. A. Dongelmans, E. de Jonge, M. J. Schultz, and A. Abu-Hanna, “Effects of two different levels of computerized decision support on blood glucose regulation in critically ill patients,” *Int. J. Med. Inf.*, vol. 81, no. 1, pp. 53–60, Jan. 2012, doi: 10.1016/j.ijmedinf.2011.10.004.
- [12] R. Kwok, M. Dinh, D. Dinh, and M. Chu, “Improving adherence to asthma clinical guidelines and discharge documentation from emergency departments: implementation of a dynamic and integrated electronic decision support system,” *Emerg. Med. Australas. EMA*, vol. 21, no. 1, pp. 31–37, Feb. 2009, doi: 10.1111/j.1742-6723.2008.01149.x.
- [13] M. D. Cabana *et al.*, “Why Don’t Physicians Follow Clinical Practice Guidelines? A Framework for Improvement,” *JAMA*, vol. 282, no. 15, pp. 1458–1465, Oct. 1999, doi: 10.1001/jama.282.15.1458.
- [14] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, “An overview of clinical decision support systems: benefits, risks, and strategies for success,” *Npj Digit. Med.*, vol. 3, no. 1, p. 17, Feb. 2020, doi: 10.1038/s41746-020-0221-y.
- [15] S. T. McMullin, T. P. Loneran, C. S. Rynearson, T. D. Doerr, P. A. Veregge, and E. S. Scanlan, “Impact of an Evidence-Based Computerized Decision Support System on Primary Care Prescription Costs,” *Ann. Fam. Med.*, vol. 2, no. 5, pp. 494–498, Sept. 2004, doi: 10.1370/afm.233.
- [16] C. A. Algaze, M. Wood, N. M. Pageler, P. J. Sharek, C. A. Longhurst, and A. Y. Shin, “Use of a Checklist and Clinical Decision Support Tool Reduces Laboratory Use and Improves Cost,” *Pediatrics*, vol. 137, no. 1, p. e20143019, Jan. 2016, doi: 10.1542/peds.2014-3019.
- [17] A. Georgiou, M. Prgomet, A. Markewycz, E. Adams, and J. I. Westbrook, “The impact of computerized provider order entry systems on medical-imaging services: a systematic review,” *J. Am. Med. Inform. Assoc.*, vol. 18, no. 3, pp. 335–340, May 2011, doi: 10.1136/amiajnl-2010-000043.
- [18] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, “An overview of clinical decision support systems: benefits, risks, and strategies for success,” *Npj Digit. Med.*, vol. 3, no. 1, p. 17, Feb. 2020, doi: 10.1038/s41746-020-0221-y.

- [19] H. Greenspan, B. van Ginneken, and R. M. Summers, “Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique,” *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1153–1159, May 2016, doi: 10.1109/TMI.2016.2553401.
- [20] K. Suzuki and Y. Chen, Eds., *Artificial Intelligence in Decision Support Systems for Diagnosis in Medical Imaging*. in Intelligent Systems Reference Library. Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-319-68843-5.
- [21] V. Gulshan *et al.*, “Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs,” *JAMA*, vol. 316, no. 22, pp. 2402–2410, Dec. 2016, doi: 10.1001/jama.2016.17216.
- [22] A. Y. Hannun *et al.*, “Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network,” *Nat. Med.*, vol. 25, no. 1, pp. 65–69, Jan. 2019, doi: 10.1038/s41591-018-0268-3.
- [23] S. Davis, A. Roudsari, R. Raworth, K. L. Courtney, and L. MacKay, “Shared decision-making using personal health record technology: a scoping review at the crossroads,” *J. Am. Med. Inform. Assoc.*, vol. 24, no. 4, pp. 857–866, July 2017, doi: 10.1093/jamia/ocw172.
- [24] D. Ben-Israel *et al.*, “The impact of machine learning on patient care: A systematic review,” *Artif. Intell. Med.*, vol. 103, p. 101785, Mar. 2020, doi: 10.1016/j.artmed.2019.101785.
- [25] R. P. Evans, L. D. Bryant, G. Russell, and K. Absolom, “Trust and acceptability of data-driven clinical recommendations in everyday practice: A scoping review,” *Int. J. Med. Inf.*, vol. 183, p. 105342, Mar. 2024, doi: 10.1016/j.ijmedinf.2024.105342.
- [26] J. S. Ash, D. F. Sittig, E. M. Campbell, K. P. Guappone, and R. H. Dykstra, “Some unintended consequences of clinical decision support systems,” *AMIA Annu. Symp. Proc. AMIA Symp.*, vol. 2007, pp. 26–30, Oct. 2007.
- [27] M. Khalifa and I. Zabani, “Improving Utilization of Clinical Decision Support Systems by Reducing Alert Fatigue: Strategies and Recommendations,” *Stud. Health Technol. Inform.*, vol. 226, pp. 51–54, 2016.
- [28] K. Goddard, A. Roudsari, and J. C. Wyatt, “Automation bias - a hidden issue for clinical decision support system use,” *Stud. Health Technol. Inform.*, vol. 164, pp. 17–22, 2011.

- [29] S. J. Leslie *et al.*, “Clinical decision support software for management of chronic heart failure: development and evaluation,” *Comput. Biol. Med.*, vol. 36, no. 5, pp. 495–506, May 2006, doi: 10.1016/j.compbimed.2005.02.002.
- [30] E. Murray *et al.*, “Why is it difficult to implement e-health initiatives? A qualitative study,” *Implement. Sci. IS*, vol. 6, p. 6, Jan. 2011, doi: 10.1186/1748-5908-6-6.
- [31] J. S. Ash, D. F. Sittig, E. M. Campbell, K. P. Guappone, and R. H. Dykstra, “Some unintended consequences of clinical decision support systems,” *AMIA Annu. Symp. Proc. AMIA Symp.*, vol. 2007, pp. 26–30, Oct. 2007.
- [32] J. Kabachinski, “A look at clinical decision support systems,” *Biomed. Instrum. Technol.*, vol. 47, no. 5, pp. 432–434, 2013, doi: 10.2345/0899-8205-47.5.432.
- [33] A. C. Scanzera *et al.*, “Optometrist’s perspectives of Artificial Intelligence in eye care,” *J. Optom.*, vol. 15, pp. S91–S97, Jan. 2022, doi: 10.1016/j.optom.2022.06.006.
- [34] F. Ghaffar, N. M. Furtado, I. Ali, and C. Burns, “Diagnostic Decision-Making Variability Between Novice and Expert Optometrists for Glaucoma: Comparative Analysis to Inform AI System Design,” *JMIR Med. Inform.*, vol. 13, no. 1, p. e63109, Jan. 2025, doi: 10.2196/63109.
- [35] D. S. W. Ting *et al.*, “Deep learning in ophthalmology: The technical and clinical considerations,” *Prog. Retin. Eye Res.*, vol. 72, p. 100759, Sept. 2019, doi: 10.1016/j.preteyeres.2019.04.003.
- [36] J. D. Akkara and A. Kuriakose, “Role of artificial intelligence and machine learning in ophthalmology,” *Kerala J. Ophthalmol.*, vol. 31, no. 2, p. 150, Aug. 2019, doi: 10.4103/kjo.kjo_54_19.
- [37] Z. Li *et al.*, “Artificial intelligence in ophthalmology: The path to the real-world clinic,” *Cell Rep. Med.*, vol. 4, no. 7, p. 101095, July 2023, doi: 10.1016/j.xcrm.2023.101095.
- [38] P. Harasymowycz *et al.*, “Medical Management of Glaucoma in the 21st Century from a Canadian Perspective,” *J. Ophthalmol.*, vol. 2016, p. 6509809, 2016, doi: 10.1155/2016/6509809.
- [39] N. Mahabadi, M. Zeppieri, and K. Tripathy, “Open Angle Glaucoma,” in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2025. Accessed: May 19, 2025. [Online]. Available: <http://www.ncbi.nlm.nih.gov/books/NBK441887/>

