Understanding Human Decision Variability and the Effects of AI System Data Modality on Trust and Acceptance in Human–AI Collaboration

A Human-Centred Approach to Designing AI Systems in Healthcare Contexts

by

Faisal Ghaffar

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2025

© Faisal Ghaffar 2025

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

Supervisor:	Catherine Burns Professor, Faculty of Engineering, Dept. of System Design Engineering, University of Waterloo
External Examiner:	Carolyn McGregor Professor & Dean, Faculty of Business & IT, Ontario Tech University
Internal-External Member:	Nadine Furtado Professor, School of Optometry and Vision Sciences, University of Waterloo
Internal Member:	Alex Wong Professor, Faculty of Engineering, Dept. of System Design Engineering, University of Waterloo
Internal Member:	Siby Samuel Professor, Faculty of Engineering, Dept. of System Design Engineering, University of Waterloo

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

Papers included in this thesis

- Ghaffar F, Furtado N, Ali I, & Burns, C. M. (2024), Decision-Making Variability in Glaucoma: Comparative Analysis of Novice and Expert Optometrists to Inform AI System Design," Published in *JMIR Medical Informatics*.
  This paper is incorporated in Chapter 4 of this thesis.

Contributor	**Statement of contribution**
Ghaffar F. (Candidate)	Conceptual design (70%) Interviews (80%) Data analysis (75%) Writing and editing (70%)
Furtado N.	Conceptual design (10%) Interviews (15%) Data analysis (0%) Writing and editing (10%)
Ali I.	Conceptual design (0%) Interviews (0%) Data analysis (15%) Writing and editing (5%)
Burns, C. M.	Conceptual design (20%) Interviews (5%) Data analysis (10%) Writing and editing (15%)

- Ghaffar F, Furtado N, & Burns, C. M. (submitted), Different thoughts to the same decision: Experts and Novice optometrists' differences in reasoning and providing

evidences against their decisions." Submitted to Elsevier *J. of Health Policy*.

This paper is incorporated in Chapter 5 of this thesis.

Contributor	**Statement of contribution**
Ghaffar F. (Candidate)	Conceptual design (70%) Data analysis (85%) Writing and editing (75%)
Furtado N.	Conceptual design (10%) Data analysis (0%) Writing and editing (5%)
Burns, C. M.	Conceptual design (20%) Data analysis (15%) Writing and editing (20%)

- Ghaffar F, Islam, Y. Z, Furtado N, & Burns, C. M. (Submitted), Trust and Accuracy in AI: Optometrists Favor Multimodal AI Systems Over Unimodal for Glaucoma Diagnosis. Submitted and under-review in Elsevier *J. of Computers in Biology and Medicine*.

  This paper is incorporated in Chapter 6 of this thesis.

Contributor	Statement of contributions
Ghaffar F. (Candidate)	Conceptual design (75%) Interface development (60%) Participants Interviews (70%) Data analysis (90%) Writing and editing (70%)
Islam, Y. Z	Conceptual design (0%) Interface development (30%) Participants Interviews (30%) Data analysis (0%) Writing and editing (5%)
Furtado N.	Conceptual design (10%) Interface development (5%) Participants Interviews (10%) Data analysis (0%) Writing and editing (5%)
Burns, C. M.	Conceptual design (15%) Interface development (5%) Participants Interviews (0%) Data analysis (10%) Writing and editing (20%)

- Ghaffar F, Woima R, Furtado N, & Burns, C. M. (Submitted), I am not accepting this AI System as my Collaborator: Modality Matters in Clinical Decision-Making. Submitted to IEEE *J. of Translational Engineering in Health and Medicine*, This paper is incorporated in Chapter 7 of this thesis.

Contributor	Statement of contribution
Ghaffar F. (Candidate)	Conceptual design (75%) Data analysis (90%) Writing and editing (60%)
Woima, R	Conceptual design (0%) Data analysis (0%) Writing and editing (30%)
Furtado N.	Conceptual design (10%) Data analysis (0%) Writing and editing (5%)
Burns, C. M.	Conceptual design (15%) Data analysis (10%) Writing and editing (5%)

Use of Generative AI

I acknowledge the use of generative AI tools, specifically Grammarly and ChatGPT-4o, for grammar checks and flow of writing. It is important to note that generative AI tools are not used in Conceptualization, Methodology, Information Collection, Data Collection Methodology Design, Execution, Data Curation or Data analysis part of the PhD study.

Abstract

Effective decision-making is a complex cognitive process that plays a crucial role in high-stakes domains such as healthcare, where inconsistencies in judgment can significantly impact outcomes. It varies widely between individuals, particularly across levels of expertise [Rasmussen, 1983], often leading to variability and inefficiencies [Curran et al., 2022], a phenomenon well-documented in cognitive and decision sciences. As Artificial Intelligence (AI) becomes increasingly integrated into various domains, including finance, transportation, and healthcare, it presents new opportunities to enhance human decision-making. With recent advancements, AI has the potential to mitigate these decision-making challenges, such as inter-user variation, by providing standardized, data-driven recommendations. It can support novice reasoning by guiding decision processes and complement expert intuition when aligned with users' strengths and limitations [Inkpen et al., 2023].

However, other research has shown that standalone AI systems cannot be fully relied upon due to inherent biases, limited contextual understanding, and an inability to adapt dynamically to the complexities of human decision-making. These limitations necessitate a shift toward human-AI collaborative systems, where AI serves as a complementary tool to enhance human performance. Effective collaboration, however, depends on calibrated trust and high acceptance of AI systems, influenced by numerous factors, including the underexplored effect of AI system data modality on user trust and acceptance.

This thesis investigates the variations in decision-making strategies between novices and experts and examines how AI systems can bridge these gaps through human-centric design. It further explores how data modality in AI systems, particularly unimodal versus multimodal data, affects human-AI collaboration. A two-phase study was conducted in the healthcare domain, focusing on glaucoma diagnosis as a specific case study, and employed

a mixed-method approach combining qualitative interviews and quantitative evaluations.

The first phase examined variations in decision-making strategies between novices and experts. It was found that experts adopted more dynamic and efficient approaches by integrating a wider range of factors, emphasizing progression analysis, identifying complex patterns and correlations, and dynamically balancing positive and negative decision factors based on contextual severity. They demonstrated cognitive efficiency by filtering out extraneous information and prioritizing critical data points. In contrast, novices relied on more structured and analytical methods, often overemphasizing explicit indicators and struggling to balance conflicting evidence. Their decision-making factors also varied significantly across different scenarios. Furthermore, the impact of data availability was evident, with novices being more adversely affected by limited data compared to experts.

The second phase evaluated user interactions with unimodal and multimodal AI systems designed for glaucoma diagnosis, measuring trust and acceptance using statistical methods. Multimodal systems consistently outperformed unimodal systems by integrating diverse data sources that mirrored how clinicians process information, leading to better alignment with their workflow. This, in turn, led to significantly higher trust and acceptance for multimodal systems, compared to unimodal systems with significant differences. ($p < 0.01$). Decision-making performance of optometrists also improved, with multimodal systems achieving higher performance than unimodal systems. An interaction effect between user factors (expertise, gender) and system type was also observed, with notable differences in accuracy and confidence levels. These findings show the need for human-centric AI systems that support novice learning and expert decision-making, leveraging data modalities that align with cognitive processes to foster calibrated trust and high acceptance. By enhancing human-AI collaboration, such systems can improve decision-making consistency and optimize outcomes across diverse contexts.

This thesis makes interdisciplinary contributions to human factors, AI, and optometry by investigating cognitive variability in glaucoma diagnosis and examining how data modality influences trust, acceptance in human-AI collaborative environments. In the domain of human factors, it offers insights into how clinical expertise shapes diagnostic reasoning, revealing distinct approaches to data use, heuristic application, and uncertainty management between novices and experts. In the domain of AI, it proposes the design of human-centered AI systems that support optometrists with varying levels of expertise, and also evaluates unimodal and multimodal decision-support systems, demonstrating that multimodal systems more closely align with clinicians' workflows, resulting in higher trust, acceptance, and diagnostic performance. In optometry, the research examines how clinicians interpret glaucoma data, highlighting differences in reasoning strategies such as progression analysis and data prioritization. It informs the development of AI tools that align more closely with optometrists' decision-making processes and integrate seamlessly into routine clinical workflows.

In conclusion, this thesis explores variations in human decision-making and demonstrates the value of human-centric AI systems in supporting both novices and experts. It identifies data modality as a key factor influencing trust and acceptance, providing a rationale for using the same data clinicians rely on to achieve better workflow alignment, leading to higher trust and greater system acceptance. These insights offer a strong foundation for future AI integration in optometry and other high stakes medical domains.

Acknowledgements

This PhD journey has been a transformative and rewarding experience, made possible through the support, guidance, and encouragement of many individuals, institutions, and funding agencies.

First and foremost, I express my deepest gratitude to my supervisor, Professor Catherine Burns, for her unwavering support, insightful guidance, and constant encouragement. Her mentorship has been instrumental in shaping my research, refining my ideas, and helping me grow both as a researcher and as an individual. I am also grateful to Professor Burns and the University of Waterloo for providing funding support through the Natural Sciences and Engineering Research Council of Canada (NSERC CREATE Grant), the Graduate Research Studentship, and the International Doctoral Student Award, all of which were vital to the completion of this work.

I am deeply thankful to Dr. Nadine Furtado, whose collaboration and expertise significantly enriched this research. Her guidance on study design, assistance with participant recruitment, and facilitation of access to critical data were invaluable. I also extend my thanks to Dr. Stanley Woo, Director of the School of Optometry and Vision Science, for enabling and supporting these collaborations.

I am grateful to my committee members, Dr. Alex Wong, Dr. Nadine Furtado, Dr. Siby Samuel, and Dr. Carolyn McGregor, for their thoughtful feedback and valuable recommendations that strengthened this dissertation.

This research was made possible by the generous support of funding agencies and institutions. I sincerely thank the Natural Sciences and Engineering Research Council of Canada, its CREATE program, and the University of Waterloo for additional financial

support through the Graduate Research Studentship and the International Doctoral Student Award.

I would also like to thank all the participants who gave their time and insights to this study. Your contributions provided the essential data that made this work possible.

To my lab colleagues at the University of Waterloo, past and present, thank you for your camaraderie, constructive discussions, and collaborative spirit throughout this journey.

On a personal note, I owe my deepest gratitude to my family. To my parents, Abdul Ghaffar and Jamila Bibi, your unwavering love, prayers, and sacrifices have been the foundation of everything I have achieved. To my wife, Uzma Iqbal, your patience, understanding, and steadfast support during the most challenging moments gave me strength when I needed it most. To my brother, Nasar Ghaffar, and my sisters, Hina Ghaffar and Lubna Ghaffar, thank you for your love and belief in me.

Finally, I thank my friends and extended family who have supported me in countless ways throughout this journey. To everyone who has contributed to this milestone, whether directly or indirectly, I am deeply and forever grateful.

Dedication

To my Family.

Table of Contents

Examining Committee Membership ii

Author's Declaration iii

Statement of Contributions iv

Abstract viii

Acknowledgements xi

Dedication xiii

List of Figures xxix

List of Tables xxxiv

List of Abbreviations xxxvii

1 Introduction **1**

1.1 Artificial Intelligence (AI) 4

1.1.1 Definitions of AI 4

1.1.2 Evolution of AI 5

1.1.3 AI's Expansion into Diverse Fields 6

1.1.4 Human-centred AI and Collaborative Decision-Making 8

1.2 Decision-Making in Healthcare with AI 9

1.3 Challenges in AI Adoption: Variability in Decision-Making, Trust, and User Acceptance 10

1.3.1 Variability in Human Decision-Making 11

1.3.2 Trust 12

1.3.2.1 Definition 12

1.3.2.2 Trust in Automation 13

1.3.2.3 Challenges in Trusting AI 15

1.3.2.4 Designing a Trustworthy AI System 17

1.4 User Acceptance 18

1.4.1 Definitions 18

1.4.2 User Acceptance in Automation, Technology and Human-Centred AI 19

1.4.3 Challenges in User Acceptance of AI 22

1.5 Data Modality and AI: Implications for User Trust and Acceptance 23

1.5.1 Data Modality and Trust in AI 24

1.5.2 Data Modality and User Acceptance of AI . . . 25
1.6 Research Questions . . . 25
1.7 Objectives . . . 28
1.8 Case study of Glaucoma Diagnosis Decision-Making . . . 31
1.9 Thesis Structure . . . 31

2 Case Study of Glaucoma **33**

2.1 Introduction . . . 33
2.2 Objective of the Preliminary Study . . . 34
2.3 Glaucoma Disease and its Diagnosis . . . 35
2.4 Glaucoma Pathways . . . 36
2.5 Apparatus and Exams . . . 39
2.6 Outcome of the Preliminary Study . . . 39
2.6.1 Decision-Making Process of Glaucoma . . . 39
2.6.2 Modeling of observational findings with CWA approaches . . . 41
2.6.3 Data collection . . . 42
2.7 Data Modalities used in Glaucoma Diagnosis . . . 45
2.8 Generalization of Studying Glaucoma to other Contexts . . . 45
2.9 Conclusion . . . 47

3 Diagnostic Decision-Making Variability in Glaucoma: Comparative Analysis of Novice and Expert Optometrists to Inform AI System Design 49

3.1 Abstract . . . 50

3.1.1 Background . . . 50

3.1.2 Objective . . . 51

3.1.3 Methods . . . 51

3.1.4 Results . . . 51

3.1.5 Conclusion . . . 52

3.2 Introduction . . . 52

3.3 Methodology . . . 55

3.3.1 Ethical Consideration . . . 56

3.3.2 Data Refinement and Processing . . . 58

3.3.3 Optometrist Interviews . . . 59

3.3.4 Data Coding . . . 59

3.3.5 Data Analysis . . . 61

3.4 Results . . . 63

3.4.1 Variability in Diagnostic Concordance and Approach Evolution in the Clinical Journey . . . 63

3.4.2 Comprehensive Overview of Examination Data Across all Visits . . . 68

3.4.2.1 All optometrists focus more on optic nerve exams and risk factors. . . . 68

3.4.2.2 Experts carry out progression and change analysis more than novices. 68
3.4.2.3 Experts integrate clinical findings and patient background more than novices. 70
3.4.2.4 Experts tend to emphasize structural function correlation more than novices. 71
3.4.2.5 Novices exhibit a higher tendency for inter-case and longitudinal comparisons, less for intra-case. 72
3.4.3 Variation in Patient Follow-up Time Suggestion 73
3.5 Discussion 74
3.5.1 Decision-Making Processes 75
3.5.2 Data Prioritization in Clinical Assessments 75
3.5.3 Heuristic Approaches and Cognitive Efficiency 76
3.5.4 Guidelines for Developing AI System to Minimize Diagnostic Variations 77
3.5.4.1 Minimizing the Effect of Limited Data Availability 77
3.5.4.2 Evolving AI system for Initial and Subsequent Visits . . . 78
3.5.4.3 Follow-up Time Suggestion 78
3.5.4.4 AI System Features and Behavior 79
3.6 Conclusion 80
3.7 Acknowledgment 80
3.8 Data Availability 81

3.9 Extended Discussion 81
3.9.1 Participant Classification Criteria 81
3.9.2 Clarification on the Number of Cases 83
3.9.3 Clarification on Sample Size 84

4 Different Thoughts to the Same Decision: A Qualitative and Quantitative Investigation of Experience-Based Clinical Reasoning 87

4.1 Abstract 88
4.2 Introduction 89
4.3 Methodology 92
4.3.1 Data Collection 92
4.3.2 Data Coding 93
4.3.3 Data Analysis 93
4.4 Results 95
4.4.1 Variations in Decision-Making: Approaches and Confidence of Novices Versus Experts and the Reasons Behind Their Decisions 95
4.4.1.1 Diagnostic Approach and Decision Reasons Divergence in Glaucoma Cases 98
4.4.1.2 Diagnostic Approaches and Decision-Reasons Divergence in Suspected Cases 102
4.4.1.3 Diagnostic Approaches and Decision-Reasons Divergence in Glaucoma Exclusion 106

4.4.2 Variation in Glaucoma Staging Decision . . . 109
4.4.3 Limitation and Future Work . . . 111

5 From Diagnostic Variation to AI System Design 113

5.1 Introduction . . . 113
5.2 What Kind of AI System Do Optometrists Want? . . . 114
5.2.1 AI System Design Suggestions Derived from the Observed Decision-Making Variation . . . 114
5.2.1.1 Addressing Data Availability Limitations . . . 114
5.2.1.2 Adaptive Support Based on Visit Type . . . 114
5.2.1.3 Standardizing Follow-Up Time Frames . . . 115
5.2.1.4 Prioritizing Relevant Clinical Information . . . 115
5.2.1.5 Optic Nerve Evaluation and Risk Factor Integration . . . 115
5.2.1.6 Supporting Progression Analysis . . . 115
5.2.1.7 Output Interpretability Beyond Binary Decisions . . . 116
5.2.2 Additional Key Insights from Participant Discussion to Inform AI Design . . . 116
5.2.2.1 Data Prioritization and Modality Preferences . . . 116
5.2.2.2 Optometrists Expectations for AI Output and Decision Support . . . 117
5.2.2.3 Preferences for Data Visualization and Presentation . . . 117
5.2.2.4 Explainability and Decision Transparency . . . 118

5.2.2.5 Trust in Unimodal AI Systems . . . 118
5.3 Mockup AI System Design . . . 119
5.3.1 AI System Interface for Unimodal Data . . . 120
5.3.1.1 AI Diagnostic Summary Panel . . . 121
5.3.1.2 Information Specific to Fundus Imaging . . . 123
5.3.2 The Multimodal AI Mockup System Interface . . . 124
5.3.2.1 OCT Tab . . . 124
5.3.2.2 Visual Field . . . 125
5.3.2.3 Patient Background and History Tab . . . 125
5.3.3 Summary Panel . . . 126
5.4 Alignment of AI Systems with Optometrist Needs . . . 126
5.4.1 Adaptive Support According to Visits . . . 126
5.4.2 Follow-Up Time Suggestion . . . 128
5.4.3 Prioritization of Risk Factors and Relevant Clinical Information . . . 128
5.4.3.1 Providing Additional Information Beyond Binary Output . . . 128
5.4.4 AI System Output Format . . . 129
5.4.5 Data Presentation . . . 129
5.4.6 Decision Transparency . . . 129
5.5 Technical Feasibility of the AI Mock-ups . . . 130
5.6 Study Design: Explaining the Participants, Conditions, and Measures . . . 131

5.6.1 Participants . . . 131
5.6.2 Study Conditions . . . 132
5.6.3 Study Procedure . . . 133
5.6.4 Measures . . . 134
5.7 Methodological Choice . . . 135
5.7.1 Mock-Up vs. Real AI . . . 135
5.7.2 Use of Fundus images in unimodal AI system . . . 136
5.7.3 Potential Biases and Mitigation Strategies . . . 137
5.8 Next Chapters . . . 138

6 Trust and Accuracy in AI: Optometrists Favor Multimodal AI Systems Over Unimodal for Glaucoma Diagnosis 139

6.1 Abstract . . . 140
6.2 Introduction . . . 142
6.3 Methodology . . . 148
6.3.1 Ethical Consideration . . . 148
6.3.2 Patients' Data Collection . . . 149
6.3.3 Data Refinement and Processing . . . 150
6.3.4 AI Mock-Up System Development . . . 150
6.3.5 Optometrist Interviews . . . 153
6.3.6 Data Analysis . . . 155

6.4 Results . . . 159
6.4.1 Demographic Characteristics of Optometrists . . . 159
6.4.2 Assessment of Trust Questionnaire Reliability, Validity, and Inter-class consistency . . . 160
6.4.3 Trust Disparities Between Unimodal and Multimodal AI Systems . 161
6.4.3.1 Descriptive Statistics of Trust Targets . . . 163
6.4.4 Inferential Analysis of Trust Variations Between Unimodal and Multimodal AI . . . 164
6.4.5 Influence of Optometrists and System Factors on Trust and Confidence in Unimodal and Multimodal AI . . . 167
6.4.5.1 Effect of Experience on Trust . . . 167
6.4.5.2 Effect of Gender on Trust . . . 167
6.4.5.3 Effect of AI System Accuracy on Optometrists Confidence 169
6.4.6 Performance of Optometrists in Clinical Decision-Making using Unimodal and Multimodal AI Systems . . . 170
6.5 Discussion . . . 171
6.5.1 Development and Validation of the Trust Instrument . . . 172
6.5.2 Interpretation of Trust Targets . . . 174
6.5.3 Impact of Modality on Optometrists' Confidence and Performance . 176
6.5.4 User Characteristics and Trust . . . 177
6.5.5 Limitations and Future Research . . . 178

6.6 Conclusion . . . 178
6.7 Acknowledgment . . . 179

7 Behavioral Determinants of Acceptance of an Artificial intelligent (AI) System: The Impact of Data Modality in human-AI collaboration in Optometry 180

7.1 Abstract . . . 181
7.2 Introduction . . . 183
7.3 Methodology . . . 187
7.3.1 Study Design . . . 187
7.3.2 Ethical Consideration . . . 188
7.3.3 AI Mock-up System Development . . . 189
7.3.4 Optometrist Interviews . . . 191
7.3.5 Data Analysis . . . 193
7.4 Results . . . 194
7.4.1 Demographic Characteristics of Optometrists . . . 194
7.4.2 Descriptive Statistics of User Acceptance Dimensions . . . 195
7.4.3 Variation in User Acceptance of Unimodal vs Multimodal system . . . 196
7.4.3.1 Descriptive Statistics . . . 199
7.4.3.2 Significance of Variation . . . 200
7.4.4 Progressive Changes in User Acceptance: Pre-AI System Interaction (PAI), Post-Unimodal AI system Interaction (PUI), and Post-Multimodal AI System Interaction (PMI) . . . 202

7.4.5 Sustained User Acceptance Despite AI System Errors in Multimodal AI Systems . . . 205
7.4.6 Conclusion . . . 206

8 Discussion and Conclusion 208

8.1 Discussion and Interpretation of Key Findings . . . 208
8.1.1 Variability in Decision-Making . . . 208
8.1.1.1 Decision Making Processes . . . 209
8.1.1.2 Data Prioritization in Clinical Assessments . . . 212
8.1.1.3 Heuristic Approaches and Cognitive Efficiency . . . 212
8.1.2 Decision Factors Variations . . . 213
8.1.2.1 Differences in Factors that Influence the Decisions of Experts and Novices. . . . 214
8.1.2.2 Balancing Positive (in Favor of Glaucoma) and Negative (Against Glaucoma) Factors/Reasons . . . 215
8.1.3 Human-Centric AI System Guideline to Support Variations . . . 217
8.1.3.1 Minimizing the Effect of Limited Data Availability . . . 217
8.1.3.2 Evolving AI System for Initial and Subsequent Visits . . . 218
8.1.3.3 Follow-up Time Suggestion . . . 219
8.1.3.4 Enable Efficient Data Filtration and Prioritization . . . 219
8.1.3.5 Balance Positive and Negative Justifications Dynamically 220
8.1.3.6 Educational Module . . . 220

8.1.3.7 General AI system Features and Behavior 221
8.1.4 Effect of AI System Data Modality on Trust and Implications for AI System Design 222
8.1.4.1 Difference in User Trust on Unimodal and Multimodal AI System 222
8.1.4.2 Demographics Effect on User's Trust and AI Design 224
8.1.4.3 User's has Higher Performance While Interacting with Multimodal System 225
8.1.5 Effect of AI System Data Modality on User's Acceptance 226
8.1.5.1 Overall Acceptability of AI for Glaucoma Diagnosis 226
8.1.5.2 User's Acceptance of Multimodal is Higher Than Unimodal AI Systems 227
8.1.5.3 Progressive Changes to User's Acceptance 228
8.1.5.4 Error May Not Impact User Acceptance 230
8.1.5.5 Final Reflections on Data Modality for AI System Design and Generalizability of Findings 231
8.2 Limitations and Future Directions 232
8.2.1 Participant Diversity 232
8.2.2 Limitation of Case Numbers 233
8.2.3 Limitation of Novice and Expert Clusters 233
8.2.4 Artificial Research Setting 234

8.2.5 Generalizability Across Clinical Specialties 235
8.2.6 Mock AI System and Real-World Deployment 235
8.2.7 Idealized Design and Limitations of the AI Prototype 236
8.2.8 Focus on Multimodal and Unimodal Systems 236
8.2.9 Limited Longitudinal Interaction 237
8.3 Contributions of the Study 238
8.3.1 Contributions to Human Factors 238
8.3.1.1 Insights into Variability in Decision-Making 238
8.3.1.2 Data Prioritization and Cognitive Efficiency 239
8.3.1.3 Heuristic Approaches and Clinical Judgment 239
8.3.1.4 Design of Human-Centric Decision Support Systems . . . 240
8.3.2 Contributions to Artificial Intelligence 240
8.3.2.1 Bridging Expertise Gaps with AI Systems 241
8.3.2.2 Clinician-Driven Approach to AI Design 241
8.3.2.3 Enhancing diagnostic accuracy through multimodal AI systems 242
8.3.3 Contributions to optometry 244
8.3.3.1 Understanding Decision-Making in Glaucoma Diagnosis . 244
8.3.3.2 Guidelines for AI Integration in Optometric Practice . . . 244
8.3.3.3 Enhancing Training and Education 245
8.3.3.4 Practical Implications for Glaucoma Management 245

8.3.4 Final Contribution Remarks 246
8.4 Conclusion 246

References **248**

APPENDICES **288**

A Variabilty study additional data **289**

A.1 Ethics application for varibility study 289
A.2 Participants consent letter for variability study 290
A.3 Demographic questionnaire 292
A.4 Questions asked in the interview 294
A.5 Thematic categorization of factors in glaucoma assessment 295
A.6 Descriptive statistics for thematic assessments across visits 296

B Modality study additional data **298**

B.1 Ethics application for modality study 298
B.2 Participants consent letter for modality study 299
B.3 Questionnaire for assessing general AI acceptance 300
B.4 Descriptive statistics of Individual items of each dimension in trust 312
B.5 Descriptive statistics of confidence score 314
B.6 Additional statistics for user acceptance 315

List of Figures

2.1 A generalized view of the glaucoma diagnosis pathway 38

2.2 Glaucoma diagnosis decision ladder 43

3.1 The figure shows stages and flow of the study. It started with patient data collection and selection, followed by optometrist interviews and data analysis. Each stage is further subdivided into steps. Each of these steps is further explained in the methodology section. The circle and asterisk symbol shows the data obtained at a specific stage in the interview and how it was analyzed. 56

3.2 The figure shows the percentage increase or decrease in glaucoma-related eye examinations and assessments conducted by optometrists, tracked from the initial visit to subsequent visits through longitudinal patient history. The * with name indicate that there is significant difference with $P<.05$ and the error bar shows the 95% confidence interval. 65

3.3 The figure shows findings from the whole dataset, revealing a strong emphasis on optic nerve structural and functional assessments. It also shows moderate attention to clinical findings, medical history, and family history, while progression analysis and structure-function correlation analysis were performed less frequently. The error bar indicates 95% confidence interval. 69

3.4 The figure shows the percentages of glaucoma-related eye examinations conducted by novices and experts, highlighting differences in focus on clinical findings, patient history, progression and change analysis, and structure-function correlation. The * with name indicate that there is significant difference with P <.05 and the error bar shows the 95% confidence interval. 71

3.5 The figure shows the higher follow-up times in months suggested by expert and novice optometrists for confirmed glaucoma cases, suspects, and patients without glaucoma. 74

3.6 The figure shows classification of experts and novice optometrists 82

4.1 Comparative analysis of glaucoma-related eye examinations across different diagnostic decisions. 96

4.2 Comparison of justifications provided in favor of Glaucoma Diagnosis: Expert and Novice Optometrists 101

4.3 Negative indicators against glaucoma confirmation in suspected cases: Expert vs. Novice optometrists' 104

4.4 Positive indicators in favour of glaucoma confirmation in suspected cases: Expert vs. Novice optometrists' 106

4.5 Negative indicators against glaucoma in no glaucoma cases: Expert vs. Novice optometrists' 108

4.6 Distribution of confirmed glaucoma cases stages by experts and novices. . . 110

4.7 Distribution of suspected cases stages by experts and novices. 111

4.8 Comparison and trends of positive and negative reasons in staging: Experts vs Novices 112

5.1 AI mock-up system interfaces (Left-hand side shows the AI global output and the right-hand side includes all the patient information) 121

5.2 Unimodal and multimodal AI mock-up system, side-by-side comparison of all sections. The left-hand side shows a unimodal system and the right-hand side shows a multimodal system 127

6.1 AI mock-up system interfaces (Left-hand side shows the AI global output and the right-hand side includes all the patient information) 152

6.2 Unimodal and multimodal AI mock-up system, side-by-side comparison of all sections. The left-hand side shows a unimodal system and the right-hand side shows a multimodal system 154

6.3 Interview flow, detail of cases, and patient 156

6.4 Demographic distribution of optometrists 160

6.5 Comparison of the distribution of mean score of each target of unimodal and multimodal AI mock-up systems 163

6.6 Distribution of responses for each question in the data modality target, with means for each group marked. 165

7.1 Screenshot of proposed mock interface used. Section A highlights the diagnosis and systems confidence, Section B provides details on the decision shown in Section A. Section C is where different data types can be selected for review, and Section D is the main display for whichever section is under review 190

7.2 Different interface views for the data types available: Fundus imaging, OCT and visual fields. The left side is from a unimodal mock-up and the right side is images from a multimodal mock-up 192

7.3 Demographic information (gender, ethnicity and age) of recruited optometrists 195

7.4 Distribution of year of experience (left), patients seen in a week (middle), and glaucoma patients seen in a week (right) in relation to gender 195

7.5 Distribution of responses against each item of all dimensions of user acceptance except modality preference. 197

7.6 Distribution of responses by optometrist averaged across dimensions for both uni-modal and multi-modal systems 202

7.7 The figure shows distribution of acceptance ratings from optometrists from before interacting with the AI system (green), after interacting with the unimodal system (red) and after interacting with the multimodal system (purple) for four dimensions. 206

7.8 Distribution of acceptance ratings from optometrists before error introduction into the system (green) and after (red) for all four dimensions. 207

A.1 The figure shows the accepted ethics application by Delegated Research Ethics Committee (DERC) of University of Waterloo 289

A.2 Enter Caption 291

A.3 Novices' descriptive statistics for first visits and subsequent visits 296

A.4 Experts' descriptive statistics for first visits and subsequent visits 297

B.1 The figure shows the accepted ethics application by Delegated Research Ethics Committee (DERC) of University of Waterloo for the Modality study 298

List of Tables

2.1 Risk factors of glaucoma 39

2.2 Tests and apparatuses used in glaucoma diagnosis 40

2.3 Distribution of Status, Gender, and Ethnicity 44

2.4 Summary of Attributes, History, Internal Attributes, Images, and Reports . 44

3.1 The table shows data features for glaucoma diagnosis collected from the electronic health record system from the School of Optometry, University of Waterloo. 58

3.2 The table shows the variation in glaucoma-related eye examinations throughout a patient's clinical journey, highlighting the differences between novice and expert optometrists, tracked from initial to subsequent visits. 67

4.1 Percentages of decisive factors used by optometrists in their evaluation . . 99

6.1 Summary of psychometric properties for Uni-modal and Multi-modal constructs. 162

6.2 The table shows Shapiro-Wilk (SW) test for data normality and Wilcoxon Signed-Rank (WSR) test for variability analysis in unimodal and multimodal systems across the items of the user's trust questionnaire. 168

7.1 Summary of reliability tests for Uni-modal and Multi-modal Systems for all dimensions of user acceptance 198

7.2 Descriptive statistic of user acceptance of unimodal and multimodal AI system across all dimensions 199

7.3 Significance of variation of user acceptance of unimodal and multimodal system across all dimensions 201

7.4 The table shows comparative descriptive statistics of user acceptance at pre and post interaction stages across all dimensions 203

7.5 ANOVA Results 205

7.6 Bonferroni Post Hoc Comparisons (t, p-value) 205

B.1 Summary of input data modality across different modalities and questions. 312

B.2 Summary of information quality across different modalities and questions. . 313

B.3 Summary of technical accuracy across different modalities and questions. . 313

B.4 Summary of output explainability across different modalities and questions. 314

B.5 Summary of system quality across different modalities and questions. . . . 314

B.6 Mean ± SD confidence scores for novice and expert optometrists under Unimodal and multimodal conditions, separated by AI decision correctness. . . 314

B.7 Performance of optometrists while intercating with unimodal and multimodal systems, separated by AI decision correctness. 315

B.8 Descriptive statistics of modified UTAUT items 315

B.9 Pairwise comparison of modified UTAUT items for unimodal and multimodal systems 316

List of Abbreviations

α Cronbach's α 156

ACC Accuracy 157, 158

ECT Expectation-Confirmation Theory 229

EE effort expectancy 186, 187, 195, 198–200, 203–206

F1 F1-Score 157–159

FC facilitating conditions 186, 196, 198–200, 203–206

FN False Negative 158

FP False Positive 157

ICC Inter-item correlation 156

MP modality preference 186, 198–200, 202, 206

PAI Pre-AI system Interaction 202–204, 228, 229

PE performance expectancy 186, 187, 195, 198–200, 203–206

PMI Post-multimodal AI system Interaction 202–204, 228, 229

PREC Precision 157, 158

PUI Post-Unimodal AI system Interaction 202–204, 228, 229

REC Recall 157–159

SI social influence 186, 196, 198–200, 203–206

SW Shapiro-Wilk test 157

TN True Negative 158

TP True Positive 157

UTAUT Unified Theory of Acceptance and Use of Technology 135, 182, 184–186, 191, 206

WSR Wilcoxon Signed-Rank test 157

Chapter 1

Introduction

Integrating Artificial Intelligence (AI) into clinical decision-making is changing the healthcare landscape. In the last few decades, AI has become so intense that its decision-making accuracy has surpassed the accuracy of expert human beings in some fields [Silver et al., 2016, Silver et al., 2017]. These highly accurate systems have encouraged humans to adopt AI into daily life systems so that they can help in the decision-making process. According to the AI Index Report 2022, about 56% of organizations use AI capabilities in at least one business function. Due to high adoption, the global AI market is currently estimated at USD 93.5 billion [Zhang et al., 2021]. Aside from the benefits of AI, AI is anticipated to have substantial societal impacts and ethical challenges when interacting with people [Hagerty and Rubinov, 2019]. A deeper understanding of human-AI interaction (HAI) and experience is crucial to studying these impacts and challenges of an AI system.

To avoid the negative impacts of AI, it needs to work closely in collaboration with the human to assist in decision-making processes. For an efficient collaborative environment

and successful AI system deployments, a comprehensive understanding of human decision-making processes, aligning AI systems with user needs and mental models, and cultivating trust and user acceptance are necessary. Moreover, the AI system design should be human-centred and designed so that the AI system aligns with human mental models and should foster calibrated trust and high acceptance.

Before designing an AI system to assist humans in decision-making, the human decision-making process needs to be understood. Effective human-AI collaboration requires not only recognizing diverse decision-making patterns but also ensuring a smooth cognitive handoff between human users and AI systems. This handoff is influenced by factors such as experience, expertise and preferred practices, which shape how users perceive, interpret, and integrate AI system decisions. Designing AI systems that account for these differences ensures that information is presented in a way that aligns with users' cognitive processes, promoting appropriate reliance and effective collaboration. Moreover, an AI system is deemed successful if the users have high confidence in its decisions (user trust), are willing to involve the system decision in his/her own decisions (user acceptance) and are satisfied with its behaviour (explainable, fair, safe). The user trust and acceptance of an AI system are highly affected by various factors, such as system accuracy, explainability, efficiency, data modality, etc.

This thesis examines these interconnected aspects, with a particular focus on human decision-making variability and the role of data modalities in shaping user trust and acceptance. By addressing these factors, the research aims to inform the design of AI systems in healthcare that are better aligned with clinical workflows and user expectations, ultimately facilitating smooth integration into healthcare practice.

Integrating AI into clinical workflows requires more than achieving high diagnostic accuracy; it must support clinicians' diverse cognitive needs. Medical decision-making varies

significantly between novices and experts, with novices relying on structured guidelines while experts use pattern recognition and intuition. These differences shape how clinicians interact with AI. Understanding these variations is essential, not as a problem to be solved, but as a factor that informs AI design to support clinicians across experience levels, address diverse cognitive needs, and enhance human-AI collaboration without disrupting established reasoning processes. In addition, user's trust and acceptance play a critical role in determining whether clinicians adopt and rely on AI tools. One Key factor influencing these factors is the data modality of an AI system and its alignment with decision-making processes

Data modalities play an important role in shaping user perceptions of AI, influencing trust and acceptance in clinical settings. The term data modality refers to the type or format of input data that an AI system processes, such as imaging data, text data, or data generated with different devices. Multimodal AI systems, which integrate multiple data types, such as merging imaging with text data or merging multiple device data, offer richer and more comprehensive insights. In contrast, unimodal systems, which rely on a single data type, are simpler and more straightforward. Investigating the data-modality factor is essential for designing and creating AI systems that align with diverse user needs, foster trust, enhance acceptance, does not disrupt domain workflows, and ultimately enable smooth human-AI collaboration.