- [40] T. S. Vajaranant, S. Wu, M. Torres, and R. Varma, “The Changing Face of Primary Open-Angle Glaucoma in the United States: Demographic and Geographic Changes from 2011–2050,” *Am. J. Ophthalmol.*, vol. 154, no. 2, pp. 303–314.e3, Aug. 2012, doi: 10.1016/j.ajo.2012.02.024.
- [41] D. S. Jacobs, “Open-angle glaucoma: Epidemiology, clinical presentation, and diagnosis,” UptoDate. Accessed: May 19, 2025. [Online]. Available: https://www-uptodate-com.proxy.lib.uwaterloo.ca/contents/open-angle-glaucoma-epidemiology-clinical-presentation-and-diagnosis?search=glaucoma&source=search_result&selectedTitle=1~150&usage_type=default&display_rank=1
- [42] B. E. Prum *et al.*, “Primary Open-Angle Glaucoma Preferred Practice Pattern(®) Guidelines,” *Ophthalmology*, vol. 123, no. 1, pp. P41–P111, Jan. 2016, doi: 10.1016/j.ophtha.2015.10.053.
- [43] S. Thomas, W. Hodge, and M. Malvankar-Mehta, “The Cost-Effectiveness Analysis of Teleglaucoma Screening Device,” *PLoS ONE*, vol. 10, no. 9, p. e0137913, Sept. 2015, doi: 10.1371/journal.pone.0137913.
- [44] S. Kingman, “Glaucoma is second leading cause of blindness globally,” *Bull. World Health Organ.*, vol. 82, no. 11, pp. 887–888, Nov. 2004.
- [45] D. B. Rein *et al.*, “The economic burden of major adult visual disorders in the United States,” *Arch. Ophthalmol. Chic. Ill 1960*, vol. 124, no. 12, pp. 1754–1760, Dec. 2006, doi: 10.1001/archophth.124.12.1754.
- [46] J. R. Ehrlich *et al.*, “Prevalence of Glaucoma Among US Adults in 2022,” *JAMA Ophthalmol.*, vol. 142, no. 11, pp. 1046–1053, Nov. 2024, doi: 10.1001/jamaophthalmol.2024.3884.
- [47] L. M. Young, “Glaucoma Facts And Stats - Glaucoma Research Foundation,” Glaucoma Research Foundation. Accessed: May 19, 2025. [Online]. Available: <https://glaucoma.org/articles/glaucoma-facts-and-stats>
- [48] Y.-C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C.-Y. Cheng, “Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis,” *Ophthalmology*, vol. 121, no. 11, pp. 2081–2090, Nov. 2014, doi: 10.1016/j.ophtha.2014.05.013.

- [49] F. Kyari *et al.*, “A Population-based survey of the prevalence and types of glaucoma in Nigeria: results from the Nigeria National Blindness and Visual Impairment Survey,” *BMC Ophthalmol.*, vol. 15, p. 176, Dec. 2015, doi: 10.1186/s12886-015-0160-6.
- [50] S.-M. Thomas, M. Jeyaraman, W. G. Hodge, C. Hutnik, J. Costella, and M. S. Malvankar-Mehta, “The Effectiveness of Teleglaucoma versus In-Patient Examination for Glaucoma Screening: A Systematic Review and Meta-Analysis,” *PLoS ONE*, vol. 9, no. 12, p. e113779, Dec. 2014, doi: 10.1371/journal.pone.0113779.
- [51] D. S. Jacobs, “Open-angle glaucoma: Epidemiology, clinical presentation, and diagnosis,” in *UpToDate*, Wolters Kluwer, 2024. Accessed: July 16, 2025. [Online]. Available: https://www-uptodate-com.proxy.lib.uwaterloo.ca/contents/open-angle-glaucoma-epidemiology-clinical-presentation-and-diagnosis?search=glaucoma&source=search_result&selectedTitle=1~150&usage_type=default&display_rank=1#H1576165147
- [52] F. Topouzis, E. Anastasopoulos, Department of Ophthalmology, Aristotle University of Thessaloniki, Greece, and Department of Ophthalmology, Aristotle University of Thessaloniki, Greece, “Glaucoma — The Importance of Early Detection and Early Treatment,” *Eur. Ophthalmic Rev.*, vol. 00, no. 00, p. 13, 2007, doi: 10.17925/eor.2007.00.00.13.
- [53] J. Scheetz, K. Koklanis, M. McGuinness, M. Long, and M. E. Morris, “A Randomized Trial to Increase the Assessment Accuracy of Glaucoma and Optic Disc Characteristics by Orthoptists,” *J. Contin. Educ. Health Prof.*, vol. 39, no. 3, p. 161, Summer 2019, doi: 10.1097/CEH.0000000000000258.
- [54] N. Yoshioka *et al.*, “Influence of education and diagnostic modes on glaucoma assessment by optometrists,” *Ophthalmic Physiol. Opt.*, vol. 35, no. 6, pp. 682–698, 2015, doi: 10.1111/opo.12247.
- [55] F. Z. Zulfira, S. Suyanto, and A. Septiarini, “Segmentation technique and dynamic ensemble selection to enhance glaucoma severity detection,” *Comput. Biol. Med.*, vol. 139, p. 104951, Dec. 2021, doi: 10.1016/j.combiomed.2021.104951.
- [56] M. Christopher *et al.*, “Deep Learning Approaches Predict Glaucomatous Visual Field Damage from OCT Optic Nerve Head En Face Images and Retinal Nerve Fiber Layer Thickness Maps,” *Ophthalmology*, vol. 127, no. 3, pp. 346–356, Mar. 2020, doi: 10.1016/j.ophtha.2019.09.036.