In this thesis work in particular, we explore the variability in clinical decision making of optometrists in glaucoma diagnosis and examine the effect of data modalities on user trust and acceptance. An understanding of these factors is critical to overcoming barriers to adoption and ensuring that AI systems help clinicians in their decision-making processes. Developing AI systems that align with the diverse styles of clinicians will promote calibrated trust, improve user acceptance, and enable AI to act as a reliable and supportive

collaborator rather than a disruptive addition to clinical practice.

Ultimately, this study contributes to the broader effort of providing guidelines on developing human-centered AI systems that not only ease clinician mental load but foster calibrated trust, improve clinician acceptance of the system and improve diagnostic consistency. It further lays a foundation for the development of AI systems that adapt to complex clinical decisions by examining the interplay of decision variability, data modality, user trust, and acceptance.

1.1 Artificial Intelligence (AI)

1.1.1 Definitions of AI

From the beginning, the goal of machines has been to match human intelligence. In the 1950s, Alan Turing's famous question, "Can machines think?" [Turing, 2009] laid the foundation for artificial intelligence (AI). AI is broadly defined as the simulation of human intelligence in such a way that machines think and exhibit behaviour like humans. The term was coined by [McCarthy et al., 1955], who defined AI as "the science and engineering of making intelligent machines." This definition provides the foundational goal of AI, which is to replicate human intelligence in machines through computational approaches. Recently, AI is defined as "systems that perceive their environment and take actions that maximize their chances of achieving their goals" [Russell and Norvig, 2021]. This latest definition shows the adaptive and goal-oriented nature of AI, emphasizing its ability to interact with dynamic environments and operate autonomously to achieve desired outcomes.

1.1.2 Evolution of AI

Artificial intelligence (AI) has evolved from rigid, rule-based systems to adaptive, data-driven algorithms that are capable of handling complex tasks. Early AI systems, often called symbolic AI or Good Old-Fashioned AI, relied on predefined rules [McCarthy et al., 1955]. Although these systems worked well in controlled settings, they did not have the flexibility and scalability that are needed for real-world applications [Nilsson, 2009]. Due to this constraint, machine learning (ML) emerged in the 1980s, when AI started to recognize patterns in data instead of depending exclusively on static programming [Mitchell, 1997].

This shift was further sped up in the 1990s by developments in computing power and access to big datasets, as well as by algorithms like support vector machines, decision trees, and Bayesian networks, random forest. By the 2000s, advances in neural networks, supported by scientists like Yoshua Bengio, Yann LeCun, and Geoffrey Hinton, had completely changed artificial intelligence [Lecun et al., 1998, LeCun et al., 1989]. Computer vision and natural language processing, mainly driven by convolutional neural networks [Krizhevsky et al., 2012] and recurrent neural networks [Hochreiter and Schmidhuber, 1997], allowed systems to process complex, high-dimensional data through hierarchical architectures, eliminating the need for handcrafted features.

Recently, the advancements in deep learning [LeCun et al., 2015] and transformer architectures [Vaswani et al., 2017] have enabled the development of large language models (LLMs) such as OpenAI's GPT series and Google's Gemini. These models utilize extensive text data to achieve contextual fluency in language-related tasks. Similarly, Large visual models (LVM) and large behaviour models (LBM) have been developed that can integrate

diverse data types to deliver comprehensive insights. Additionally, multi-agent systems, supported by frameworks from organizations such as DARPA and Microsoft, are facilitating decentralized decision-making in dynamic environments, including robotics, autonomous vehicles, and smart grids.

This progression from rule-based systems to sophisticated, autonomous, and context-aware AI highlights the field's rapid advancement. Modern AI systems are poised to engage in complex interactions, adapt to dynamic contexts, and collaborate across domains, shaping the next era of transformative innovation.

1.1.3 AI's Expansion into Diverse Fields

AI systems have simplified human life, decreased the burden of making complicated decisions and become a transformative force in many industries, such as healthcare finance, education, agriculture, environmental management, transportation, etc.

In healthcare, it is used for the diagnosis of life-threatening diseases such as Cancer [Sun et al., 2016], diabetes detection [Swapna et al., 2018], cardiovascular disease detection [Lih et al., 2020], eyes diseases such as diabetic retinopathy [Gulshan et al., 2016] [Ghaffar et al., 2016] and glaucoma [Mursch-Edlmayr et al., 2020], skin diseases detection [Goceri, 2019], beauty and cosmetics surgeries [Jarvis et al., 2020] and the 3-dimension convolution neural networks are helping to understand better the 3D images created by magnetic resonance imaging (MRI) and CT scans [Singh et al., 2020]. Companies like Google DeepMind and IBM Watson have achieved specialist-level accuracy in detecting diseases. Additionally, DeepMind's AlphaFold [Jumper et al., 2021] has revolutionized drug discovery by predicting protein structures with high precision, which is helping in the development of new therapies.

In finance, AI is used for improving fraud detection, automating processes, and enhancing personalization. Machine learning models like deep neural networks identify anomalies in large datasets, enabling robust fraud prevention [Mishra et al., 2024]. Financial institutions utilize AI for efficient underwriting [Sachan et al., 2024] and customer support through chatbots [Nirala et al., 2022]. It also ensures compliance by automating regulatory monitoring [Chao et al., 2022] and improves investment strategies with predictive analytics [Goel et al., 2023].

In education, AI is used to enhance students' engagement, improve learning of students and support teachers. It has been used for customizing content to optimize outcomes [Extance, 2023], and provides targeted feedback to enhance understanding [Pan et al., 2023]. Chatbots have been developed to streamline administrative tasks, freeing educators to focus on teaching [Nouhan et al., 2021].

In agriculture, it has automated labour-intensive tasks and facilitates real-time monitoring of crop health and soil conditions. This resulted in higher yields and more efficient use of resources. Moreover, it forecasts insect infestations, which aids in prompt interventions and lessens the need for pesticides [Mittal et al., 2024]. Additionally, AI-powered robots automate tasks such as planting, harvesting, and weeding, improving productivity and addressing labour shortages [Wakchaure et al., 2023].

AI is advancing transportation by enhancing efficiency and safety and enabling autonomous vehicles. It is used to optimize traffic flow, and reduce congestion [Chen et al., 2023], and has significantly enhanced safety by enabling collision-avoidance technologies and accident prediction, particularly in autonomous vehicles [Muzahid et al., 2023].

AI continues to transform all these fields by solving difficult problems, furthering social

objectives, and driving innovation through automation, data-driven decision-making, and real-time insights, This transformation is increasingly guided by the principles of human-centred AI, fostering collaborative decision-making that emphasizes the integration of AI capabilities with human expertise to address complex challenges effectively.

1.1.4 Human-centred AI and Collaborative Decision-Making

Human-centred AI (HCAI) and human-AI Collaboration are interconnected yet distinct concepts within the AI domain. HCAI is an approach that focuses on designing AI systems to enhance human skills and capabilities so that they are in line with human needs, values and ethical standards. Schneiderman defines HCAI as a framework in which humans control AI systems while ensuring and enhancing their safety, reliability, and trustworthiness [Shneiderman, 2020]. Being centred on humans means the one that aligns with societal values and ethical principles [Dignum, 2019].

By adopting these principles, HCAI not only creates ethical, user-friendly and trustworthy systems but also establishes AI as a collaborative partner in decision-making. In this collaboration, AI contributes analytical power and data-driven insights, while humans provide contextual understanding, moral reasoning, and creative problem-solving. This collaboration is critical for addressing complex challenges in domains like healthcare, where human-AI collaboration can enhance diagnostic precision and increase the clinician's efficiency.

1.2 Decision-Making in Healthcare with AI

AI is assisting clinicians in their decision-making by providing tools that enhance the accuracy, efficiency, and reliability of decisions. It provides diagnostic support, predicts patient outcomes, and recommends treatment strategies. For example, machine learning models can analyze medical imaging to identify conditions such as cancer with higher accuracy and speed than traditional methods [Bai et al., 2024]. Predictive analytics tools enable clinicians to anticipate complications, such as sepsis or cardiac events, allowing for timely interventions [Valik et al., 2023, fei Sun et al., 2024]. Furthermore, AI reduces diagnostic errors by identifying patterns in data that might be overlooked by human practitioners [Khalifa and Albadawy, 2024]. Beyond diagnostics, AI optimizes operational workflows, streamlining patient scheduling and resource allocation, ultimately improving healthcare delivery [Pierre et al., 2023].

However, despite these benefits, the integration of AI into clinical decision-making presents significant challenges.

Clinicians employ diverse decision-making approaches shaped by differences in expertise, heuristics, and cognitive strategies. This variability allows flexibility in clinical reasoning, enabling clinicians to navigate complex cases using methods suited to their training and experience. AI systems must be designed to complement this diversity, ensuring that they provide meaningful support across different decision-making styles. A solution designed for one clinical context may not integrate smoothly with others, creating inconsistencies in the way AI systems are used in practice. Overcoming these complexities requires an AI that adapts to diverse reasoning processes and skill sets.

Similarly, clinicians rely on specific data formats in their decision-making, and when AI processes input data using unfamiliar or non-traditional modalities, it may introduce

uncertainty about the validity and reliability of its outputs. The type of input data AI utilizes, whether imaging, text reports, or multimodal combinations, affects how clinicians interpret and trust its recommendations. If AI systems rely on data modalities that do not align with clinicians' expectations or typical diagnostic workflows, skepticism may arise regarding the accuracy and applicability of AI-generated decisions, potentially impacting clinicians' trust and acceptance. While reliability, explainability, and bias have been widely studied in fostering trust, the role of input data modality in AI systems remains comparatively underexamined in healthcare. This gap raises critical questions about whether certain types of input data inherently inspire more confidence than others and how mismatches between AI data processing and clinical reasoning might further hinder trust and acceptance

Addressing these challenges is essential for developing AI systems that are accurate and aligned with the cognitive processes of clinicians. Smooth integration into clinical workflows requires a human-centered design, which encourages collaboration between AI systems and clinicians to build trust, enhance acceptance, and ultimately improve clinical outcomes. The following section discusses variability in decision making, trust in AI, and user acceptance, all fundamental to successful adoption of AI in healthcare.

1.3 Challenges in AI Adoption: Variability in Decision-Making, Trust, and User Acceptance

Human-AI collaboration has emerged as a transformative approach across industries that has the ability to improve decision-making, streamline and streamline workflows, and achieve outcomes that drive efficiency and value. However, achieving an effective inte-

gration of AI systems into decision-making presents significant challenges. In this collaboration, where humans retain ultimate authority, the decision-making process is shaped by factors such as variability in human judgment, trust and acceptance of AI systems, the reliability of AI recommendations, and the alignment of AI outputs with clinical workflows and needs.

1.3.1 Variability in Human Decision-Making

Variability in human decision making refers to the differences in how individuals perceive, process and act on information when solving problems or making decisions. It is shaped by cognitive factors (mental processes such as perception, memory, and problem-solving), contextual factors (external elements like the environment, task complexity, and time constraints), and experiential factors (individual's prior knowledge, skills, and experiences) all of which contribute to diverse responses to the same set of circumstances[Judd et al., 2024].

Rather than being a limitation, decision-making variability is an asset, allowing individuals to adapt flexibly to different situations. This adaptability manifests through heuristics and biases, which function as cognitive shortcuts for managing uncertainty and are shaped by an individual's background and experiences [Gigerenzer and Gaissmaier, 2011]. These shortcuts influence how individuals recognize patterns and make judgments, with their effectiveness depending on prior knowledge and exposure. While novices often rely on structured guidelines or rule-based approaches due to limited experience, experts draw on intuition and pattern recognition refined through practice. Recognizing these differences within a specific domain is crucial for designing AI systems that accommodate diverse decision-making styles and user needs. By providing meaningful support, AI-driven decision-support tools can help less experienced users develop effective heuristics while en-

suring that expert users benefit from systems that complement their advanced strategies. This approach enhances decision-making across varying expertise levels and strengthens human-AI collaboration.

1.3.2 Trust

1.3.2.1 Definition

Trust is a multifaceted concept defined and studied across numerous domains, including psychology, human-computer interaction, and artificial intelligence (AI). Its definitions have evolved to reflect the complexities of interpersonal relationships and relationships with automated systems in a collaborative environment. It involves reliance on the actions, intentions, or outcomes of another entity, often in situations involving risk or uncertainty.

In everyday life, trust is commonly understood as reliance and belief in the reliability, ability, or integrity of someone or something. The Merriam-Webster Dictionary defines trust as "reliance on the integrity, ability, or truth of someone or something" [Merriam, 2024] while the Oxford Dictionary describes it as a "firm belief in the reliability, truth, or ability of someone or something" [Oxford, 2024b]. In academic literature, Deutsch (1958) defines trust as "the confidence that an individual has in the intentions and actions of others, such that they will act in ways beneficial or at least not detrimental to the trustor," highlighting its foundational role in interpersonal relationships. Expanding on this, Gambetta (1988) and Luhmann (1979) conceptualize trust as a risk acceptance process, where individuals rely on others based on expectations about their behavior. Together, these perspectives establish trust as both a belief in reliability and an evaluation of risk in human relationships.

1.3.2.2 Trust in Automation

With the advancement of technology, trust extended beyond interpersonal contexts to include human interactions with automated systems. Trust in automation is the user's confidence in system performance and reliability. It is the extent to which a person is willing to rely on an automated system to perform a task based on expectations of its performance and reliability. [Barber and House, 1985] defines trust in automation (T) as the expectation (E), held by a member of a system (i), of persistence (P) of the natural (n) and moral social (m) orders, and of technically competent performance (TCP), and of fiduciary responsibility (FR), from a member (j) of the system, and is related to, but is not necessarily isomorphic with, objective measures of these properties.

In trust research, researchers have approached human behaviour towards automated systems in a similar way as they have approached trust in an interpersonal relationship. Researchers are divided on its similarities with interpersonal trust. Some claim that there are significant similarities between interpersonal and trust in automation and have attempted to extend interpersonal trust concepts into research on trust in automation [Dzindolet et al., 2003, MUIR, 1994]. While others have been stricter in claiming a need for separate investigations and modelling [Lee and See, 2004, Madhavan and Wiegmann, 2007]

The foundational research on trust in automation began in the 1990s, with notable contributions from [Dzindolet et al., 2003, MUIR, 1994, MUIR and MORAY, 1996, LEE and MORAY, 1992, Lee and Moray, 1994] focused on understanding how trust affects human reliance on automated systems. They examined the dynamics of trust formation, factors influencing trust, and the consequences of inappropriate trust, such as over-trust or distrust. These studies also explored decision-making processes in automated environ-

ments and how trust evolves over time in human-machine interactions. Subsequent studies, including those by [Lee and See, 2004, Dzindolet et al., 2003, Jian et al., 2000a] expanded the understanding of trust in automation by introducing frameworks to model and measure trust and by exploring design strategies to foster appropriate trust. These works examined how system design influences user behaviour, why reliable automation is sometimes rejected, and how trust can be measured using surveys and behavioural analysis, emphasizing the psychological and contextual factors shaping trust.

Recently, with advancements in AI systems, trust has emerged as a critical factor, as the decision to use a particular AI system is heavily influenced by trust [Afroogh et al., 2024]. This dynamic introduces a new dimension of HAI. HAI differs from traditional HCI as it faces new challenges related to uncertainty about AI capabilities and the complexity of its outputs [Yang et al., 2020]. HAI thus introduces new challenges in trust and collaboration as humans interact with intelligent decision-making systems. In this interaction, the trustor is the human user, and the trustee is the AI system. This interaction or relationship can be seen in diverse fields. For example, AI can process medical images to diagnose cancer, with physicians as the trustors who adjust or base their decisions on the AI model's outcomes [Asan et al., 2020]. Similarly, Robo-advisors provide investment advice [Leon, 2020] in which the system acts as an advisor– a trustee. Similarly, human drivers interact with AI-based systems in safety-critical situations in autonomous vehicles [Ajenaghughrure et al., 2020]. Building on these examples, HAI emerges as a multidisciplinary field that explores how humans engage with computers and automated systems across such diverse contexts. Trust becomes a pivotal factor in this interaction, influencing user behaviour, technology adoption, and the effectiveness of human-AI collaboration. In HAI, trust is not only affected by reliability, performance and efficiency but also affected by the design of systems. The design of the system incorporates the quality of the infor-

mation and system, the accuracy of the system, type/types of data used by the system, interpretability, transparency, ethical accountability, and the user's ability to understand and predict system behaviour [Doshi-Velez and Kim, 2017, Tucci et al., 2022a]

1.3.2.3 Challenges in Trusting AI

In Human-AI Interaction (HAI), trust is shaped not only by the reliability, performance, and efficiency of AI systems but also by their design. Several challenges related to trust arise when using AI, including issues of interpretability, transparency, ethical accountability, and the quality, accuracy, and modality of data. These factors are interrelated and influence how users perceive and rely on AI-driven decision-making, ultimately affecting human-AI collaboration.

Trust in AI systems is strongly influenced by their interpretability, transparency, and the clarity with which they explain their decisions [Yang et al., 2023]. Many AI-driven systems, particularly those based on deep learning, generate decisions through complex models (termed as blackbox model) and opaque processes [Hassija et al., 2024]. Without clear explanations of how an AI system arrives at its outputs, users may struggle to assess its reasoning, leading to low confidence on the system. This lack of transparency is particularly problematic in high-stakes domains such as healthcare and finance, where explainability is necessary for informed decision-making.

Another factor that affects trust is uncertainty in AI performance [Tomsett et al., 2020]. AI models are trained on historical data, which may not always generalize well to novel or ambiguous situations. Users who experience inconsistencies, where AI performs well in some cases but poorly in others, may hesitate to trust its recommendations though different users may respond differently. Communicating uncertainty clearly and designing

AI systems that adapt to changing conditions can mitigate this issue.

Beyond uncertainty, trust in AI is also shaped by ethical considerations such as fairness and accountability. Users may tolerate some degree of uncertainty if they perceive the system as fair and transparent, but if users believe an AI system is unfair, biased, or lacks accountability, particularly when AI operates in sensitive areas such as hiring, law enforcement, and medical diagnosis, trust erodes. Ensuring fairness in algorithmic decision-making, improving transparency around ethical considerations, and implementing mechanisms for accountability are crucial for building user confidence.

Trust in AI is not just about ensuring users have confidence in its decisions, it is also about preventing excessive reliance. That's why researchers often use the term calibrated trust or optimal trust . [Khan et al., 2025, Tomsett et al., 2020]. Some users may distrust AI due to concerns over errors or unpredictability, while others may develop excessive reliance on AI recommendations without critically evaluating them . The over-trust can have severe consequences in some domain specifically in healthcare. AI systems should incorporate mechanisms that encourage users to critically engage with recommendations rather than accept them blindly.

AI decision-making is heavily dependent on data. The quality, diversity, and representativeness of the data used to train AI systems directly influence the factors discussed above. Using higher-quality, diverse, and well-curated data can enhance interpretability and improve the explainability of AI outputs. When models are trained on comprehensive and representative datasets, they are more likely to produce reliable predictions that align with real-world scenarios. Additionally, well-structured data allows AI systems to generate clearer, more transparent explanations, enabling users to better understand the reasoning behind AI-driven decisions and fostering a more calibrated level of trust.

1.3.2.4 Designing a Trustworthy AI System

Trust in the HAI is affected by several factors. To design a trustable AI system according to user needs, these factors need to be explored and studied. Earlier studies revealed that users often reject automation when its design rationale is unclear, or they cannot understand the reasons behind its decisions or errors [Dzindolet et al., 2003]. It is still valid in the context of HAI. It is widely accepted that trust in models, which are explainable, transparent and interpretable, outperforms trust in black-box models [Ferrario and Loi, 2022, Zolanvari et al., 2023]. Some researchers also claim that outcome feedback may be more effective than interpretability in building trust [Ahn et al., 2024]. It is also noted that multi-modal explanations that provide reliable and personalized solutions enhance users' understanding and support trust development in AI systems [Papagni et al., 2023].

Other factors that affect user trust include information and system quality, and data quality. [Nicolaou et al., 2013, Ogawa et al., 2023]. With better quality, the user trust increases and decreases with bad quality. The reliability of input data affects user trust, as insufficient or subjective training data can result in biased or overfitted outcomes [Asan et al., 2020]. Also, it's not about the quality of data but the type or types of used by the system and provided to the user.

These factors explain the reasons behind the AI system decisions and provide a rationale for the design of the system. Considering these factors in the design of the system can leads to a high trustable system. However, researchers also caution against risks such as over-trust and automation bias, where users excessively rely on AI despite potential errors [Parasuraman and Manzey, 2010]. [Holbrook et al., 2024] illustrated how over-reliance on AI systems can lead to critical errors in decision-making, even in life-

or-death scenarios, emphasizing the risks of overtrusting these technologies under uncertainty, and highlighting the need for careful trust calibration. Similarly, we have earlier discussed that explanations enhances trust, but overly complex explanations can erode trust [Poursabzi-Sangdeh et al., 2021a]. Therefore system should be designed in such a that it creates a balance in all these factors so that the user can have an optimum or calibrated trust in the system.

1.4 User Acceptance

Like trust, user acceptance is also a multidimensional concept and has been explored in many disciplines such as psychology, human-computer interaction and automation, and AI. Broadly, user acceptance refers to an individual's willingness to engage with and adopt a system, influenced by cognitive, emotional, and contextual factors and the extent to which the system meets their needs and expectations.

1.4.1 Definitions

Acceptability or acceptance is a term that is used widely. Oxford Dictionary defines acceptance as "the process of allowing someone to join something or be a member of a group". According to Merriam-Webster, acceptance is defined as "the quality or state of being acceptable" [Merriam-Webster, 2024]. The term acceptable is further described as "capable or worthy of being accepted" while accepted is defined as "to make a favourable response" [Oxford, 2024a], Connecting these definitions shows that acceptance involves acknowledging or approving something (giving a positive response) based on its perceived value or suitability.

In the technology domain, user acceptance has been discussed in various contexts. In an earlier study, [Gall et al., 1977] associated user acceptance with tangible benefits perceived by users. However, a concrete definition of user acceptance is provided by [Dillon and Morris, 1996], which defines user acceptance as "the demonstrable willingness within a user group to employ information technology for the tasks it is designed to support. This definition emphasizes the behavioural aspect of acceptance, focusing on users' willingness to adopt technology to accomplish specific goals. However, the concept is quite big and complete frameworks and models were needed to explain it.

1.4.2 User Acceptance in Automation, Technology and Human-Centred AI

The concept of user acceptance was not explicitly defined in the context of automation within human factors literature before the 90's. However, related constructs such as user satisfaction and system acceptance were explored. Early work emphasized the degree to which individuals perceive a system as meeting their needs or supporting their tasks effectively.

A study by [Swanson, 1974] discussed the concept of appreciation and relate it with involvement. They hypothesized that involvement leads to the appreciation of the system which is closely related concept to acceptance. Similarly, [Maish, 1979] explored how factors such as management support, user attitudes, and organizational context shape user interactions with information systems. They focused on different user behaviours like requesting new system features, and increasing terminal usage, examining how these behaviours reflect the alignment of the system with user needs. Further, [Ives et al., 1983] introduced the concept of User Information Satisfaction, defining it as "the extent to

which users believe the information system available to them meets their information requirements". Although these studies did not explicitly define user acceptance, they lies in close proximity.

In modern information system research, user acceptance has been one of the most thoroughly explored topics. Its concept was formally introduced by [Davis, 1989] through the Technology Acceptance Model (TAM), which became a foundational framework in the field. TAM identified two key constructs: perceived usefulness, the belief that a system will improve job performance, and perceived ease of use, the belief that using the system will require minimal effort. Since its introduction, researchers have extended TAM by adding constructs and relationships to better explain user acceptance. These models draw from fields like information systems, psychology, and sociology, with notable contributions from many researchers [Taylor and Todd, 1995, Bajaj and Nidumolu, 1998, Dishaw and Strong, 1999, Venkatesh and Davis, 2000, Venkatesh et al., 2003, Yi et al., 2006, Ghazizadeh et al., 2012].

The work by [Venkatesh et al., 2003] is very prominent, they introduced the Unified Theory of Acceptance and Use of Technology (UTAUT) model. The core concept underlying UTAUT is illustrated in Figure 1, which highlights the progression from individual reactions to using information technology to intentions to use technology, ultimately leading to actual use. This linear flow emphasizes the psychological and behavioural mechanisms driving user acceptance, offering a foundation for structured frameworks. In the model, they identified four determinants (performance expectancy, effort expectancy, social influence, and facilitating conditions) of user acceptance that influence behavioural intention and subsequent use behaviour. Performance expectancy refers to the perceived usefulness of technology in improving performance, while effort expectancy focuses on the ease of use. Social influence captures how others' opinions shape users' intentions, and facilitating

conditions address the support and resources available to the user. These determinants are moderated by variables such as gender, age, experience, and voluntariness of use.

The UTAUT model is further extended by [Venkatesh et al., 2012] to study user acceptance in a consumer context, laying the groundwork for its adaptation to AI and XAI. Similarly [Gursoy et al., 2019] modified UTAUT to develop the AI Device Use Acceptance model to explain customer acceptance of AI devices in service interactions through a three-stage process involving appraisal, evaluation, and outcomes.

In healthcare, [Huang et al., 2024] adapted the UTAUT2 model to assess patients' acceptance of AI-driven healthcare services, focusing on factors like performance expectancy and social influence. UTAUT is also employed by [Tran et al., 2021] to analyze the cognitive determinants influencing medical students' intentions to adopt AI-based clinical decision support systems (CDSS). [Fujimori et al., 2022] further used UTAUT to study physicians' acceptance of AI-based CDSS in emergency care, examining facilitators and barriers to adoption while focusing on their perceptions of real-time AI tools.

More recently UTAUT has also been applied to Generative AI. [Du and Lv, 2024] used UTAUT to measure which constructs have a strong effect on user acceptance of generative artificial intelligence among students. Similarly, [Chen et al., 2024] extended UTAUT with trust and perceived risk to study AI chatbot adoption among research scholars, finding that social influence, trust, and facilitating conditions were key predictors.

This shows that user acceptance remains a critical factor in the adoption and integration of AI and XAI technologies across various domains, including healthcare. The adaptability of models like UTAUT and its extensions demonstrates their relevance in capturing the complex dynamics of user behavior and technology adoption. In Chapter 7, we further explore the relationship between data modality and user acceptance by employing

a modified version of UTAUT, tailored to the specific context of our study.

1.4.3 Challenges in User Acceptance of AI

In Human-AI Interaction (HAI), user acceptance is influenced by various factors, including trust, transparency, ethical considerations, usability, and data modality alignment. These challenges are deeply interconnected, collectively shaping the willingness of users to adopt AI-driven tools across different domains. Addressing these challenges effectively is crucial for ensuring successful and streamlined HAI.

A fundamental factor in AI adoption is trust, which is closely tied to transparency. Many AI-driven systems, particularly those based on deep learning, operate as black boxes, making it difficult for users to understand how decisions are made [Hassija et al., 2024]. This lack of interpretability can lead to skepticism and hesitation in user acceptance [Wanner et al., 2022]. Explainable AI (XAI) seeks to address this by making AI decision processes more transparent, thereby enhancing user trust and ultimately increasing user acceptance [Baroni et al., 2022].

Bias and fairness issues is another important factor that affects user acceptance. AI systems learn from historical data, and if this data contains biases, the AI model may inadvertently reinforce existing inequalities [Ntoutsi et al., 2020]. This can lead to unfair treatment of certain user groups, diminishing confidence in AI-driven decision-making. Addressing this bias through diverse data representation, rigorous testing, and continuous monitoring is essential to ensuring fair and unbiased AI applications.

These biases if not addressed, give rise to ethical and privacy concerns that also play a significant role in user acceptance [Ostrom et al., 2019]. AI systems often require access to vast amounts of personal data, raising concerns about data misuse, unauthorized access,

and potential biases in decision-making. These issues can deter users from engaging with AI technologies, particularly in sensitive domains such as healthcare and finance. Implementing data protection measures, adhering to ethical guidelines, AI-driven decisions are essential to increase user acceptance.

Along with the above factors, an AI system must be useful and easy to use to have a high acceptance by user according to TAM [Kelly et al., 2023]. However, if AI systems are overly complex, require extensive training, or fail to provide clear advantages over existing methods, users may resist adoption. Designing intuitive, user-friendly interfaces and demonstrating the tangible benefits of AI integration are critical strategies for increasing acceptance.

Resistance to change is another common barrier to AI adoption. Many users, particularly those in traditional industries, may fear job displacement or disruption to established workflows [Dwivedi et al., 2021]. If AI systems are perceived as replacements rather than augmentations of human capabilities, resistance can grow. To counteract this, it is crucial to emphasize the collaborative nature of AI, illustrating how it enhances human decision-making rather than replacing it.

1.5 Data Modality and AI: Implications for User Trust and Acceptance

A critical but often overlooked aspect of trust and user acceptance in AI is data modality, the type of data an AI system processes, such as text, images, audio, or sensor data. Different modalities introduce distinct challenges in AI performance, user trust, and user acceptance. Addressing these challenges is essential to ensure that AI systems align with

user expectations and decision-making processes.

1.5.1 Data Modality and Trust in AI

Trust in AI systems is influenced by the relationship between the data modalities an AI model utilizes and those valued by human decision-makers. The use of the same or different data modalities may impact decision-making processes, potentially shaping user trust in AI recommendations. For instance, in ophthalmology, AI models for glaucoma diagnosis may analyze fundus images, optical coherence tomography (OCT) scans, patient history, and demographic background data. However, clinicians may prioritize patient history and symptoms in their diagnostic process. When an AI system heavily relies on imaging data while overlooking contextual patient information, even if its conclusions align with those of a doctor, trust issues may still arise due to differences in how decisions are reached. A clinician may become skeptical of the AI's conclusions if they perceive it as disregarding crucial non-imaging factors. Understanding and addressing how data modality affects trust is crucial for designing AI systems that foster well-calibrated trust in their recommendations.

While existing literature extensively addresses trust challenges in AI generally as mentioned in Section 1.3.2.3, the specific impact of data modality alignment or mismatch on user trust remains relatively unexplored. This gap shows the need to investigate how using similar or different modalities in AI affects user trust. In Chapter 6, we explore this gap by examining whether the modality of an AI system influences user trust and accuracy and if trust levels differ when users interact with unimodal versus multimodal AI systems.

1.5.2 Data Modality and User Acceptance of AI

User acceptance of an AI system is influenced by the data modalities the AI utilizes. Different user groups rely on specific types of data in various domains, and when an AI system employs certain modalities over others, its acceptance may be affected. The impact of this alignment is not necessarily positive or negative, it depends on the context and user expectations.

For example, in healthcare, clinicians typically integrate multiple sources of information, such as patient history, symptoms, and imaging results, whereas an AI system may rely primarily on imaging data. If this approach aligns with clinician expectations and decision-making processes, it may enhance acceptance. However, in some cases, a system that focuses on a single data modality might streamline decision-making, reduce cognitive load, and improve usability, leading to higher acceptance.

Although user acceptance is extensively studied as mentioned in Section 1.4, the effect of AI system modality on user acceptance is insufficiently explored. This gap highlights the importance of investigating how using similar or different data modalities impacts AI acceptance. In Chapter 7, we aim to address this gap through a modified UTAUT model. Specifically, we explore whether and how the modality of AI systems affects user acceptance, comparing acceptance levels between unimodal and multimodal AI systems.

1.6 Research Questions

Although AI systems demonstrate high accuracies as independent systems, they are often unsuitable for deployment as standalone solutions due to associated risks, such as biases in decision-making, vulnerability to attacks, lack of explainability, and ethical concerns

regarding accountability. Moreover, using these systems in a collaborative environment also does not come without a challenge, including integration issues with human decision-making, potential over-trust or under-trust on AI, non-acceptance of AI decisions, and the complexity of balancing automation with human oversight.

Users with varying levels of expertise, such as novices and experts, tend to approach problem-solving and decision-making differently. Extensive research has been done on decision making differences, learning patterns and behavior of novices and experts and expertise development [Chi et al., 1981, Tilden and Tilden, 1985, Sweller, 1988, Benner et al., 1984, Dreyfus and Dreyfus, 1980, Ericsson et al., 1993, Campitelli and Gobet, 2010, Burns, 2019, Uereten, 2023]. These studies has shown that experts rely on deep domain knowledge, pattern recognition, and heuristics, while novices depend more on structured guidance, rule-based learning, and trial-and-error strategies. However, existing AI-driven decision-support systems often adopt a one-size-fits-all approach, failing to accommodate these fundamental differences in cognitive processes. While many studies have explored AI-assisted decision-making, they primarily focus on improving accuracy and efficiency without considering how AI should adapt to users' varying expertise levels specifically while using AI system in collaboration with humans in healthcare domain. There is a significant gap in research on designing human-centered AI systems that dynamically adjust their interactions based on user expertise.

In the first part of this thesis work comprised of chapters 2, 3 and 4, we address this gap by examining the differences between novices and experts in the decision-making process of glaucoma diagnosis and management. However, our goal is not merely to identify these differences but to derive actionable insights for designing AI systems that effectively support both user groups. By understanding how expertise influences decision-making, we aim to establish guidelines for developing human-centered AI systems that dynamically

adapt to users' needs. Ultimately, this research will contribute to the design of AI-driven decision-support systems that foster more effective collaboration between novices and experts, particularly in healthcare, and can be extended to other domains.

Another significant challenge lies in the misalignment between the data modalities used by AI systems and those clinicians rely on during decision-making. For example, an AI system trained primarily on Fundus images may generate outputs that are difficult to interpret for clinicians who make diagnostic judgments by integrating multiple sources such as fundus images, visual fields, intraocular pressure, and patient history. When the AI system does not incorporate or reflect the full range of information typically used in clinical reasoning, its outputs can appear incomplete or disconnected from real-world workflows. This disconnect reduces the relevance and interpretability of the AI's recommendations, leading to lower trust and reduced likelihood of adoption. We hypothesize that alignment between AI data modalities and the clinician's decision-making process is critical for trust and acceptance. Systems that reflect the actual inputs clinicians use are more likely to be perceived as supportive rather than intrusive, fostering better human-AI collaboration.

Based on these challenges, this thesis seeks to investigate the complex relationships between user expertise, AI system data modalities, trust, and acceptance. The specific research questions I aim to address are as follows:

1. How do decision-making processes and data prioritization differ between novice and expert optometrists when analyzing the same information, and what role does experience play in adapting to varying levels of data availability?

2. What factors influence the decision-making strategies of novice and experts and how can AI systems help bridge the gap in their approaches to decision making in complex scenarios?

3. How a human centered AI system should be developed to serves users with diverse range of expertise?

4. Does modality of an AI system affect users' trust and accuracy and whether there is a difference in trust when user interact with unimodal and multimodal AI systems?

5. Does modality of an AI system affect users' acceptance and whether there is a difference in acceptance when user interact with unimodal and multimodal AI systems?

By answering these questions, this research aims to provide insights into how AI systems can be designed to better support diverse user groups, fostering trust, acceptance, and effective collaboration.

1.7 Objectives

This thesis investigates the differences in decision-making strategies between novice and expert optometrists and examines how AI system data alignment can enhance human-AI collaboration. Specifically, we explore how decision-making processes and data prioritization vary between these groups and the role of experience in adapting to different levels of data availability. Additionally, we analyze the factors that shape their decision-making strategies and assess how AI can bridge the gap in complex scenarios.

Furthermore, we examine how human-centered AI systems can be designed to accommodate users with diverse expertise levels. A key focus is the influence of AI system modality on user trust, accuracy, and acceptance, particularly in the context of unimodal versus multimodal AI interactions. By addressing these research questions, this study

aims to provide insights for designing AI systems that foster effective collaboration across expertise levels.

To systematically explore the research questions, this thesis employs a structured approach that integrates empirical studies and theoretical analysis.

1. Decision-Making Processes and Data Prioritization Across Expertise Levels To address question 1, this thesis investigates how novice and expert optometrists differ in decision-making strategies and data prioritization when analyzing the same information in the context of glaucoma in chapter 3. It examines how experience influences reliance on heuristics, particularly when data availability varies. Through controlled experiments, where both groups perform identical glaucoma diagnostic tasks, and task-based evaluations of patient data, we assess how expertise shapes decision adaptation and the prioritization of critical clinical information. These insights will enhance the understanding of expertise-driven differences in glaucoma assessment and inform the development of AI systems that effectively support decision-making across varying levels of expertise.

2. Factors Shaping Decision-Making and AI's Role in Bridging the Expertise Gap To address question 2, this thesis examines the cognitive, contextual, and experiential factors that shape decision-making in novices and experts in the context of glaucoma in chapter 4. It investigates differences in the factors influencing decisions related to glaucoma diagnosis and staging, identifying key themes in the justifications provided by both groups. Additionally, it explores how novices and experts balance positive (supporting) and negative (opposing) factors in their decision-making process, offering insights into variations in staging decisions. Finally, the study assesses how AI can bridge the expertise gap by supporting novices in decision-making while en-

hancing consistency among experts. Through empirical analysis, these findings will contribute to the development of AI-driven decision-support systems that improve clinical reasoning and diagnostic accuracy across expertise levels.

3. The Influence of AI System Modality on Trust and Accuracy To address question 3, this thesis examines whether the modality of an AI system (unimodal vs. multimodal) impacts user trust and decision accuracy in chapter 6. We evaluate how trust varies with different AI inputs and whether expertise levels moderate this relationship. By analyzing user interactions with unimodal and multimodal AI system in controlled experiments, we assess the measurable impact of AI modality on decision accuracy and user trust.