- [57] S. I. Niwas *et al.*, “Automated anterior segment OCT image analysis for Angle Closure Glaucoma mechanisms classification,” *Comput. Methods Programs Biomed.*, vol. 130, pp. 65–75, July 2016, doi: 10.1016/j.cmpb.2016.03.018.
- [58] Z. Li, Y. He, S. Keel, W. Meng, R. T. Chang, and M. He, “Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs,” *Ophthalmology*, vol. 125, no. 8, pp. 1199–1206, Aug. 2018, doi: 10.1016/j.ophtha.2018.01.023.
- [59] S. Andersson, A. Heijl, D. Bizios, and B. Bengtsson, “Comparison of clinicians and an artificial neural network regarding accuracy and certainty in performance of visual field assessment for the diagnosis of glaucoma,” *Acta Ophthalmol. (Copenh.)*, vol. 91, no. 5, pp. 413–417, Aug. 2013, doi: 10.1111/j.1755-3768.2012.02435.x.
- [60] O. Perski and C. E. Short, “Acceptability of digital health interventions: embracing the complexity,” *Transl. Behav. Med.*, vol. 11, no. 7, pp. 1473–1480, July 2021, doi: 10.1093/tbm/ibab048.
- [61] L. Westerbeek *et al.*, “Barriers and facilitators influencing medication-related CDSS acceptance according to clinicians: A systematic review,” *Int. J. Med. Inf.*, vol. 152, p. 104506, Aug. 2021, doi: 10.1016/j.ijmedinf.2021.104506.
- [62] T. Nadarzynski, O. Miles, A. Cowie, and D. Ridge, “Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study,” *Digit. Health*, vol. 5, p. 2055207619871808, Jan. 2019, doi: 10.1177/2055207619871808.
- [63] A. A. AlQudah, M. Al-Emran, and K. Shaalan, “Technology Acceptance in Healthcare: A Systematic Review,” *Appl. Sci.*, vol. 11, no. 22, p. 10537, Jan. 2021, doi: 10.3390/app112210537.
- [64] F. D. Davis, “A technology acceptance model for empirically testing new end-user information systems: Theory and results,” Massachusetts Institute of Technology, 1985.
- [65] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, “User Acceptance of Information Technology: Toward a Unified View,” *MIS Q.*, vol. 27, no. 3, pp. 425–478, 2003, doi: 10.2307/30036540.
- [66] M. Laka, A. Milazzo, and T. Merlin, “Factors That Impact the Adoption of Clinical Decision Support Systems (CDSS) for Antibiotic Management,” *Int. J. Environ. Res. Public Health*, vol. 18, no. 4, Art. no. 4, Jan. 2021, doi: 10.3390/ijerph18041901.

- [67] H. Hashemian *et al.*, “Application of Artificial Intelligence in Ophthalmology: An Updated Comprehensive Review,” *J. Ophthalmic Vis. Res.*, vol. 19, no. 3, pp. 354–367, Sept. 2024, doi: 10.18502/jovr.v19i3.15893.
- [68] A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr, “Basic concepts and taxonomy of dependable and secure computing,” *IEEE Trans. Dependable Secure Comput.*, vol. 1, no. 1, pp. 11–33, Jan. 2004, doi: 10.1109/TDSC.2004.2.
- [69] O. O’Neill, “Linking Trust to Trustworthiness,” *Int. J. Philos. Stud.*, vol. 26, no. 2, pp. 293–300, Mar. 2018, doi: 10.1080/09672559.2018.1454637.
- [70] R. C. Mayer, J. H. Davis, and F. D. Schoorman, “An Integrative Model of Organizational Trust,” *Acad. Manage. Rev.*, vol. 20, no. 3, pp. 709–734, 1995, doi: 10.2307/258792.
- [71] H. M. Tun, H. A. Rahman, L. Naing, and O. A. Malik, “Trust in AI-Based Clinical Decision Support Systems Among Healthcare Workers: A Systematic Review (Preprint),” *J. Med. Internet Res.*, Dec. 2024, doi: 10.2196/69678.
- [72] V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qual. Res. Psychol.*, vol. 3, no. 2, pp. 77–101, Jan. 2006, doi: 10.1191/1478088706qp063oa.
- [73] S. Yoon *et al.*, “Assessing the Utility, Impact, and Adoption Challenges of an Artificial Intelligence–Enabled Prescription Advisory Tool for Type 2 Diabetes Management: Qualitative Study,” *JMIR Hum. Factors*, vol. 11, no. 1, p. e50939, June 2024, doi: 10.2196/50939.
- [74] V. Sivaraman, L. A. Bukowski, J. Levin, J. M. Kahn, and A. Perer, “Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, in CHI ’23. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–18. doi: 10.1145/3544548.3581075.
- [75] E. R. Burgess *et al.*, “Healthcare AI Treatment Decision Support: Design Principles to Enhance Clinician Adoption and Trust,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, in CHI ’23. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–19. doi: 10.1145/3544548.3581251.
- [76] M. Nair, J. Andersson, J. M. Nygren, and L. E. Lundgren, “Barriers and Enablers for Implementation of an Artificial Intelligence–Based Decision Support Tool to Reduce the Risk of

Readmission of Patients With Heart Failure: Stakeholder Interviews,” *JMIR Form. Res.*, vol. 7, p. e47335, Aug. 2023, doi: 10.2196/47335.

[77] S. G. Anjara *et al.*, “Examining explainable clinical decision support systems with think aloud protocols,” *PLOS ONE*, vol. 18, no. 9, p. e0291443, Sept. 2023, doi: 10.1371/journal.pone.0291443.

[78] A. K. P. Bach, T. M. Nørgaard, J. C. Brok, and N. van Berkel, “‘If I Had All the Time in the World’: Ophthalmologists’ Perceptions of Anchoring Bias Mitigation in Clinical AI Support,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, in CHI ’23. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–14. doi: 10.1145/3544548.3581513.

[79] W. Van Biesen, D. Van Cauwenberge, J. Decruyenaere, T. Leune, and S. Sterckx, “An exploration of expectations and perceptions of practicing physicians on the implementation of computerized clinical decision support systems using a Qsort approach,” *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 185, July 2022, doi: 10.1186/s12911-022-01933-3.

[80] X. Liu *et al.*, “Discrepancy between perceptions and acceptance of clinical decision support Systems: implementation of artificial intelligence for vancomycin dosing,” *BMC Med. Inform. Decis. Mak.*, vol. 23, no. 1, p. 157, Aug. 2023, doi: 10.1186/s12911-023-02254-9.

[81] D. Wang *et al.*, “‘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*, Yokohama Japan: ACM, May 2021, pp. 1–18. doi: 10.1145/3411764.3445432.

[82] M. Yin, J. Wortman Vaughan, and H. Wallach, “Understanding the Effect of Accuracy on Trust in Machine Learning Models,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland Uk: ACM, May 2019, pp. 1–12. doi: 10.1145/3290605.3300509.

[83] K. Yu, S. Berkovsky, R. Taib, J. Zhou, and F. Chen, “Do I trust my machine teammate?: an investigation from perception to decision,” in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, Marina del Ray California: ACM, Mar. 2019, pp. 460–468. doi: 10.1145/3301275.3302277.