4. The Role of AI System Modality in User Acceptance In response to question 5, this research investigates how AI modality affects user acceptance, particularly comparing unimodal and multimodal AI systems in Chapter 7. We also explore whether initial acceptance levels shift after system interaction and whether multimodal AI maintains user acceptance despite occasional errors. Longitudinal studies will help determine how user perceptions evolve over time.

5. Developing a Human-Centered AI System for Users with Diverse Expertise Addressing question 3, this thesis synthesizes findings from the above investigations to propose design recommendations for AI systems that effectively support users with varying expertise through the discussion in Chapter 8. By integrating insights on decision-making, AI trust, and user acceptance, we outline principles for developing AI systems that align with human cognitive processes and facilitate optimal human-AI collaboration.

By systematically answering these research questions, this thesis provides a compre-

hensive understanding of decision-making variations and the role of AI in fostering trust, acceptance, and collaboration. The findings will offer actionable insights for designing AI systems that enhance decision-making support across different expertise levels in critical domains.

1.8 Case study of Glaucoma Diagnosis Decision-Making

For this thesis work and to answer the research questions, we chose glaucoma as a case study. In Chapter 2 we further provide details on glaucoma and basis for using glaucoma as a case study for answering our research question.

1.9 Thesis Structure

This thesis is organized into eight chapters, each building on the preceding ones to explore the role of user expertise, trust, and AI system modalities in collaborative decision-making within the context of glaucoma diagnosis. Following this introductory chapter, Chapter 2 explores glaucoma as a clinical condition, its causes, effects, and the complexities of its diagnosis. It also examines the diagnostic process, the types of data collected, and the apparatuses used, justifying its selection as the focus of this research.

Chapter 3 investigates diagnostic decision-making variability among novice and expert optometrists, analyzing how their differing approaches can inform the design of AI systems tailored to diverse expertise levels. This chapter 3 is already published as a paper in *JMIR medical Informatics*. Building on this, Chapter 4 dives deeper into the reasoning patterns of

novices and experts, exploring how they justify their decisions and weigh evidence for and against their choices. This chapter highlights the variability in reasoning that influences how users interact with AI systems. This chapter is submitted to Elsevier *j. of Health Policy*.

Chapter 5 summarizes the findings in Chapters 3 and 4 and make a case for AI system design and connect it with study in Chapters 6 and 7 and further design the methodology for this next study.

Chapter 6 shifts focus to trust and accuracy in AI, comparing clinicians' perceptions of unimodal and multimodal AI systems for glaucoma diagnosis. It examines why multimodal systems tend to inspire greater trust and achieve higher acceptance among clinicians. This chapter is submitted to Elsevier *J. of Computer in Biology and Medicine* and currently under review. Chapter 7 further explores the role of modality in clinician acceptance, investigating how interactions with unimodal and multimodal AI systems, including those involving errors, affect their willingness to adopt AI as a collaborator.This chapter is submitted as a paper to IEEE *J. of Translational Engineering in Health and Medicine* currently under review. The basic idea of Chapters 6 and 7 is published as an abstract in *HFES health care conference 2025*.

Finally, Chapter 8 synthesizes the findings from the study, discussing their implications for human-centered AI system design in healthcare. This chapter also reflects on the study's contributions, acknowledges its limitations, and provides recommendations for future research aimed at improving human-AI collaboration. Together, these chapters provide a comprehensive exploration of how clinicians engage with AI systems in a complex, high-stakes domain like glaucoma diagnosis, offering insights that can inform the design of more effective and trustworthy AI tools.

Chapter 2

Case Study of Glaucoma

The chapter provides details on glaucoma diseases and justifies its selection as the case study for this research through a preliminary study.

It provides a foundational understanding of glaucoma diagnosis, highlighting its complexity and the reliance on diverse diagnostic tools and data. This complexity makes glaucoma an ideal case study for exploring the variance in decision making and effect of AI system data modality on user trust and acceptance.

2.1 Introduction

The World Health Organization reported in 2019 that about 2.2 billion people have some form of visual impairment [Organization, 2019]. About half of these cases could have been prevented. It is expected that these numbers will increase everywhere as the size of the aged population increases. For instance, the number of people in Canada was approximately 0.5 million in 2013, with an additional 50,000 individuals expected to have

sight loss yearly [Sarhan et al., 2019]. Currently, about 1.5 million Canadians have vision loss. It was estimated that in 2016, about C$ 19 billion was spent yearly on blindness-related problems. The projected expense is C$ 25 billion by 2024. In their pre-budget submissions in 2016 and 2021, the Canadian Association of optometrists have stated that 80% and 75% of vision loss cases can be prevented with early diagnosis and by providing quality eye care.

The population over 60 is also increasing at almost double the rate of ophthalmologists, which, coupled with the already overburdened healthcare system, will further intensify the demand for glaucoma care [Resnikoff et al., 2012]. Ophthalmologists will face high volumes and workloads. This increase will ultimately lead to delays in timely glaucoma diagnosis and care. Artificial intelligence (AI) can provide a possible solution to the growing demand as it has the potential to make faster diagnoses with high accuracy. It is fatigue-free and looks for patterns and features that are either not visible to the human eye or humans can miss them easily. However, we cannot allow AI to make standalone decisions because of the potential drawbacks discussed in the above sections. Clinicians can team up with AI systems to help in the fast diagnosis process of diseases. A collaborative environment is needed where humans and AI can make decisions. The collaborative environment poses the same challenges related to data modality, accuracy, trust, and acceptance. These challenges need to be investigated and solved to get the benefits of Human-AI collaboration.

2.2 Objective of the Preliminary Study

The main objective of the study was to understand the disease, and lay a foundation for exploring variations in doctor's decisions. We specifically answered the following questions in this prestudy:

1. What is glaucoma and what are its causes and effects?
2.
3. How is glaucoma diagnosed, and what is the general diagnosis pathway?
4. What makes the glaucoma decision-making process complex?
5. What kind of data do they collect from a patient and investigate?
6. Which apparatuses do they use?

2.3 Glaucoma Disease and its Diagnosis

Glaucoma is a slow-progressing disease of an eye's optic nerve caused by retinal ganglion cell degeneration. Structural changes in the optic nerve head and functional failure of the visual field (VF) occur with progression. If it is not diagnosed early or left untreated, it can lead to blindness. it is the leading cause of blindness after cataracts worldwide. About 12% of global blindness is because of glaucoma [Fu et al., 2017]. The glaucoma research society of Canada reports that nearly 400,000 Canadians have glaucoma today. Out of these, 250000 have chronic open-angle glaucoma.

Optometrists/ophthalmologists diagnose glaucoma using manual observations of tests. The diagnosis and classification of glaucoma stages solely depend upon an individual optometrist's decision. These decisions are highly biased, so globally, half of glaucoma cases go undetected [Soh et al., 2021]. Glaucoma diagnosis involves several devices and performing a series of tests. It then combines the current states of the eyes with historical health data, ocular data, and external factors. A comparison with patients with the same conditions is then carried out. Finally, with all these in mind, the state of the disease is decided.

The mental model of every optometrist works in a completely different way as their decisions vary even for a patient in the same state. The involvement of so many variables, the slow progression of the disease, non-programmed decisions, and the involvement of the optometrist's mental model make the diagnosis decisions a very complex process.

Because of the complexities, manual observation and decision-making without the help of any intelligent agent are difficult. Therefore, many researchers have developed artificial intelligence systems for the early diagnosis of Glaucoma [Chai et al., 2018, Akter et al., 2022]. These systems can process massive amounts of data, observe the trends and changes invisible to the human eye and therefore it can be used in a clinical environment to assist optometrists in decision-making.

These systems range from using single data sources to multimodal data and accuracies ranging from low to high. Besides data and accuracy, the models trained in these systems also vary from simple models, such as decision trees and support vector machines, to deep learning models. Recently many companies have approved the deployment of AI screening systems for various eye diseases. For example, Eyenuk got approval for diabetic retinopathy screening systems called EyeArt AI from the FDA, CE marking, and Health Canada (Eyenuk, 2020). In addition, iPredict recently got CE certification as an automated system for diabetic retinopathy, age-related macular degeneration, and glaucoma suspect (iHealthScreen Inc., 2022).

2.4 Glaucoma Pathways

To understand glaucoma and its complex diagnosis procedure, we conducted a small observational study and extensively discussed it with an optometrist (Dr. Furtado, an expert

in glaucoma diagnosis). Glaucoma diagnosis is a complicated process involving many diagnostic parameters, and different optometrists take different pathways.

A generalized view of the glaucoma diagnosis pathway is shown in Figure 2.1. The glaucoma diagnosis involves sequential exams. In an initial eye exam, there are two steps. In the first step, the optometrist gathers the history of the patient. In this phase, if a optometrist finds any eye-related abnormalities, the optometrist then moves to the second step of the exam, where vision tests take place if the patient needs glasses/lenses or if there is any change in glasses. The initial examination often serves as a baseline reading for glaucoma-related diagnosis and creates an initial alert for suspicion. In the following subsequent exam, optometrists repeat the eye pressure exam, carry out an Optical Coherence Tomography (OCT), examine the anterior segment of the eye, the optic nerve head and measure the corneal thickness.

Based on the patient's longitudinal data and comparison with similar patients, optometrists then decide on the patient's glaucoma suspicion level. The subsequent visit of the patient is determined based on the suspicion level. If they notice any changes in subsequent visits and the patient's glaucoma suspicion increases, they classify the patient as an active glaucoma case. The treatment plan is then decided according to the patient's severity and the reading changes due to glaucoma.

Based on the previous studies [Leske et al., 2008, Le et al., 2003, de Voogd et al., 2006, Gordon et al., 2002] and our preliminary study, we can say that both the physical state of the eye and external factors hugely influences the decision-making about glaucoma diagnosis. The factors are shown in Table 2.1.

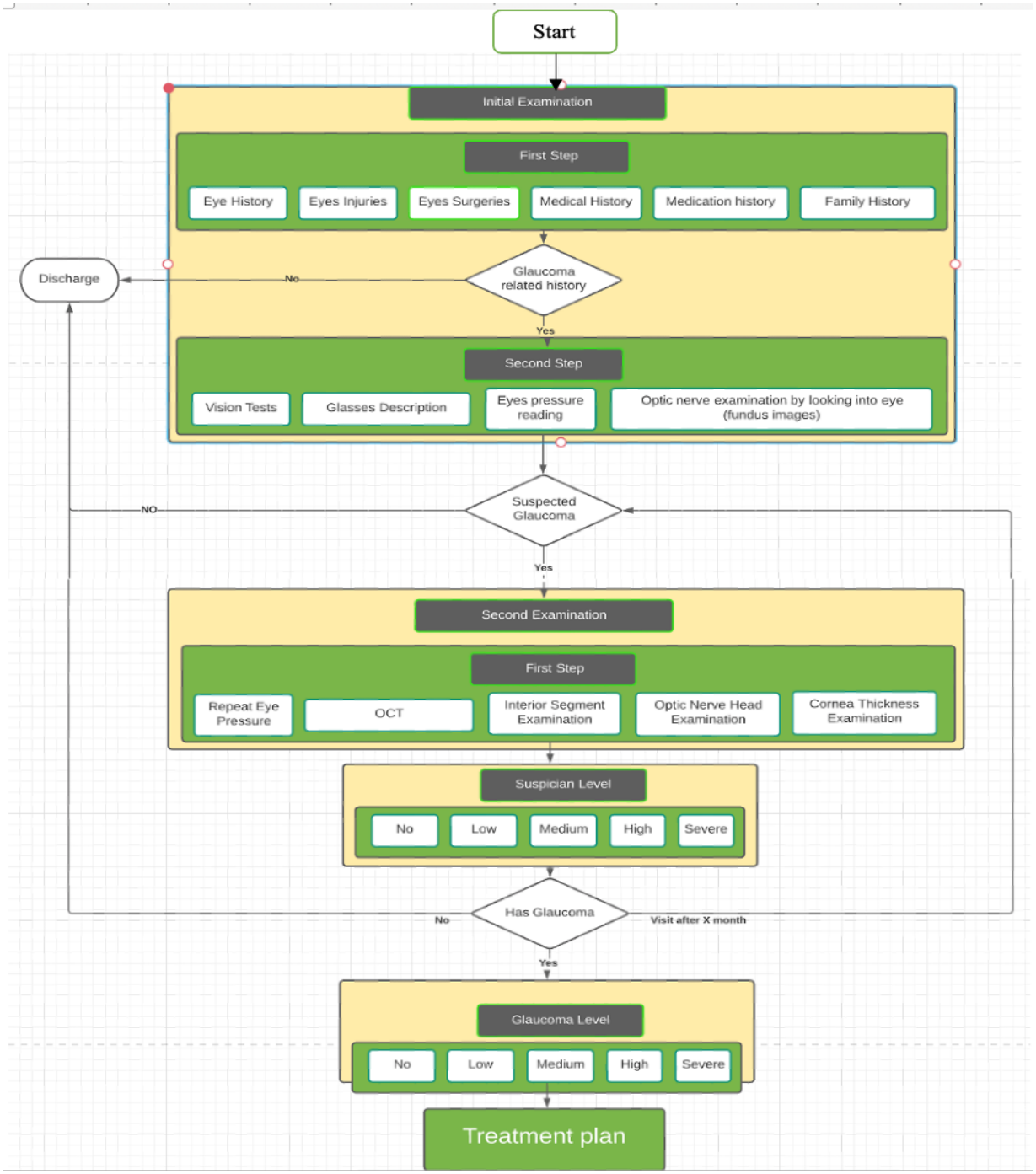


Figure 2.1: A generalized view of the glaucoma diagnosis pathway

Category	Factors
External	Risk group (ethnicity), Family history, Medical conditions, Glaucoma suspicion level, Glaucoma severity level, Age, Medication history
Internal	High intraocular pressure, Corneal thickness/thinness, Extremely farsighted or near-sighted / changes in peripheral vision, Eye injuries/surgeries, Nerve cell thinness, Disease progression / changes in comparison to the previous visit, Optic disc to optic nerve ratio

Table 2.1: Risk factors of glaucoma

2.5 Apparatus and Exams

During the glaucoma diagnosis examination, many different types of apparatus are used. These are shown in Table 4.2.

2.6 Outcome of the Preliminary Study

Through the preliminary study, we explored the general pathway of glaucoma diagnosis, found out about the exams and tests a patient must go through and the data the optometrists collect about patient ocular and general health.

2.6.1 Decision-Making Process of Glaucoma

The preliminary study revealed that the diagnosis of glaucoma involves many tests. There are external and internal variables, and the decision of every optometrist is very subjective and varies highly. We predict that expert and novice optometrists' decisions will differ for

Tests	**Apparatus**	**Measures/Examine**	**Normal Range**
Tonometry	Tonometer	Intra-ocular pressure	10-21 mm Hg
Gonioscopy	Lens	Anterior chamber angle	Van Herick method, 0-slit-1-2-3-4
Pachymetry	Pachymeter	Corneal thickness	Periphery: 700-900 µm, Center: 544 µm
VF test	Perimeter	Peripheral and central vision	Glaucoma hemifield test: Outside normal limits, Borderline, General reduction of sensitivity, Abnormally high sensitivity.
Optical coherence tomography	Spectralis	Posterior segment	Subjective analysis
Optical nerve head	Ophthalmoscope	Nerve head	Mild, early, moderate, advanced, late, and severe stages

Table 2.2: Tests and apparatuses used in glaucoma diagnosis

a single patient. The glaucoma suspicion level decision will also vary among Optometrists, and finally, deciding the severity level of glaucoma will also be different. A patient with the same state may be classified as having open-angle glaucoma by one optometrist and not glaucomatous by another one.

There are several reasons behind these decisions variations:

- The slow progression of glaucoma. Usually, glaucoma takes 10-15 years, from the early to the advanced stage.
- There is no clear boundary between the different stages of the disease.
- The diagnosis requires multiple tests.

- The diagnosis involves many devices.
- Various types of data are to be analyzed for diagnosis decisions.
- The influence of external factors such as age, race, and family history. These factors affect the likelihood of developing glaucoma and may also influence diagnostic decisions. For instance, individuals of African or Hispanic descent have a higher risk of developing glaucoma, and a positive family history increases susceptibility. Additionally, older age is a well-known risk factor, potentially affecting how aggressively an optometrist interprets test results or recommends treatment.

On every visit, the optometrist investigates the patient and analyzes the disease's current state and features. The optometrist then compares new data with the patient's previous visit data related to glaucoma. Along with these, the optometrist keeps the ocular and family history in mind and decides the disease state. As there is the involvement of multiple factors, this decision varies from optometrist to optometrist and might be highly dependent on expertise level. The decision agreement and disagreement are also documented in the literature [Azuara-Blanco et al., 2007, Marks et al., 2012, Moreno-Montañés et al., 2017, Pourjavan et al., 2024, Wang et al., 2024]. Also, manual data interpretation is both time-consuming and a potential source of medical error [Chen et al., 2022a].A unified, universal decision for a patient with the same state is challenging, and that is why we call it a complex decision-making process.

2.6.2 Modeling of observational findings with CWA approaches

Based on our initial discussion of the glaucoma diagnosis decision-making process, we describe the steps on a decision ladder template as shown in Figure 2.2. The decision ladder

is a template and does not correspond to any group of people based on expertise but differences in novice and expert decision making are often modeled on the ladder [Burns, 2019]. The boxes represent the information processing activities, while the oval shows the states of knowledge. The ladder begins with an activation operation on the left leg of the ladder when a patient comes to the clinic for a glaucoma checkup. The optometrist then observes the patient by gathering information about the patient's condition, and the optometrist reaches a state with all the required information. The optometrist then further identifies the risk factors, populates his belief with the new information, and merges the previous knowledge with the latest knowledge. The agent here reaches a new system state. The ladder then moves forward to enter a loop where the optometrists evaluate all the options, choose a goal, and then explore the consequences of the selected goal. The circle breaks, and the ladder moves to the right leg once the optometrist reaches a target which is understanding the patient's glaucoma situation. Based on the understanding, the optometrist selects a task to achieve the target. The tasks are classifying the glaucoma risk, clinic revisiting schedule, and treatment plan. When the task is selected, the next step then is to plan the process to carry out the selected task. The task is executed in the last stage after the procedure is selected. The ladder also shows that some shortcuts can be taken based on heuristics. These shortcuts reveal the difference between handling familiar and unfamiliar situations or the agent's different expertise levels.

2.6.3 Data collection

During the pre-study period, we collected data from patients to know more about the features and attributes involved and prepare data for further stages. As a pre-processing stage, a training session was arranged to learn about the electronic medical record system.

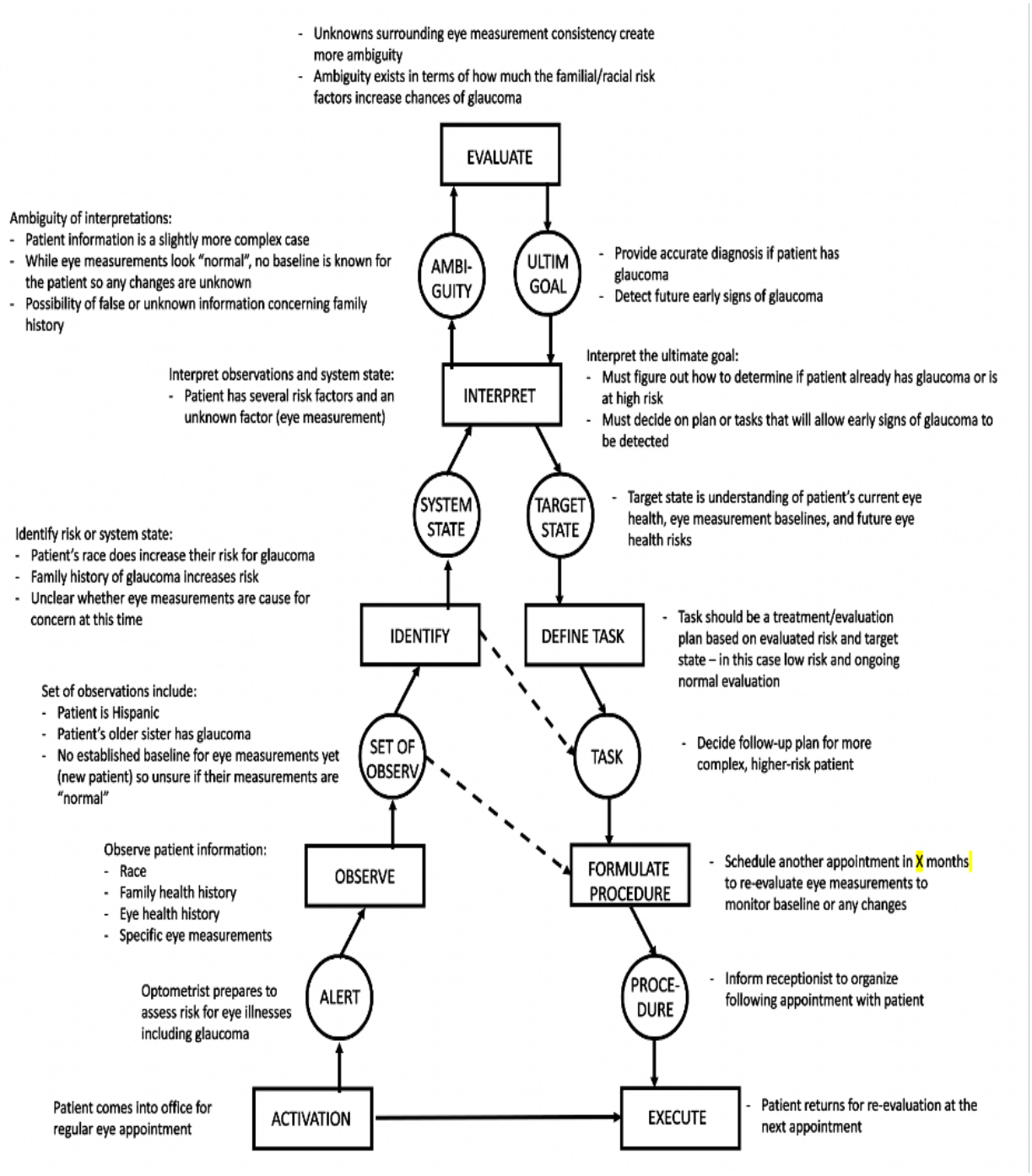


Figure 2.2: Glaucoma diagnosis decision ladder

Those patient records from the system are selected who have granted permission to use the data for research purposes. Initially, 20 patients were selected, but in the end, only eight patients' data was collected. The reason for collecting only seven patients' data is that we cannot review a more significant number of records by optometrists because of time constraints. On averagely an optometrist spends about 10-15 minutes reviewing a single patient exam and writing the assessment. For seven patients, data from 3 visits were collected, The first visit, the last visit, and one in the middle. For one patient, the complete visit data was collected. The collected patient features are shown in Table 2.3. The primary data attributes are included in the collected data below in Table 2.4.

Category	**Sub-category**	**N**
Status	Glaucomic	3
	Suspected	5
Gender	Male	4
	Female	4
Ethnicity	Caucasian	5
	African American	2
	Asian	1

Table 2.3: Distribution of Status, Gender, and Ethnicity

Attributes	**Details**
External Attributes	Age, Gender, Race
History	Ocular History, Medical History, Medications history, Social history, Family history
Internal Attributes	Habitual Spectacle Rx, Unaided Distance VA, Aided Near VA, Pupils, EOMs, Anterior Segment, Goldman Tonometry, Corneal Thickness, Gonioscopy
Images	Fundus image, OCT images
Reports	Posterior Segment, VF, OCT

Table 2.4: Summary of Attributes, History, Internal Attributes, Images, and Reports

2.7 Data Modalities used in Glaucoma Diagnosis

The data collected is of multiple modalities and these different types of data guide optometrist's diagnostic and treatment decisions for glaucoma. Since glaucoma is a progressive condition that affects the optic nerve, its assessment requires both structural and functional evaluations, supported by contextual patient information. The primary modalities that this collected data can be classified as into 4 groups including Optical Coherence Tomography (OCT), Visual Field (VF) tests, Fundus Imaging, and Textual Data. Each of these provides unique insights, helping optometrists assess disease presence, progression, and impact.

OCT plays a critical role in detecting early structural changes in the retina, particularly in the retinal nerve fiber layer (RNFL) and ganglion cell complex, where glaucoma damage often begins. VF tests, on the other hand, assess functional vision loss, allowing clinicians to understand how the disease affects a patient's sight. Fundus Imaging offers a broader view of the optic nerve head, capturing important details such as cup-to-disc ratio changes and neuroretinal rim thinning, which are key indicators of glaucomatous damage. Textual Data, encompassing medical history, systemic health conditions, medications, background information, and past clinical evaluations, provides a broader context that helps optometrists tailor their decisions to individual patient needs.

2.8 Generalization of Studying Glaucoma to other Contexts

One of the aims of this pre-study was to explore whether glaucoma can serve as an effective case study for our research questions and can be generalized to other domains. Through

this exploration, it became evident that the characteristics of glaucoma diagnosis align with challenges seen in other domains, allowing for the generalization of insights beyond optometry.

Given the involvement of multiple data modalities, various diagnostic devices, and different types of clinical assessments, it was anticipated that, as in other domains, decision-making in glaucoma diagnosis might vary based on expertise. If properly explored, these variations can provide valuable direction for developing AI systems that cater to different expertise levels and can be applied to other fields. As similar variability is observed in disciplines such as pathology, where specialists interpret biopsy slides differently, cybersecurity, where analysts assess threats based on their experience, and legal decision-making, where attorneys apply different reasoning strategies when weighing evidence, the variability in glaucoma decision making and AI system design guidelines for different expertise levels can be generalized to these other domains.

Another key characteristic is multimodal data integration, where decision-making relies on multiple sources of information. In glaucoma, optometrists analyze structural imaging (OCT, Fundus Imaging), functional assessments (VF tests), and patient history to reach a diagnosis. Similar challenges exist in radiology, where MRI and CT scans must be interpreted alongside patient records, and in financial risk assessment, where analysts consider economic indicators, historical data, and market trends to inform decisions. AI systems for all these domains, including glaucoma, can be developed using different modality data. Studying the data modality of glaucoma detection and management AI systems and its effect on trust and acceptance provides a foundation in these other domains and can be closely correlated, shaping the design and adoption of AI-driven support systems across various fields.

Additionally, progressive and uncertain conditions requiring continuous monitoring are

a defining feature of glaucoma diagnosis. Since glaucoma develops gradually and requires long-term assessment, optometrists must track subtle changes over time to determine disease progression. This characteristic is found in other medical fields, such as neurology (monitoring Alzheimer's or Parkinson's disease progression) and oncology (tracking tumor growth and treatment response), as well as outside medicine in areas like predictive maintenance, where engineers monitor equipment degradation over time. In all these cases, early detection and trend analysis are key to effective intervention.

Finally, human-AI collaborative decision-making is an emerging theme across disciplines. In glaucoma diagnosis, AI has the potential to support optometrists by identifying patterns that might be difficult to detect manually, improving efficiency and diagnostic accuracy. Studying this approach can be generalized to other domains where human expertise and AI need to work together, such as radiology (where AI assists in detecting anomalies in medical imaging), finance (where AI-driven risk assessments complement human judgment), and cybersecurity (where AI-enhanced threat detection supports analysts in identifying cyber threats). Understanding how AI can effectively integrate into human decision-making processes is crucial for ensuring that AI-driven systems provide meaningful support while allowing experts to maintain oversight and control.

By using glaucoma as a case study, this pre-study establishes that it is a strong candidate for studying decision-making variability and AI integration in complex diagnostic settings.

2.9 Conclusion

This chapter establishes glaucoma as an ideal case study for examining decision-making variability and the role of AI data modality in shaping trust and acceptance. The prelimi-

nary study highlights the complexity of glaucoma diagnosis, driven by its slow progression, reliance on multiple tests, and the subjective nature of optometrist assessments influenced by expertise levels. By mapping the diagnostic process and identifying key risk factors and data modalities, the study provides a foundational understanding of glaucoma and optometrists' decision-making. To develop an AI system that effectively assists in glaucoma diagnosis, it is essential to first explore how optometrists with different expertise levels interpret and integrate diagnostic data, as well as how their decision-making patterns differ.

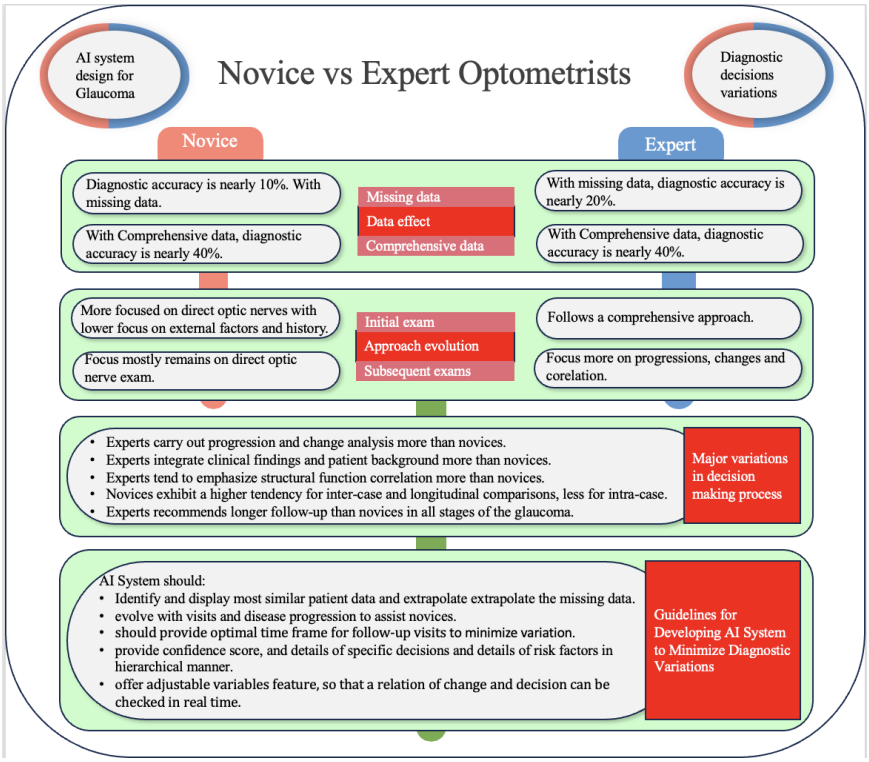
Chapter 4 further investigates these diagnostic variations, analyzing differences in decision-making between novice and expert optometrists. The insights gained from this analysis will guide the design of AI systems that foster a human-AI collaborative environment, ensuring that AI complements clinical expertise while maintaining flexibility and trust in decision-making.

Chapter 3

Diagnostic Decision-Making Variability in Glaucoma: Comparative Analysis of Novice and Expert Optometrists to Inform AI System Design

This chapter is based on the following published article:

Ghaffar F, Furtado N, Ali I, Burns C. Diagnostic Decision-Making Variability Between Novice and Expert Optometrists for Glaucoma: Comparative Analysis to Inform AI System Design. *JMIR Medical Informatics*. 2025;13:e63109. doi: 10.2196/63109



Graphical Abstract

3.1 Abstract

3.1.1 Background

This study examines the diagnostic decision-making differences between novice and expert optometrists in glaucoma diagnosis. While expert optometrists tend to rely on a deep understanding of the disease and intuitive pattern recognition, those with lesser experience may depend more on extensive data, comparisons and external guidance. Understanding these variations is important for developing AI systems that can effectively support optometrists with varying degrees of experience and minimize decision inconsistencies.

3.1.2 Objective

The main objective of this study is to identify and analyze the variations in diagnostic decision-making approaches between novice and expert optometrists. By understanding these variations, we aim to provide guidelines for the development of AI systems that can support optometrists with varying levels of expertise. These guidelines will assist in developing AI systems for glaucoma diagnosis, ultimately enhancing the diagnostic accuracy of optometrists and minimizing inconsistencies in their decisions.

3.1.3 Methods

We conducted in-depth interviews with 14 optometrists using within-subject design, including both novices and experts, focusing on their approaches to glaucoma diagnosis. The responses were coded and analyzed using a mixed-method approach incorporating both qualitative and quantitative analysis. Statistical tests such as Mann-Whitney and Chi-square tests were used to find significance in inter group variations. These findings were further supported by themes extracted through qualitative analysis, which helped to identify decision-making patterns and understand variations in their approaches.

3.1.4 Results

Both groups showed lower concordance rates with clinical diagnosis, with experts showing almost double (20%) concordance rates with limited data in comparison to novices (10%) highlighting the impact of experience and data availability on clinical judgment; this rate increased to nearly 40% for both groups when they had access to complete historical data of the patient. We also found statistically significant inter group differences between the

first visits and subsequent visits with a p-value of less than .05 on the Mann-Whitney test in many assessments.

Furthermore, approaches to the exam assessment and decision differed significantly: experts emphasized comprehensive risk assessments and progression analysis, demonstrating cognitive efficiency and intuitive decision-making, while novices relied more on structured, analytical methods and external references. Additionally, significant variations in patient follow-up times were observed, with a p-value of 0.029 on the chi-square test, showing a stronger influence of experience on follow-up time decisions.

3.1.5 Conclusion

The study highlights significant variations in the decision-making process of novice and expert optometrists in glaucoma diagnosis, with experience playing a key role in accuracy, approach, and management. These findings demonstrate the critical need for AI systems tailored to varying levels of expertise. They also provide insights for the future design of AI systems aimed at enhancing the diagnostic accuracy of optometrists and consistency across different expertise levels, ultimately improving patient outcomes in optometric practice.

Keywords: Decision-making; Human-centred AI design, Human Factors; Experts vs Novices differences; Optometry; Glaucoma diagnosis

3.2 Introduction

Artificial Intelligence (AI) has emerged as a transformative force in optometry [Scanzera et al., 2022], particularly for glaucoma diagnosis. In this context, AI in optometry refers to the advanced algorithms that analyze complex data such as fundus

photos and related patient data. These AI systems use techniques of image processing, computer vision, deep learning, and transfer learning to detect and diagnose glaucoma. Glaucoma is a progressive eye disease that often eludes early detection due to its subtle or absent symptoms, which poses a unique challenge for optometrists [Topouzis and Anastasopoulos, 2007]. This challenge along with no clear-cut tests or signs frequently results in variability in glaucoma-related decisions, influenced by factors such as the optometrists' expertise, experience, the subjective interpretation of clinical data, and the use of different equipment [Yoshioka et al., 2015, Scheetz et al., 2019]. This variability is a critical issue that AI is well-positioned to address, potentially standardizing the diagnostic process and offering a more objective analysis of clinical data. Studies have illustrated AI's advanced analytical capability and proficiency in sifting through extensive patient data, including retinal images, visual field tests, Optical Coherence Tomography (OCT) scans, and medical histories [Chaurasia et al., 2022]. This advanced analytical capability enables AI to identify signs of glaucoma with consistency and accuracy, often surpassing that of human optometrists [Scanzera et al., 2022, Mursch-Edlmayr et al., 2020]. For example, in [Akter et al., 2022], researchers found that AI algorithms could detect glaucoma with up to 96% accuracy, markedly higher than the 80% accuracy rate commonly associated with optometrists.

Similarly, in another study [Rogers et al., 2019], 83.4% accuracy is reported in identifying glaucomatous optic neuropathy using AI. These findings show AI's potential to reduce misdiagnoses and address the prevalent decision variability. In a similar vein, in another study [Wang et al., 2021], it is demonstrated that an AI system helped junior radiologists perform at levels close to that of a senior radiologist. This finding is particularly relevant to optometry, with AI having the potential to play a similar role in enhancing the diagnostic abilities of less experienced optometrists. As in [Jammal et al., 2020], researcher have

highlighted that advocating for a more integrated use of AI tools in optometric practices could be a significant step toward standardizing care and reducing diagnostic variability.

However, the effective integration of AI into optometric practice hinges on a comprehensive understanding of optometrists' decision-making processes [Cook, 1993, Lai et al., 2023]. AI promises to offer standardization in clinical decision-making. Yet, the crux of its successful integration lies in deeply comprehending the subtle intricacies inherent in human decision-making. This understanding is especially critical in diagnosing conditions like glaucoma, where subjective assessments and clinical expertise play significant roles. Highlighting the variability in diagnostic approaches between novice and expert optometrists, as demonstrated by studies [Curran et al., 2022, Stainer et al., 2015], shows the importance of tailored AI support. Therefore, AI systems must adapt to different levels of clinical expertise and integrate the diverse decision-making processes of novice and expert optometrists.

Diagnostic decisions in this field vary significantly, influenced by an optometrist's expertise, experience, and cognitive strategies. These differences affect how patient data is interpreted, the importance assigned to various clinical findings, and the resulting diagnostic outcomes.

In this study, we investigate the variance of diagnostic decisions made by optometrists in the context of glaucoma, focusing on how the availability and nature of data influence these decisions. Several key research questions guide our investigation:

1. How do the decision-making processes of expert and novice optometrists differ when they are presented with the same patient data?

2. What specific data do optometrists at different experience levels tend to prioritize in a clinical journey?

3. Are experienced optometrists more inclined to use heuristic approaches when confronted with limited data instead of scenarios with comprehensive data availability?

To address our research questions, we conducted in-depth interviews with novice and expert optometrists, focusing on their diagnostic approaches to glaucoma. Following these interviews, we performed a detailed coding of the responses, enabling us to conduct qualitative and quantitative analyses. This structured approach allowed us to comprehensively examine the decision-making patterns and their implications for AI integration in optometric practice.