- [84] A. Rechkemmer and M. Yin, "When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models," in *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, Apr. 2022, pp. 1–14. doi: 10.1145/3491102.3501967.
- [85] L. Zheng *et al.*, "Clinical Needs Assessment of a Machine Learning–Based Asthma Management Tool: User-Centered Design Approach," *JMIR Form. Res.*, vol. 8, no. 1, p. e45391, Jan. 2024, doi: 10.2196/45391.
- [86] J. Stacy *et al.*, "Qualitative Evaluation of an Artificial Intelligence–Based Clinical Decision Support System to Guide Rhythm Management of Atrial Fibrillation: Survey Study," *JMIR Form. Res.*, vol. 6, no. 8, p. e36443, Aug. 2022, doi: 10.2196/36443.
- [87] X. Wang *et al.*, "A deep learning algorithm for automatic detection and classification of acute intracranial hemorrhages in head CT scans," *NeuroImage Clin.*, vol. 32, 2021, doi: 10.1016/j.nicl.2021.102785.
- [88] D. V. Gunasekeran *et al.*, "Acceptance and Perception of Artificial Intelligence Usability in Eye Care (APPRAISE) for Ophthalmologists: A Multinational Perspective," *Front. Med.*, vol. 9, Oct. 2022, doi: 10.3389/fmed.2022.875242.
- [89] K. E. Henry *et al.*, "Human–machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system," *Npj Digit. Med.*, vol. 5, no. 1, pp. 1–6, July 2022, doi: 10.1038/s41746-022-00597-7.
- [90] D. Jones, A. Cameron, D. J. Lowe, S. M. Mason, C. A. O'Keeffe, and E. Logan, "Multicentre, prospective observational study of the correlation between the Glasgow Admission Prediction Score and adverse outcomes," *BMJ Open*, vol. 9, no. 8, p. e026599, Aug. 2019, doi: 10.1136/bmjopen-2018-026599.
- [91] A. Choudhury, O. Asan, and J. E. Medow, "Effect of risk, expectancy, and trust on clinicians' intent to use an artificial intelligence system -- Blood Utilization Calculator," *Appl. Ergon.*, vol. 101, p. 103708, May 2022, doi: 10.1016/j.apergo.2022.103708.
- [92] C. Jones, J. Thornton, and J. C. Wyatt, "Artificial intelligence and clinical decision support: clinicians' perspectives on trust, trustworthiness, and liability," *Med. Law Rev.*, vol. 31, no. 4, pp. 501–520, Nov. 2023, doi: 10.1093/medlaw/fwad013.

- [93] J. Amann, E. Vayena, K. E. Ormond, D. Frey, V. I. Madai, and A. Blasimme, “Expectations and attitudes towards medical artificial intelligence: A qualitative study in the field of stroke,” *PLOS ONE*, vol. 18, no. 1, p. e0279088, Jan. 2023, doi: 10.1371/journal.pone.0279088.
- [94] C.-W. Chiang and M. Yin, “Exploring the Effects of Machine Learning Literacy Interventions on Laypeople’s Reliance on Machine Learning Models,” in *27th International Conference on Intelligent User Interfaces*, Helsinki Finland: ACM, Mar. 2022, pp. 148–161. doi: 10.1145/3490099.3511121.
- [95] D. Long and B. Magerko, “What is AI Literacy? Competencies and Design Considerations,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA: ACM, Apr. 2020, pp. 1–16. doi: 10.1145/3313831.3376727.
- [96] M. Jacobs, M. F. Pradier, T. H. McCoy, R. H. Perlis, F. Doshi-Velez, and K. Z. Gajos, “How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection,” *Transl. Psychiatry*, vol. 11, no. 1, p. 108, Feb. 2021, doi: 10.1038/s41398-021-01224-x.
- [97] A. Ankolekar *et al.*, “Clinician perspectives on clinical decision support systems in lung cancer: Implications for shared decision-making,” *Health Expect.*, vol. 25, no. 4, pp. 1342–1351, 2022, doi: 10.1111/hex.13457.
- [98] A. Choudhury, O. Asan, and J. E. Medow, “Clinicians’ Perceptions of an Artificial Intelligence–Based Blood Utilization Calculator: Qualitative Exploratory Study,” *JMIR Hum. Factors*, vol. 9, no. 4, p. e38411, Oct. 2022, doi: 10.2196/38411.
- [99] Z. Lu, D. Wang, and M. Yin, “Does More Advice Help? The Effects of Second Opinions in AI-Assisted Decision Making,” *Proc. ACM Hum.-Comput. Interact.*, vol. 8, no. CSCW1, pp. 1–31, Apr. 2024, doi: 10.1145/3653708.
- [100] Z. Lu and M. Yin, “Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, May 2021, pp. 1–16. doi: 10.1145/3411764.3445562.

- [101] O. Vereschak, G. Bailly, and B. Caramiaux, “How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies,” *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, pp. 1–39, Oct. 2021, doi: 10.1145/3476068.
- [102] J. S. Chen *et al.*, “Usability and Clinician Acceptance of a Deep Learning-Based Clinical Decision Support Tool for Predicting Glaucomatous Visual Field Progression,” *J. Glaucoma*, vol. 32, no. 3, p. 151, Mar. 2023, doi: 10.1097/IJG.0000000000002163.
- [103] A. V. D. Brandt *et al.*, “GLANCE: Visual Analytics for Monitoring Glaucoma Progression,” *Eurographics Workshop Vis. Comput. Biol. Med.*, p. 12 pages, 2020, doi: 10.2312/VCBM.20201175.
- [104] O. Deperlioglu, U. Kose, D. Gupta, A. Khanna, F. Giampaolo, and G. Fortino, “Explainable framework for Glaucoma diagnosis by image processing and convolutional neural network synergy: Analysis with doctor evaluation,” *Future Gener. Comput. Syst.*, vol. 129, pp. 152–169, Apr. 2022, doi: 10.1016/j.future.2021.11.018.
- [105] Md. S. Kamal, N. Dey, L. Chowdhury, S. I. Hasan, and K. Santosh, “Explainable AI for Glaucoma Prediction Analysis to Understand Risk Factors in Treatment Planning,” *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–9, 2022, doi: 10.1109/TIM.2022.3171613.
- [106] “Explainable AI Enhances Glaucoma Referrals, Yet the Human-AI Team Still Falls Short of the AI Alone.” Accessed: June 12, 2025. [Online]. Available: <https://arxiv.org/html/2407.11974v1>
- [107] F. Ghaffar, N. M. Furtado, I. Ali, and C. Burns, “Diagnostic Decision-Making Variability Between Novice and Expert Optometrists for Glaucoma: Comparative Analysis to Inform AI System Design,” *JMIR Med. Inform.*, vol. 13, no. 1, p. e63109, Jan. 2025, doi: 10.2196/63109.
- [108] F. Ghaffar, “Understanding Human Decision Variability and the Effects of Data Modality on Trust and Acceptance in Human-AI Collaborative Environment. A Human-Centred Approach to Designing AI Systems in Healthcare Contexts,” University of Waterloo.
- [109] M. C. Leske, S.-Y. Wu, A. Hennis, R. Honkanen, and B. Nemesure, “Risk Factors for Incident Open-angle Glaucoma: The Barbados Eye Studies,” *Ophthalmology*, vol. 115, no. 1, pp. 85–93, Jan. 2008, doi: 10.1016/j.ophtha.2007.03.017.
- [110] M. O. Gordon *et al.*, “The Ocular Hypertension Treatment Study: Baseline Factors That Predict the Onset of Primary Open-Angle Glaucoma,” *Arch. Ophthalmol.*, vol. 120, no. 6, pp. 714–720, June 2002, doi: 10.1001/archopht.120.6.714.