Our research aims to comprehensively understand how optometrists make decisions when diagnosing glaucoma, exploring the diversity in their approaches. By exploring the variations in decision-making between novice and expert optometrists, our study aims to pave the way for the future design of AI systems. These envisioned systems are intended to bridge the expertise gap in glaucoma diagnosis. They will complement human judgment to reduce biases and enhance diagnostic accuracy, leading to more consistent and reliable evaluations. The insights gained from this exploration are anticipated to significantly contribute to optometric practice and education, setting the stage for future advancements in glaucoma diagnosis and management.

3.3 Methodology

This section details our research methodology investigating variations in optometrists' decision-making. It begins with an in-depth explanation of the data collection process, emphasizing the criteria and methods used for selection. Subsequently, the stages of data refinement and processing are outlined. Optometrist interviews are discussed, covering

outreach strategies, interview platforms, and participant demographics. A within-subjects design was used for the interviews, allowing each participant to engage with all patient cases. The section culminates with insights into the data coding process, involving transcript cleaning and multidisciplinary analysis. For the data analysis, a mixed methods approach was used, integrating both qualitative and quantitative data to provide a comprehensive understanding of the decision-making processes. Figure 3.1 summarises the flow of the study.

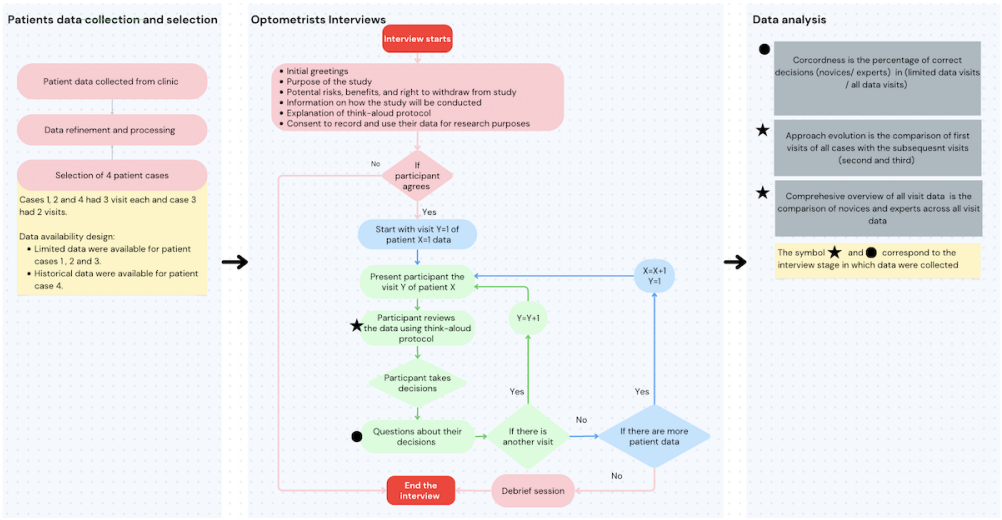


Figure 3.1: The figure shows stages and flow of the study. It started with patient data collection and selection, followed by optometrist interviews and data analysis. Each stage is further subdivided into steps. Each of these steps is further explained in the methodology section. The circle and asterisk symbol shows the data obtained at a specific stage in the interview and how it was analyzed.

3.3.1 Ethical Consideration

This study with application number 43288 is reviewed and approved by the Delegated Research Ethics Committee (DERC) of University of Waterloo on May 27, 2021. The

details are given in Appendix A.1. The informed consent from the participants was sought before the data collection and is given in Appendix A.2. They were informed about the procedure, purpose, benefits, risks, and their right to withdraw from study. Further, they were informed that data acquired from the interview would be anonymized, deidentified and stored on a protected server and used anonymously in all publication that arise from the study. Participants were remunerated $150 CAD for their two-hour commitment, recognizing their professional contribution and encouraging detailed participation.

Data for this study was acquired from the School of Optometry and Vision Sciences at the University of Waterloo, Canada. The selection process entailed scrutiny of patient records within the institution's database, explicitly targeting those records wherein patients had granted explicit consent for their information to be utilized in research. The preliminary phase of our data collection involved gathering records from eight distinct patients. Each patient's record provided a longitudinal perspective, encompassing multiple clinical visits and offering an in-depth view of their interactions with the optometrist over time.

The data attributes collected from patient records were informed by research from previous similar studies [Leske et al., 2008, Le et al., 2003, de Voogd et al., 2006, Gordon et al., 2002], and the European Glaucoma Prevention Study [Group et al., 2007]. Our data collection encompasses a comprehensive range of patient characteristics. It includes essential ocular health metrics and relevant external factors to ensure a comprehensive glaucoma evaluation by optometrists. The patient features we collected are detailed in 3.1. These features reflect the multifaceted factors highlighted in the referenced studies.

Table 3.1: The table shows data features for glaucoma diagnosis collected from the electronic health record system from the School of Optometry, University of Waterloo.

Collected data	**Features**
Demographic factors	Age, Gender, Race
History	Ocular history, Medical history, Medication history, Social history, Family history
Images	Habitual spectacle prescription, Unaided visual acuity, Aided acuity, Pupils, Anterior segment examination, Goldman applanation Tonometry, Central Corneal thickness, Gonioscopy
Reports	Posterior segment examination, Visual field, OCT

3.3.2 Data Refinement and Processing

The data refinement and processing were executed to facilitate the interview component of the study. Out of the initial pool of patients, four records were chosen to be reviewed for the in-depth interviews. This selection process involved comprehensive discussions with an expert in the field to identify cases that could provide diverse diagnostic perspectives. The criteria for case selection were centred on identifying instances that would likely elicit differing opinions among optometrists, thereby enabling a focused examination of the variation in clinical decision-making. The rationale behind limiting the selection to four patient records stemmed from practical constraints associated with the review process. It was anticipated that participants would take about 10-15 minutes to review each patient visit. To manage this effectively within time constraints, we selected three patients with three visits (the initial, final, and intermediary visits) and one patient with two visits. This approach allowed for a detailed review of each patient's clinical journey while staying within the practical limits of the allotted time. In this exercise, historical data from previous visits was not provided for three of the four patients. However, more exhaustive data was provided for one patient, including the complete set of previous visits data, offering a more

detailed case study within the research. This approach provided a balanced perspective between a broad overview and an in-depth case analysis, enhancing the study's ability to capture general trends and specific factors in clinical decision-making.

3.3.3 Optometrist Interviews

In this study, we interviewed 14 optometrists from the Waterloo region, primarily affiliated with the University of Waterloo's School of Optometry and Vision Science, Canada. Each optometrist participated in a 100–120-minute online interview conducted in English via Microsoft Teams and Zoom. The choice of online interviews over physical ones was in response to the participants' preferences.

The participants had an average of 7.54 years of experience (SD = 8.34 years) and saw an average of 32.57 patients weekly (SD = 22.39). Each optometrist scrutinized data from at least ten patient visits, translating to a comprehensive review spanning about two hours for the entire set of patient records. We aimed to present all the visits data to optometrists but because of time constraint, we had to omit some visits. Throughout the interviews, participants employed the 'think aloud' protocol to review patient records, aiding in understanding their decision-making processes. After the first ten interviews, we observed data saturation, which was achieved by the 14th interview, as evidenced by recurring themes and no new substantial information. All interviews were audio-recorded and transcribed using the platforms' built-in features.

3.3.4 Data Coding

Our qualitative research commenced with the cleaning of interview transcripts. This preliminary step involved a thorough review in correcting errors, eliminating irrelevant sec-

tions, and enhancing clarity and coherence. This crucial stage ensured the data's integrity before proceeding to the analysis phase.

Post-cleaning, the refined transcripts were exported into NVIVO (Version 14.23.2). In NVIVO, we carried out an open coding process led primarily by FG. The involvement of multidisciplinary expertise enriched the coding process. With a background in engineering, CB conducted a thorough review of the codes from an engineering perspective. This review added a layer of technical and analytical insight, ensuring that the coding accurately reflected relevant engineering concepts and terminologies. Simultaneously, NF, leveraging her expertise in optometry, reviewed the codes with a focus on the optometry perspective. Her review was pivotal in aligning the data with optometric principles and practices, ensuring the findings were relevant and accurately interpreted within the context of eye care and vision science.

This multidisciplinary approach to reviewing the codes enhanced the depth of our analysis and ensured that the findings were robust, well-rounded, and reflective of diverse professional viewpoints. Such a collaborative effort was instrumental in ensuring that the themes and patterns identified were relevant and applicable across different fields of study. The open coding was executed inductively, allowing the data to drive the categorization process rather than imposing pre-existing theories or frameworks. This approach facilitated a more organic and grounded understanding of the data. Each transcript line was carefully examined, with open codes assigned to summarize the content succinctly. These codes were crafted to stay true to the original text, minimizing abstraction. This level of descriptiveness in coding was instrumental in capturing the essence of the participants' narratives, laying a solid foundation for subsequent conceptualization and analysis. The codes were then grouped to form more generalized codes. Appendix A.5. gives details of the created generalized codes.

Through this process, we distilled the raw data into meaningful themes and patterns, shedding light on the underlying narratives and insights embedded within the transcripts.

3.3.5 Data Analysis

To carry out the analysis, we categorized participants into two groups: novices and experts, based on their years of experience and frequency of glaucoma patient consultations, using median values for greater precision. An 'expert' was defined as having four or more years of experience as a practicing optometrist and treating a higher than average volume of glaucoma patients per week compared to the participant pool. This classification excluded the average number of all patients examined or treated in a week for any eye condition, as it was deemed irrelevant in the context of glaucoma. Participants collectively conducted 137 evaluations, with experts reviewing 48 and novices reviewing 89. Ideally the total number of evaluations would have been 144 but in some visit instances the participants didn't provide us enough data, or we skipped some visit instances in the interview because of time constraint.

Firstly, we examined the variability of novices and experts in diagnostic concordance across all patient visits, focusing on how the availability of comprehensive versus missing data influences concordance rates. Diagnostic concordance is the rate of decisions of optometrists aligned with the original clinical decision taken at a clinic. This analysis revealed disparities between optometrists' decisions and established clinical decisions. Following this analysis, we examined the approach evolution of optometrists within the clinical journey. This analysis was conducted by calculating the assessments/exams conductance rate (the percentage of a specific assessment conducted), 95% confidence interval (CI), and performing statistical analysis using Mann-Whitney U [Mann and Whitney, 1947] test to evaluate

the significance of the variations and approach evolution from initial visits to subsequent visits. To observe a general trend of evolution, we first compared and analyzed initial visits (N=51) with follow-up visits (N=82). To observe the novice and experts' trends, we then further classified the data by experience level of optometrists, comparing initial (N=33) and follow-up (N=55) visits by novices to initial (N=18) and follow-up (N=27) visits by experts. Through this analysis, we investigated the variation in approach evolution and clinical decision-making between novices and expert optometrists.

Secondly, we employed quantitative and qualitative methodologies across all evaluations (N=137) to gain an in-depth understanding of the outcomes of these examinations. This approach enabled us to identify key themes that distinguish the ocular examination practices of novice and expert optometrists. Similar to earlier analysis, we calculated the assessments/exams conductance rate, 95% confidence interval (CI), and performing statistical analysis using Mann-Whitney U test to compare each assessment type of experts and novices. These findings were further supported with identified themes. This mixed method analysis provided insights into the consistency and variability in assessments and diagnostic reasoning. Moreover, by inter-case and intra-case comparison analysis, we identified the cognitive strategies in clinical decision-making, revealing how expertise level influences the patient case comparison.

Lastly, we calculated an average followed up time of experts and novices and then used chi-square test to observe the significance of variation. This analysis offered an understanding of how experience impacts the timelines of patient management. This finding is crucial in glaucoma, where timely intervention can significantly influence disease progression and patient quality of life.

Overall, our study provides a multifaceted look at the complexities of glaucoma management, highlighting the influence of experience on various aspects of patient care. These

insights are invaluable for shaping future educational programs, clinical guidelines, and AI systems in glaucoma, ensuring that they are tailored to address the specific needs and challenges optometrists face.

3.4 Results

3.4.1 Variability in Diagnostic Concordance and Approach Evolution in the Clinical Journey

This section investigates the variability in diagnostic concordance among novice and expert optometrists relative to established clinical decisions, emphasizing the influence of accessible patient data on diagnostic accuracy. It also delineates the evolution of diagnostic and treatment methodologies throughout the clinical journey, from initial patient encounters to subsequent visits.

Significant variability exists in the diagnostic concordance and approach evolution between experts and novices throughout the clinical journey. We compared the diagnostic decisions made by our participants with established clinical decisions, which we define as the decisions previously made by optometrists in a clinical setting where they had access to additional patient data, including comprehensive medical histories and previous examination results. The diagnostic concordance measures the extent to which our participants' decisions align with these pre-existing clinical decisions. We observed significant variations in the decision-making of optometrists, when they lacked prior patient data.

Key findings highlight the disparity in diagnostic decisions when previous patient history is not available, compared to novices (10.1% concordance, 7 out of 69 evaluations),

with experts showing a higher rate of concordance (20.0% concordance, 7 out of 35 evaluations) highlighting the impact of experience on clinical judgment. Interestingly, when provided with prior patient data, the gap in decision-making accuracy between experts (41.6% concordance, 5 out of 12 evaluations) and novices (42.0% concordance, 8 out of 21 evaluations) substantially decreased. This finding shows the critical role of comprehensive patient histories in enhancing diagnostic precision, particularly for less experienced optometrists.

Building upon the insights regarding diagnostic concordance, the optometrists' diagnostic approaches through the clinical journey of a patient reveal significant differences and shifts in practice strategy and emphasis. Initially, there is a higher engagement across all assessments except progression and change analysis and structure and function correlation analysis reflecting common practice for initial assessments. However, in subsequent visits, data reveals a significant shift: the engagement with most assessments decreased, while progression and change, and structure and function correlation, increased. Figure 3.2 provides a graphical representation of these trends in percentage, illustrating the shift from broad-based examinations to focused analytical assessments. Further analysis using the Mann-Whitney U test showed that these changes were significant when comparing initial and subsequent visits. Significant differences were observed in family history ($P < .001$, Cohen d = .41), medical history ($P < .001$, Cohen d = 0.45), patient background ($P < .001$, Cohen d = .43), other risk factors ($P < .001$, Cohen d = 0.50), clinical findings ($P = .01$, Cohen d = 0.28), optic nerve exams ($P = .01$, Cohen d = 0.29), as well as in progression and change analysis ($P = .005$, Cohen d = -0.20). There was no significant change observed in optic nerve function exam and structure-function correlation assessments. The shift in engagement with different assessments shows a pivot towards more in-depth analytical work.

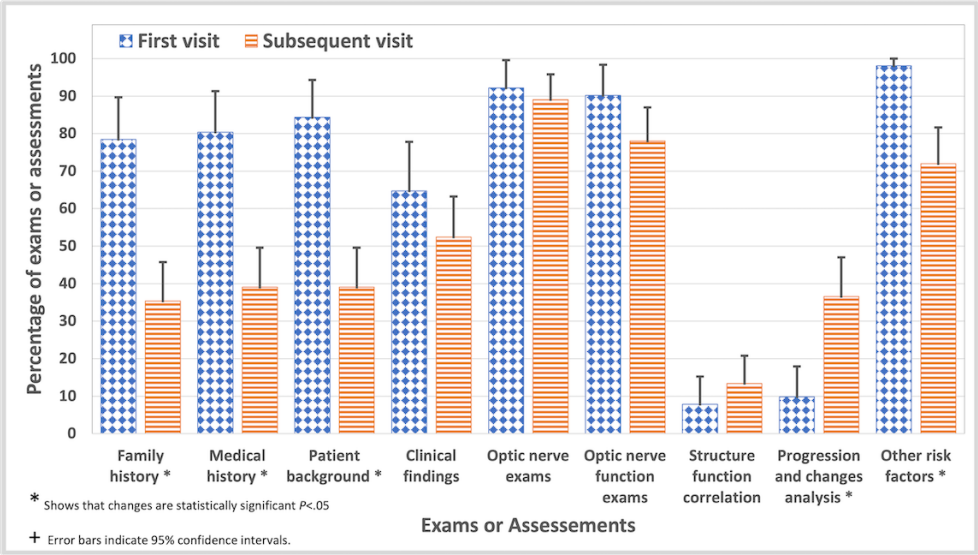


Figure 3.2: The figure shows the percentage increase or decrease in glaucoma-related eye examinations and assessments conducted by optometrists, tracked from the initial visit to subsequent visits through longitudinal patient history. The * with name indicate that there is significant difference with $P<.05$ and the error bar shows the 95% confidence interval.

To investigate further, we compared and analyzed the data of experts and novices separately, summarized in Table 3.2. Detailed descriptive statistics are provided in Appendix A.6. In the initial visit, novices emphasize direct examination techniques, focusing significantly on optic nerve assessment (93.9%) and optic nerve function (90.9%) and comparatively underemphasizing patient history and background information. While experts demonstrate a more balanced and comprehensive approach. They engage more deeply in gathering medical history (83.3%) and patient background (94.4%), while still maintaining high attention to optic nerve assessments (88.8%). Both give considerable attention to other risk factors with experts at 100%, and novices' 96.9%. These results suggest that experts prioritize a more complete picture of the patient's medical background and current health status from the onset. Regarding the analytical methods such as progression and

change analysis, novices show a surprising 12.1% engagement, which demonstrates a higher propensity to engage in advanced analytical methods early on, despite such analyses typically requiring follow-up data. This is possibly due to a misunderstanding of the optimal timing for such analyses.

In the subsequent visits shown in table3.2, there is a noticeable shift. Novices demonstrated sustained attention to optic nerve exams (87.2%) and optic nerve function exams (88.2%). However, their assessment of family history, medical history, patient background, and other risk factors dropped by almost half to that of initial visits. Experts showed a similar behavior with high attention to optic nerve exams (92%), and optic nerve function exams (74%) along with clinical findings (74%), while showing a marked decrease in assessment of family history, medical history, patient background in comparison to initial visits. In the assessment of progression and changes analysis, and structure-function correlation, both groups showed an increase. However, the extent of these increases differs between the two groups; novices showed a modest increase in progression and changes assessment (27.2%) and much smaller increase in structure function correlation (7.2%), while experts showed a robust increase to 55.5% in progression and changes assessment and modest increase to 25.9% in structure function correlation. This leap for experts could reflect a strategic approach where they dive deeper into the more complex analysis. This finding might suggest a more focused approach as specific patient issues become more apparent.

Further analysis using the Mann-Whitney U test showed significant shifts in the assessments from initial to subsequent visits for both novice and experts, as reported in Table 3.2. In this analysis, the weight of each exam was also considered, calculated based on the frequency with which an optometrist discussed a specific assessment. We observed that novices showed significant changes in more assessments with Cohen d values of moderate effect suggesting a strong shift of approach. While the experts showed fewer significant

Table 3.2: The table shows the variation in glaucoma-related eye examinations throughout a patient's clinical journey, highlighting the differences between novice and expert optometrists, tracked from initial to subsequent visits.