- [111] C. LEWIS, “Using the ‘thinking-aloud’ method in cognitive interface design,” *Res. Rep. RC9265 IBM T J Watson Res. Cent.*, 1982, Accessed: Sept. 06, 2025. [Online]. Available: <https://cir.nii.ac.jp/crid/1571135649469604736>
- [112] F. Ghaffar, N. M. Furtado, I. Ali, and C. Burns, “Diagnostic decision-making variability between novice and expert optometrists for glaucoma: Comparative analysis to inform AI system design,” *JMIR Med. Inform.*, vol. 13, p. e63109, 2025.
- [113] J. Rasmussen, *Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering*. USA: Elsevier Science Inc., 1986.
- [114] K. J. Vicente, *Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*. CRC Press, 1999.
- [115] R. P. Evans, L. D. Bryant, G. Russell, and K. Absolom, “Trust and acceptability of data-driven clinical recommendations in everyday practice: A scoping review,” *Int. J. Med. Inf.*, vol. 183, p. 105342, Mar. 2024, doi: 10.1016/j.ijmedinf.2024.105342.
- [116] A. Pouget, J. Drugowitsch, and A. Kepecs, “Confidence and certainty: distinct probabilistic quantities for different goals,” *Nat. Neurosci.*, vol. 19, no. 3, pp. 366–374, Mar. 2016, doi: 10.1038/nn.4240.
- [117] K. de Fine Licht and B. Brülde, “On Defining ‘Reliance’ and ‘Trust’: Purposes, Conditions of Adequacy, and New Definitions,” *Philosophia*, vol. 49, no. 5, pp. 1981–2001, Nov. 2021, doi: 10.1007/s11406-021-00339-1.
- [118] T. Ueno, Y. Sawa, Y. Kim, J. Urakami, H. Oura, and K. Seaborn, “Trust in Human-AI Interaction: Scoping Out Models, Measures, and Methods,” in *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, in CHI EA ’22. New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 1–7. doi: 10.1145/3491101.3519772.
- [119] J. K. Flake and E. I. Fried, “Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them,” *Adv. Methods Pract. Psychol. Sci.*, vol. 3, no. 4, pp. 456–465, Dec. 2020, doi: 10.1177/2515245920952393.
- [120] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, “Foundations for an Empirically Determined Scale of Trust in Automated Systems,” *Int. J. Cogn. Ergon.*, vol. 4, no. 1, pp. 53–71, Mar. 2000, doi: 10.1207/S15327566IJCE0401_04.

Appendix A

Participants' Demographic Related to Experience:

Participant ID	Gender	years of practicing as an optometrist	Average Number of Patients Seen per Week	Average Glaucoma Patients Screened per Week
P1	Woman	5 years	40-50	Approximately 7-8
P2	Man	11 years	75	10
P3	Man	15	40	5
P5	Man	1.5	90-100	2-4
P6	Man	1	100	20
P7	Woman	10	25	4
P8	Trans/NB	1	100	20
P9	Woman	5.5	47	1
P10	Man	0.66	80	10
P11	Woman	1	65	10
P12	Man	18	80	10
P13	Woman	1	50-60	2-5
P14	Woman	<1	35-40	5
P15	Woman	1	about 50 - 70	10
P16	Man	25	70	20
P17	Woman	6 months	60	10
P18	Woman	5	40	<1
P19	Woman	2	40	6
P20	Woman	25	8	0

Appendix B

Information Letter 1:

Information letter for individuals who are known to Dr. Nadine Furtado

University of Waterloo 04-11-2023

Dear Dr. ABC:

This letter is an invitation to participate in a 150\$ remunerated research study entitled *Glaucoma diagnosis decision variations with accuracy and data modalities of AI systems*. This study is conducted by Faisal Ghaffar (a PhD student at system engineering design department) and supervised by Dr. Catherine Burns (Executive Director, Health Initiatives and Sponsored Research) and Dr. Nadine Furtado (Head of the Ocular Disease service).

The purpose of this study is to examine how glaucoma diagnosis decision, trust and user acceptance varies when optometrist team up with AI systems that use different data modalities and have different accuracies. During the study, you will be asked to review patient records using a think-aloud protocol. At the same time, you will be asked to review AI system outputs that contain the system diagnosis decision. You will be presented with either OCT data of patients, fundus images, complete patient data or a combination of any of these. At the end of each patient record review, you will be asked some questions related to your decision. In this exercise, you will be reviewing at least 2 patient records with 4 different AI interface variants. This will take approximately 1 and a half hours $\pm$ 20 minutes. As a token of appreciation for your participation, you will receive a remuneration of \$150 for completing the interview and pro-rated remuneration (\$75/ hour) based on time spent, if you wish to leave the study early.

With your permission, the study will be recorded to facilitate the collection of information. The researchers will analyze the video recording for glaucoma decision analysis. During the study, you may decline to answer any of the questions and/or share your personal information. Further, you may withdraw from this study at any time by advising the researcher. If you decide to withdraw, we will erase the video recording and the responses that were given during the study.

We are flexible to have an online or physical session. The online interview will be carried out on Microsoft. You will be participating in a study that uses Teams or the Zoom platform. Both platforms have implemented technical, administrative, and physical safeguards to protect the information from loss, misuse, and unauthorized access, disclosure, alteration, or destruction. However, no Internet transmission is ever fully secure or error-free.

Your face will be recorded in both in-person and online interviews. The face in the videos will be blurred for anonymity purposes. After analysis, the videos will be encrypted, and the recordings will be moved to a secure computer in Advance Interface and Design Lab where they will be safely stored with an anonymous number. The recordings will be destroyed after it is transcribed, verified and when it's not needed anymore. The recordings and transcripts will not be published or made public on any platform. Your identity will remain confidential and your name or any other personal identifying information will not appear in any research papers or any publications resulting from this study. Any quotations from the interviews that will be included in the publication will be anonymous. The consent form that you signed will be stored separately from your anonymized study data. All records and transcribed data will be stored in AIDL lab securely for 7 years after which they will be destroyed according to University of Waterloo policy.

Participation in this study may not provide a personal benefit to you but will advance our understanding of teaming up an AI decision making system with optometrist. It will help us to design AI system in such a way that we minimize the risk associated with AI system deployment in critical environment. The risks involved in participating in this study are minimal. If you feel discomfort or fatigue during the study, please let the researchers know and we can stop the study or take a break at any time.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Board (REB #45248). If you have questions for the Board, contact the Office of Research Ethics, toll-free at 1-833-643-2379 (Canada and USA), 1-519-888-4440, or reb@uwaterloo.ca. For all other questions regarding this study, or if you would like additional information to assist you in reaching a decision

about participation, please contact us at f2ghaffa@uwaterloo.ca, nfurtado@uwaterloo.ca, or c4burns@uwaterloo.ca.

We very much look forward to working with you and thank you in advance for your assistance in this project.

Your Sincerely,

Faisal Ghaffar

Faisal Ghaffar, B.Sc., M.Eng,
Ph.D. Candidate
Advanced Interface Design Lab,
Systems Design Engineering,
University of Waterloo, Canada

Appendix C

Information Letter 2:

Information letter for unknown optometrists:

University of Waterloo 04-11-2023

Dear Dr. ABC:

This letter is an invitation to participate in a 150\$ remunerated research study entitled *Glaucoma diagnosis decision variations with accuracy and data modalities of AI systems*. This study is conducted by Faisal Ghaffar (a PhD student at system engineering design department) and supervised by Dr. Catherine Burns (Executive Director, Health Initiatives and Sponsored Research) and Dr. Nadine Furtado (Head of the Ocular Disease service).

The purpose of this study is to examine how glaucoma diagnosis decision, trust and user acceptance varies when optometrist team up with AI systems that use different data modalities and have different accuracies. During the study, you will be asked to review patient records using a think-aloud protocol. At the same time, you will be asked to review AI system outputs that contain the system diagnosis decision. You will be presented with either OCT data of patients, fundus images, complete patient data or a combination of any of these. At the end of each patient record review, you will be asked some questions related to your decision. In this exercise, you will be reviewing at least 2 patient records with 4 different AI interface variants. This will take approximately 1 and a half hours $\pm$ 20 minutes. As a token of appreciation for your participation, you will receive a remuneration of \$150 for completing the interview and pro-rated remuneration (\$75/ hour) based on time spent, if you wish to leave the study early.

With your permission, the study will be recorded to facilitate the collection of information. The researchers will analyze the video recording for glaucoma decision analysis. During the study, you may decline to answer any of the questions and/or share your personal information. Further, you may withdraw from this study at any time by advising the researcher. If you decide to withdraw, we will erase the video recording and the responses that were given during the study.

We are flexible to have an online or physical session. The online interview will be carried out on Microsoft. You will be participating in a study that uses Teams or the Zoom platform. Both platforms have implemented technical, administrative, and physical safeguards to protect the information from loss, misuse, and unauthorized access, disclosure, alteration, or destruction. However, no Internet transmission is ever fully secure or error-free.

Your face will be recorded in both in-person and online interviews. The face in the videos will be blurred for anonymity purposes. After analysis, the videos will be encrypted, and the recordings will be moved to a secure computer in Advance Interface and Design Lab where they will be safely stored with an anonymous number. The recordings will be destroyed after it is transcribed, verified and when it's not needed anymore. The recordings and transcripts will not be published or made public on any platform. Your identity will remain confidential and your name or any other personal identifying information will not appear in any research papers or any publications resulting from this study. Any quotations from the interviews that will be included in the publication will be anonymous. The consent form that you signed will be stored separately from your anonymized study data. All records and transcribed data will be stored in AIDL lab securely for 7 years after which they will be destroyed according to University of Waterloo policy.

Participation in this study may not provide a personal benefit to you but will advance our understanding of teaming up an AI decision making system with optometrist. It will help us to design AI system in such a way that we minimize the risk associated with AI system deployment in critical environment. The risks involved in participating in this study are minimal. If you feel discomfort or fatigue during the study, please let the researchers know and we can stop the study or take a break at any time.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Board (REB #45248). If you have questions for the Board, contact the Office of Research Ethics, toll-free at 1-833-643-2379 (Canada and USA), 1-519-888-4440, or reb@uwaterloo.ca. For all other questions regarding this study, or if you would like additional information to assist you in reaching a decision

about participation, please contact us at f2ghaffa@uwaterloo.ca, nfurtado@uwaterloo.ca, or c4burns@uwaterloo.ca.

We very much look forward to working with you and thank you in advance for your assistance in this project.

Your Sincerely,

Faisal Ghaffar

Faisal Ghaffar, B.Sc., M.Eng,
Ph.D. Candidate
Advanced Interface Design Lab,
Systems Design Engineering,
University of Waterloo, Canada

Appendix D

Study Brochure:

This brochure is for posting to various announcement boards and optometrist groups.

Join our research study and receive \$150 for participating!

Title: Human decision variation with accuracy and data modalities of AI system

Are you an optometrist looking to advance our understanding of teaming up AI decision-making systems with optometrists for glaucoma diagnosis?

We invite you to participate in our research study, Glaucoma Diagnosis Decision Variations with Accuracy and Data Modalities of AI Systems. This study aims to examine how glaucoma diagnosis decision, trust, and user acceptance vary when optometrists work with AI systems that use different data modalities and have different accuracies.

Study Details:

- You will be asked to review patient records using a think-aloud protocol.
- You will review AI system outputs that contain the system diagnosis decision.
- You will be presented with either OCT data of patients, fundus images, complete patient data, or a combination of any of these.
- You will be asked questions related to your decision at the end of each patient record review.
- You will review at least 2 patient records with 4 different AI interface variants.
- The review and interview time will take approximately 1.5 hours (+/- 20 minutes) of your time.
- You will receive a remuneration of \$150 for completing the interview and pro-rated remuneration (\$75 / hour) based on time spent if you wish to leave the study early.
- We offer in-person or online interview options.
- Your face will be recorded during the interview, but it will be blurred for anonymity purposes.

This study has been reviewed and received ethics clearance through the University of Waterloo Research Ethics Board (REB # 45248). If you have questions for the Board, contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 or reb@uwaterloo.ca.

To participate in the study or if you have any questions, please contact:

- Faisal Ghaffar: f2ghaffa@uwaterloo.ca
- Dr. Nadine Furtado: nfurtado@uwaterloo.ca
- Dr. Catherine Burns: c4burns@uwaterloo.ca

Thank you for your interest in our study!

Appendix E

Demographic Questionnaire:

Demographic Questionnaire.

You may decline to answer any of the questions and/or share your personal information. Further, you may withdraw from this study at any time by advising the researcher. If you decide to withdraw, we will erase the responses that were given during the study.

1. What gender do you identify as?
 - A. Man
 - B. Woman
 - C. Trans* / Non-binary
 - D. Prefer to self-identify (option to specify): _____
 - E. Prefer not to answer.
2. What is your age?"
 - A. 20 - 30 years old
 - B. 30 - 40 years old
 - C. 40 -50 years old
 - D. 50- 60 years old
 - E. 60 +
 - F. Prefer not to answer
3. Please specify your ethnicity.
 - A. Indigenous (First Nations, Inuit, Metis)
 - B. European
 - C. Asian (South, East, West)
 - D. African
 - E. Latin, Central and South American
 - F. Caribbean
 - G. Oceania (New Zealand, Pacific Islands)
 - H. Multiracial/ethnic (2 or more)
 - I. Prefer to self-identify: _____
 - J. Prefer not to answer
4. What is the highest degree or level of education you have completed?"
 - A. Bachelor's Degree
 - B. Master's Degree
 - C. Ph.D. or higher
 - D. D. Prefer not to say.
5. Do you have any specific glaucoma related degree/certificate or training?
 - A. Yes
 - B. No.
 - C. Prefer not to say
 - D. If yes, please give details. _____

Appendix F

Feedback Letter:

System Design Engineering,
University of Waterloo
Oct 5, 2021

Dear Participant,

I would like to thank you for your participation in this study entitled "Human decision variation with accuracy and data modalities of an AI system". As a reminder, the purpose of this research is to find out any differences in glaucoma pathways, how different pathways affect the diagnosis decisions taken by the optometrist and finding significant differences in the decision-making process of expert and novice optometrists.

The data collected during interviews will contribute to a better understanding and standardization of the Glaucoma decision making process. It will provide a basis for automating the process of glaucoma diagnosis and ultimately will pave a path for defining an explainable artificial intelligence framework.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Board (REB #45248). If you have questions for the Board, contact the Office of Research Ethics, toll-free at 1-833-643-2379 (Canada and USA), 1-519-888-4440, or reb@uwaterloo.ca.

For all other questions contact Faisal Ghaffar at faisal.ghaffar@uwaterloo.ca

Your identity will be confidential. Once all the data are collected and analyzed for this project, I plan on sharing this information with the research community through seminars, conferences, presentations, and journal articles. If you are interested in receiving more information regarding the results of this study, or would like a summary of the results, please provide your email address, and when the study is completed, anticipated by Feb 2022. I will send you the information. In the meantime, if you have any questions about the study, please do not hesitate to contact me by email as noted below.

Faisal Ghaffar
faisal.ghaffar@uwaterloo.ca

Prof. Catherine Burns
catherine.burns@uwaterloo.ca

**Advanced Interface design Lab,
System Design Engineering,
University of Waterloo**

Appendix G

Consent Form:



CONSENT FORM

By agreeing to participate in this study, you are not waiving your legal rights or releasing the investigator(s) or involved institution(s) from their legal and professional responsibilities. I have read the information presented in the information letter about the study being conducted by Faisal Ghaffar for the project Glaucoma pathways and diagnosis decision (Human decision variation with accuracy and data modalities of an AI system) *Advance interface design lab* at University of Waterloo.

I have had the opportunity to ask any questions related to this study, to receive satisfactory answers to my questions, and any additional details I wanted. I am aware that I have the option of allowing the study to be audio/video recorded to ensure an accurate analysis of my tasks. I am also aware that data collected during the online study may be included in the project papers and/or publications to come from this research, with the understanding that my identity will be kept anonymous.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Board (REB #45248). If you have questions for the Board, contact the Office of Research Ethics, toll-free at 1-833-643-2379 (Canada and USA), 1-519-888-4440, or reb@uwaterloo.ca

I was informed that I may withdraw my consent at any time during the study by informing the researchers before publishing the research. For all other questions contact Dr. Catherine Burns at email catherine.burns@uwaterloo.ca.

With full knowledge of all foregoing, I agree, of my own free will, to participate in this study.

☐ YES ☐ NO

I agree to have the study audio/video recorded.

☐ YES ☐ NO

I agree to the use of anonymous quotations or data in the course project papers and any other publications based on this research.

☐ YES ☐ NO

Participant Name: _____

Participant Signature: _____

Date: _____

Appendix H

Semi-structured Interview Script:

Note: make sure interview is being recorded

Hello, I am Faisal Ghaffar (a PhD student in the system engineering design department). Along with me is Yousuf Zia (an undergraduate student in system design engineering) both supervised by Dr. Catharin Burns (Executive Director of health Initiatives and Sponsored Research) and Dr. Nadine Furtado (Head of the Ocular Disease Service).

Thank you for giving us your time for this interview.

The research study we are conducting is titled, "Optometrist and AI Collaboration for Glaucoma Diagnosis". More specifically, we are exploring Human decision variation with accuracy and data modalities of AI system

The purpose of this study is to examine how optometrists' decisions, trust and AI system acceptance vary when they interact with AI systems having different data modalities. During the study, you will be asked to review patient records from the interface using think think-aloud protocol.

At the end of each patient record review, you will be asked some questions related to your decision and will be asked to fill out a questionnaire related to trust and user acceptance. In this exercise, you will be reviewing 2 patients each of which will have 2 different versions of AI systems. The interview will take approximately 2 hours +- 10 minutes.

With your permission, the study will be recorded to facilitate the collection of information. The researchers will analyze the audio recording for glaucoma decision analysis.

During the study, you may decline to answer any of the questions and/or share your personal information. Further, you may withdraw from this study at any time. If you decide to withdraw, we will erase the audio recording and the responses that were given during the study.

Please fill out this demographic questionnaire for us.

Demographic questionnaire

<https://tinyurl.com/DemographicQuest>

To know about your view on AI systems Please fill out this questionnaire.

General user acceptance questionnaire

https://uwaterloo.ca1.qualtrics.com/ife/form/SV_e8Seud04ezfGBsW

The same questions will be repeated.

You can now move to visit 3 and verbalize your thoughts about this patient's glaucoma status. This system also uses fundus images to detect glaucoma and provide reasoning and explanation for its decision based on fundus images only.

The same questions will be repeated.

Please click on this link and submit your responses related to user acceptance of the AI system. Your participant number is 11 and please select patient 1 unimodal in question 2

https://uwaterloo.ca1.qualtrics.com/ife/form/SV_Qj0GIEAIVdeuOpU

Please click on this link and submit your responses related to user trust in the AI system. Your participant number is 11 and please select patient 1 unimodal in question 2

https://uwaterloo.ca1.qualtrics.com/ife/form/SV_Qwx6vu6HpoirAns

2) \$8072 Uni model

This AI system used here has an overall accuracy of 90% and uses only fundus images to detect glaucoma and provide reasoning and explanation for its decision based on fundus images only.

<https://tinyurl.com/Patient2UniMod>

Please stay on visit 1 and verbalize your thoughts about this patient's glaucoma status.

Questions.

- 1) Is the diagnosis made by the AI system correct for this case? Explain your response.
- 2) How helpful is this AI system in Glaucoma decision-making? (scale 1-5). Explain how it is helpful.
- 3) Rate your confidence in the AI system's decision for this case (scale 1-5).
- 4) What diagnosis would you make for this case?
- 5) List in order of priority the risk factors and/or testing results which you considered in making your decision.
- 6) Rate your confidence in your diagnosis for this case (scale 1-5).
- 7) Do you agree with the grading assigned to risk severity (glaucoma severity) by the AI system? If not, how would you rate the risk severity (glaucoma severity) for this case?
- 8) Do you agree with the conclusion and recommended action by the AI system? If not, what is your conclusion and recommended action?
- 9) Would you trust this AI system to be used in a clinical environment? Explain.

10) UNIMODAL Do you think that this AI system should use more clinical data to make a diagnostic decision?

11) MULTIMODAL Do you think that this AI system could use less clinical data to make a diagnostic decision?

12) What output should the AI system provide for you to support the diagnostic decision so that you feel confident with the clinical diagnosis?

You can now move to visit 2 and verbalize your thoughts about this patient's glaucoma status. This system also uses fundus images to detect glaucoma and provide reasoning and explanation for its decision based on fundus images only.

The same questions will be repeated.

You can now move to visit 3 and verbalize your thoughts about this patient's glaucoma status. This system also uses fundus images to detect glaucoma and provide reasoning and explanation for its decision based on fundus images only.

The same questions will be repeated.

Please click on this link and submit your responses related to user trust in the AI system. Your participant number is 12 and please select patient 2 unimodal in question 2

https://uwaterloo.ca1.qualtrics.com/ife/form/SV_Qj0GIEAIVdeuOpU

Please click on this link and submit your responses related to user trust in the AI system. Your participant number is 12 and please select patient 2 unimodal in question 2

https://uwaterloo.ca1.qualtrics.com/ife/form/SV_Qwx6vu6HpoirAns

3) 70902 multimodal

This AI system used here has an overall accuracy of 90% and uses all data to detect glaucoma and provide reasoning and explanation for its decision based on all type of data marked with green.

<https://tinyurl.com/Patient1Multimodal>

Please stay on visit 1 and verbalize your thoughts about this patient's glaucoma status.

Questions.

- 1) Is the diagnosis made by the AI system correct for this case? Explain your response.

I am forwarding you the first interface link. This interface has all the related data for glaucoma diagnosis.

You'll see that the interface is split into two distinct parts, on the left-hand pane there's a system output, and the right-side pane lets you access all of the patient data. The patient data is divided into 4 categories: their general information, fundus imagery, OCT data, and visual fields per visit. You can view each of these categories by clicking on the respective header at the top of the page. Any piece of information indicated with a green dot represents the fact that it was used in the system decision-making process, while a pink dot indicates that it was not. All data that is highlighted in green is used to create the system diagnosis shown on the left pane.

1) 70902 Uni model

This AI system used here has an overall accuracy of 90% and uses only fundus images to detect glaucoma and provide reasoning and explanation for its decision based on fundus images only.

<https://tinyurl.com/Patient1UniModal>

Please stay on visit 1 and verbalize your thoughts about this patient's glaucoma status.

Questions.

1. Is the diagnosis made by the AI system correct for this case? Explain your response.
2. How helpful is this AI system in Glaucoma decision-making? (scale 1-5). Explain how it is helpful.
3. Rate your confidence in the AI system's decision for this case (scale 1-5).
4. What diagnosis would you make for this case?
5. List in order of priority the risk factors and/or testing results which you considered in making your decision.
6. Rate your confidence in your diagnosis for this case (scale 1-5).
7. Do you agree with the grading assigned to risk severity (glaucoma severity) by the AI system? If not, how would you rate the risk severity (glaucoma severity) for this case?
8. Do you agree with the conclusion and recommended action by the AI system? If not, what is your conclusion and recommended action?
9. Would you trust this AI system to be used in a clinical environment? Explain.
- 10) UNIMODAL Do you think that this AI system should use more clinical data to make a diagnostic decision?
- 11) MULTIMODAL Do you think that this AI system could use less clinical data to make a diagnostic decision?
- 12) What output should the AI system provide for you to support the diagnostic decision so that you feel confident with the clinical diagnosis?

You can now move to visit 2 and verbalize your thoughts about this patient's glaucoma status. This system also uses fundus images to detect glaucoma and provide reasoning and explanation for its decision based on fundus images only.

https://uwaterloo.ca1.qualtrics.com/ife/form/SV_0wx6vu6HpojrAns

4) SS072 multi modal

This AI system used here has an overall accuracy of 90% and uses all types of data to detect glaucoma and provide reasoning and explanation for its decision based on all types of data marked with green.

<https://www.figma.com/proto/QG3VwZfYmHy5KsAkCjI4MX/Patient-2-MultiModal?page-id=0%3A1&type=design&node-id=1-239&viewport=1265%2C752%2C0.22&ts=5mGSvzfaiLHx5CFW-1&scaling=scale-down&starting-point-node-id=1%3A239&show-proto-sidebar=1&mode=design>

Please stay on visit 1 and verbalize your thoughts about this patient's glaucoma status.

Questions.

- 1) Is the diagnosis made by the AI system correct for this case? Explain your response.
- 2) How helpful is this AI system in Glaucoma decision-making? (scale 1-5). Explain how it is helpful.
- 3) Rate your confidence in the AI system's decision for this case (scale 1-5).
- 4) What diagnosis would you make for this case?
- 5) List in order of priority the risk factors and/or testing results which you considered in making your decision.
- 6) Rate your confidence in your diagnosis for this case (scale 1-5).
- 7) Do you agree with the grading assigned to risk severity (glaucoma severity) by the AI system? If not, how would you rate the risk severity (glaucoma severity) for this case?
- 8) Do you agree with the conclusion and recommended action by the AI system? If not, what is your conclusion and recommended action?
- 9) Would you trust this AI system to be used in a clinical environment? Explain.
- 10) Do you agree with the local explanations provided by the system?
- 11) UNIMODAL Do you think that this AI system should use more clinical data to make a diagnostic decision?
- 12) MULTIMODAL Do you think that this AI system could use less clinical data to make a diagnostic decision?
- 13) What output should the AI system provide for you to support the diagnostic decision so that you feel confident with the clinical diagnosis?

You can now move to visit 2 and verbalize your thoughts about this patient's glaucoma status. This system also uses all types of data marked with green to detect glaucoma to detect glaucoma and provide reasoning and explanation for its decision based on fundus images only.

The same questions will be repeated.

You can now move to visit 3 and verbalize your thoughts about this patient's glaucoma status. This system also uses all types of data marked with green to detect glaucoma to detect glaucoma and provide reasoning and explanation for its decision based on fundus images only.

The same questions will be repeated.

Please click on this link and submit your responses related to user trust in the AI system. Your participant number is 14 and please select patient 2 Multimodal in question 2

https://uwaterloo.ca1.qualtrics.com/ife/form/SV_0jOGIEAiVdeuOpU

Please click on this link and submit your responses related to user trust in the AI system. Your participant number is 13 and please select patient 2 Multimodal in question 2

https://uwaterloo.ca1.qualtrics.com/ife/form/SV_0wx6vu6HpojrAns

Can you describe how your interaction differed between the unimodal and multimodal systems in terms of ease of use and efficiency? For instance, did you find multitasking easier with one system over the other, and why?

How did your cognitive load (the mental effort required) compare when using the unimodal versus the multimodal system, and how did this impact your overall satisfaction with each system?

In your experience, how did your level of trust in the accuracy and reliability of the unimodal system compare to the multimodal system?

Reflecting on your experiences with both the unimodal and multimodal systems within a clinical decision-making context, which system did you find more conducive to supporting your diagnostic or treatment decisions?