Assessment	**Novices**			**Experts**		
	First Visit	**Subsequent Visit**	**P-value (Cohen d)**	**First Visit**	**Subsequent Visit**	**P-value (Cohen d)**
Family History	78.7%	34.5%	< .001 (0.43)	77.7%	37.0%	0.025 (0.44)
Medical History	78.7%	38.1%	< .001 (0.46)	83.3%	40.7%	0.019 (0.25)
Patient Background	78.7%	32.7%	< .001 (0.51)	94.4%	51.8%	0.067 (0.08)
Clinical Findings	57.5%	41.8%	0.01 (0.35)	77.7%	74.0%	ns
Optic Nerve Exams	93.9%	87.2%	0.01 (0.36)	88.8%	92.5%	ns
Optic Nerve Function Exams	90.9%	80.0%	ns	88.8%	74.0%	ns
Structure-Function Correlation	6.0%	7.2%	ns	11.1%	25.9%	ns
Progression & Change Analysis	12.1%	27.2%	ns	6.5%	55.5%	< .001 (-0.7)
Other Risk Factors	96.9%	27.2%	< .001 (0.5)	100%	55.5%	ns

Note. For the Mann-Whitney test, effect size is given by the rank biserial correlation.

changes with Cohen d values of smaller to moderate effect, except for the progression and change analysis where the effect was large suggesting a stronger shift. These findings suggest that experts maintained a comprehensive approach, balancing all assessments with increased emphasis on analytical assessment.

This analysis highlights the critical importance of comprehensive patient data in enhancing diagnostic accuracy, particularly for novice optometrists who face more significant challenges without extensive patient histories. The findings reveal the dynamic evolution of diagnostic and treatment strategies throughout a patient's clinical journey, illustrating how adaptability in practice strategy from initial assessments to subsequent visits is crucial. novices demonstrate a reduced focus on disease progression and correlating variables, indicating a gap in their approach to long-term patient management. This result highlights the potential need for a guiding system to help novices navigate the complexities of diagnostic decisions more effectively.

3.4.2 Comprehensive Overview of Examination Data Across all Visits

We categorized the analyses carried out by optometrists into nine broad codes/categories: family history, medical history, patient background, clinical findings, optic nerve exams, optic nerve function exams, structure-function correlation, progression and change analysis, and other risk factors. The details of these categories are provided in Appendix A.5. Through these categories, we identified the following key themes that gave us insights into the factors that optometrists keep in focus while screening patients for glaucoma.

3.4.2.1 All optometrists focus more on optic nerve exams and risk factors.

The data reveals the distribution of eye examinations, as shown in Figure 3.3, within the whole dataset. Key findings include a high frequency of focus on the optic nerve assessment (87.4%) and optic nerve function assessment (80.2%), showing a significant focus on assessing the optic nerve's structural and functional aspects. Clinical findings (54.7%), medical history (53.2%), and family history (50.3%) are evaluated at moderate rates, indicating a holistic approach that incorporates patient-related and clinical information. In contrast, progression and changes analysis (18.2%) and structure-function correlation analysis (10.2%) were less frequently performed.

3.4.2.2 Experts carry out progression and change analysis more than novices.

Our findings reveal that experts are more inclined to engage in progression and change analysis than novices. A preliminary test of normality on the data distribution resulted in a $P<.001$, indicating a deviation from normal distribution. Consequently, we employed the

Mann-Whitney test to examine the differences between experts and novices in analyzing progression and changes. This test yielded a $P<.001$ (2-tailed) with U=1486, and a mean rank of 61.08 for novices and 80.54 for experts. This shows a statistically significant difference between the two cohorts. Specifically, our analysis shows that experts demonstrate a markedly higher focus on progression and change analysis (34.7% vs. 21.5%), as illustrated in Figure 3.4, reflecting their seasoned approach to monitoring disease development over time.

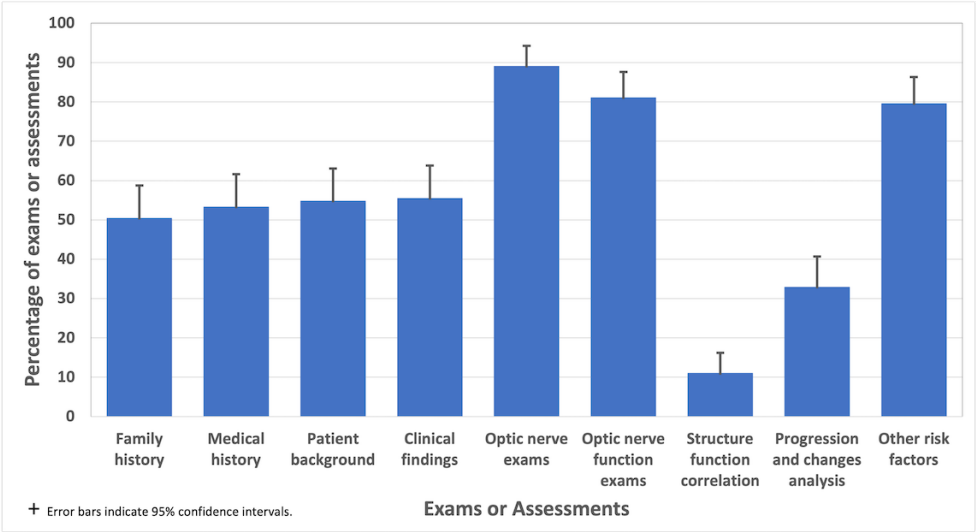


Figure 3.3: The figure shows findings from the whole dataset, revealing a strong emphasis on optic nerve structural and functional assessments. It also shows moderate attention to clinical findings, medical history, and family history, while progression analysis and structure-function correlation analysis were performed less frequently. The error bar indicates 95% confidence interval.

3.4.2.3 Experts integrate clinical findings and patient background more than novices.

There is a significant difference in the emphasis on clinical findings and patient background between novice and expert optometrists. Our initial analysis involved conducting tests of normality for both clinical findings and patient background data, which revealed $P<0.01$, indicating that the data do not follow a normal distribution. Consequently, we applied the Mann-Whitney U test to assess the differences in emphasis on these diagnostic components between novice and expert optometrists. The test results for clinical findings yielded a $P=0.039$ (2-tailed) with U=1695 and a mean rank of 63.75 for novices and 77.19 for experts. Same way the test results for patient background yielded a $P=0.045$ (2-tailed) with U=1732 and a mean rank of 64.18 for novices and 76.42 for experts, suggesting a statistically significant difference between the two groups, albeit with a modest significance level.

Specifically, novices consider these factors in approximately 50.5% and 48.2% of evaluations, respectively. In contrast, experts incorporate them at a higher rate, with about 67.3% for clinical findings and 73.9% for patient history as shown in Figure 3.4. This disparity shows the experts' more substantial focus on combining clinical signs and symptoms of glaucoma with the patient's historical data. Expert physicians' comprehensive approach showcases their in-depth understanding and extensive experience in evaluating these essential diagnostic components.

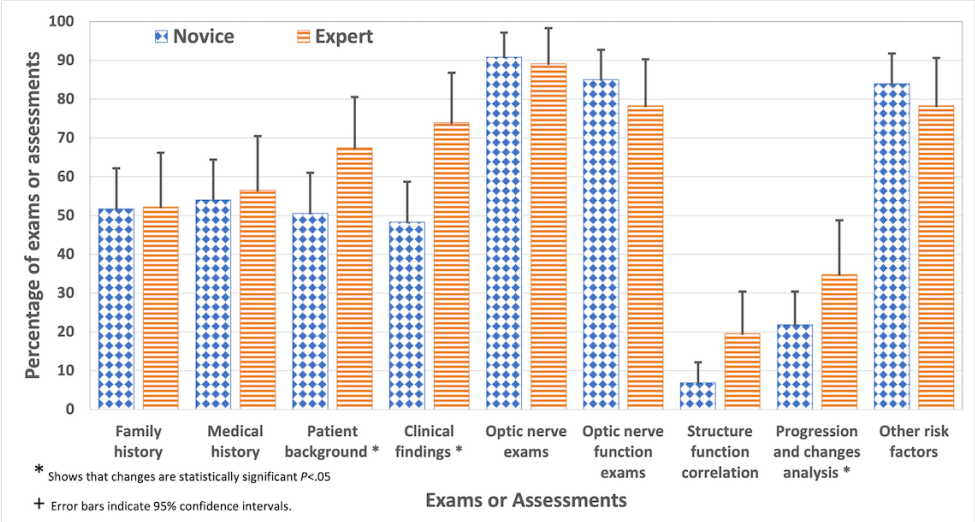


Figure 3.4: The figure shows the percentages of glaucoma-related eye examinations conducted by novices and experts, highlighting differences in focus on clinical findings, patient history, progression and change analysis, and structure-function correlation. The * with name indicate that there is significant difference with P <.05 and the error bar shows the 95% confidence interval.

3.4.2.4 Experts tend to emphasize structural function correlation more than novices.

Our analysis shows a trend that novices focus more on structure-function correlation at a rate of 6.0% and experts at a higher rate of 19.5%, as depicted in Figure 3.4. Despite the apparent disparity in percentages suggesting a more pronounced awareness of the interplay between structural and functional considerations, statistical analysis reveals that the difference in structure-function correlation between experts and novices does not achieve statistical significance (P=0.15 with U=1951.5). However, the percentage difference still indicates that there may be a trend towards a higher engagement with structure-function correlation among experts than novices.

3.4.2.5 Novices exhibit a higher tendency for inter-case and longitudinal comparisons, less for intra-case.

In such a comparison, the clinician evaluates the current case against other patient scenarios. Novices in some cases made a reference to another patient's data, denoting a reliance on external case references. This reliance may be a compensatory mechanism for the relative inexperience in glaucoma diagnosis, where pattern recognition is still being cultivated. In a representative case, a novice commented:

"I think that they would be potentially a mild suspect compared to the first patient that we had just due to the fact that this person has hypertension, and they have a family history of glaucoma..."

In stark contrast, this form of comparison was entirely absent among experts, suggesting a transition towards an intuitive, more synthesized comprehension of glaucoma presentations, informed by an extensive repository of patient encounters.

The difference between experts and novices in conducting longitudinal comparisons of patients' data is subtle, with our analysis revealing a difference that does not reach statistical significance with p¿0.05. Despite the lack of statistical significance, the observed percentages—23.9% of evaluations for experts versus 29.8% for novices—hint at an interesting trend: experts might possess a nuanced internalization of a patient's historical data. This suggests that experts are potentially more adept at retaining key details of disease progression within their working memory. With their extensive experience, they may draw upon a mental synthesis of the patient's history and current presentation without the need for frequent referrals to past records. This cognitive efficiency reflects a mastery of pattern recognition characteristic of glaucomatous progression, allowing for a more agile and comprehensive assessment that leverages their extensive experience and nuanced under-

standing of the disease. The less frequent engagement with longitudinal comparison by expert optometrists may not indicate a disregard for the patient's history but rather an advanced capacity to retain and integrate this information into their clinical judgments, a testament to their evolved diagnostic expertise.

Intracase comparisons, particularly the evaluation of asymmetry between left and right eye examinations, were utilized by novices in 17.2% of evaluations and experts in 23.9% of evaluations with no statistically significant difference. Both groups recognize the diagnostic value of these comparisons, yet the slightly higher utilization by experts may reflect a refined, systematic approach and a heightened sensitivity to the subtleties of asymmetry, honed through repeated clinical encounters.

3.4.3 Variation in Patient Follow-up Time Suggestion

The analysis reveals significant variations in follow-up times among all participating optometrists, as well as a noticeable relationship with their years of experience. The initial Chi-square statistic of 112.236 with a P=0.007 highlights a substantial difference in follow-up times among all optometrists. Further analysis, considering the years of experience, shows a Chi-square statistic of 75.359 with a P=0.029, suggesting that experience level is an influential factor in this variation. This data collectively indicates that not only do follow-up times vary among optometrists, but they also correlate significantly with their professional experience.

A further breakdown is shown in Figure 3.5. Expert optometrists, with an average follow-up time of 4.1 months for confirmed glaucoma cases and 4 for glaucoma suspects, seem to require less time to assess disease progression, likely due to their proficiency in managing the condition. In contrast, novice optometrists recommend longer follow-up

times, averaging 5.3 months for confirmed cases and 4.5 for suspects, which may indicate a need for additional time to make informed decisions about disease progression. For patients without glaucoma, both groups suggest more extended follow-up periods, with experts recommending an average of 8.15 and novices slightly more at 8.73.

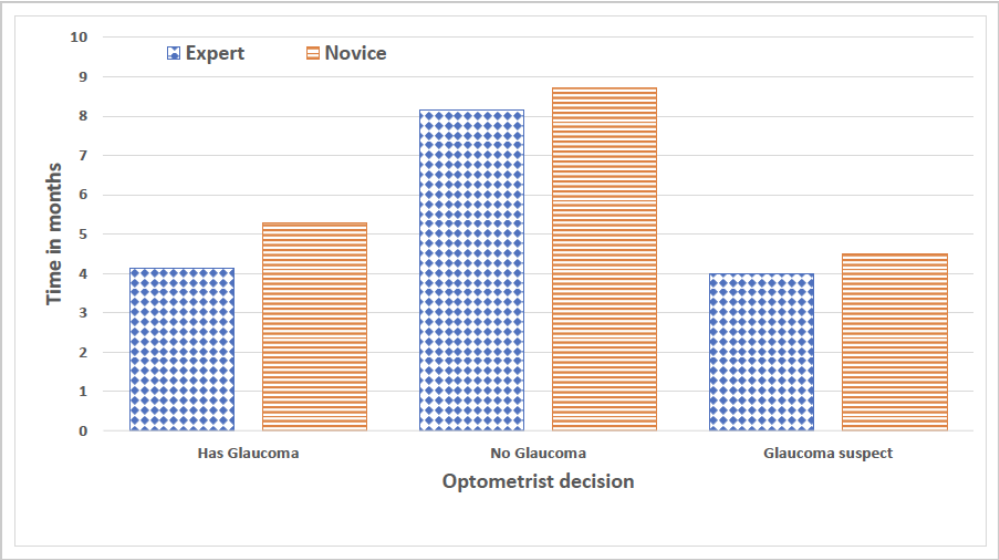


Figure 3.5: The figure shows the higher follow-up times in months suggested by expert and novice optometrists for confirmed glaucoma cases, suspects, and patients without glaucoma.

3.5 Discussion

This study investigates how optometrists' experience levels affect their decision-making in glaucoma diagnosis, a crucial factor for patient outcomes. We found significant differences in diagnostic approaches between novices and experts. The discussion ahead will dive deeper into these differences in decision-making, particularly focusing on how experience levels affect priorities in data interpretation and heuristic usage in diagnosis. This analysis

will not only correlate our findings with the central research questions but also inform the development of AI systems tailored for glaucoma diagnosis.

3.5.1 Decision-Making Processes

Our findings reveal a notable divergence in the alignment with established clinical decisions between expert and novice optometrists, particularly in the absence of prior patient data, offering crucial insights into the role of experience in clinical judgment. The higher alignment of expert diagnoses with original clinical decisions in scenarios lacking previous patient data highlights the depth of expertise honed through years of clinical practice. This expertise results in more accurate and intuitive diagnostic skills. This aligns with Hammond's Cognitive Continuum Theory, [Hammond and R, 1981], which posits that as clinicians gain experience, their decision-making process shifts from a predominantly analytical approach to a more intuitive one [McCartney, 2017, Yang et al., 2012]. Experienced practitioners are theorized to navigate and utilize this spectrum effectively, adapting their approach depending on the situation [Tsang et al., 2023]. The reduced gap in decision-making accuracy when prior patient data was provided suggests that access to comprehensive patient history significantly enhances diagnostic accuracy across experience levels highlighting the critical role of patient history in clinical decision-making. These results suggest that novices can make competent decisions when guided by detailed information.

3.5.2 Data Prioritization in Clinical Assessments

The study also explored how optometrists at different experience levels prioritize data during clinical assessments. Initially, both experts and novices engaged comprehensively across various diagnostic areas, but with a lower focus on progression, change, and correlation

analysis. This trend reflects common practice for initial assessments, where establishing a broad understanding of the patient's condition is prioritized over detailed analysis. Interestingly, subsequent visits revealed a significant increase in attention to progression and change, particularly among experts. Experts demonstrated a more comprehensive approach, integrating clinical findings and patient history from the outset and emphasizing structural-function correlation more than novices. This comprehensive risk assessment aligns with the notion that experienced clinicians can synthesize complex information more effectively [McCartney, 2017] and that is why their concordance was quite high than novices. Novices, despite a higher initial engagement progression and change analysis, might misunderstand the optimal timing for such analyses, typically requiring follow-up data. Although this increased for novices in subsequent visits, the experts showed a robust increase almost double that of novices. This indicates a strategic approach to deeper analysis in the later stages of patient management by experts.

3.5.3 Heuristic Approaches and Cognitive Efficiency

Our results suggest that experienced optometrists may employ heuristic approaches more frequently, particularly when data is limited. The nuanced internalization of a patient's historical data by experts, evidenced by their subtle differences in conducting longitudinal comparisons and intercase comparisons, suggests cognitive efficiency. This efficiency allows experts to retain and integrate key details of disease progression into their clinical judgments without frequent referrals to past records or references to other cases. This cognitive agility is indicative of their evolved diagnostic expertise and mastery of pattern recognition in glaucomatous progression [Tsang et al., 2023]. The reliance on external case references by novices, as noted in some cases, maybe a compensatory mechanism for their relative

inexperience. This behaviour highlights the gap in their approach to long-term patient management and the development of pattern recognition skills.

3.5.4 Guidelines for Developing AI System to Minimize Diagnostic Variations

In this section, we propose a set of guidelines for developing an AI system for glaucoma diagnosis based on our findings. These guidelines are intended for a human-AI collaborative environment, where the AI system will be used as a decision support tool rather than a standalone decision maker. We suggest that by following these guidelines, the AI system will effectively support optometrists with different experience levels, bridging the gap between novices and experts, and ultimately improving patient outcomes through enhanced diagnostic accuracy. The collaborative model will also uphold ethical and professional standards, mitigating any legal, ethical, and professional barriers posed by standalone system. As a result, it will facilitate effective integration of AI into clinical practice.

3.5.4.1 Minimizing the Effect of Limited Data Availability

Our findings show variability in diagnostic concordance between novice and expert optometrists when comprehensive patient histories are not available. These findings suggest that AI systems should have a robust module that provides detailed patient histories. If a patient has limited historical data, the system should present novices with two features. It should identify and display data from the most similar patient cases with the same conditions [Parimbelli et al., 2018] along with the decisions made in that case. It should also highlight the dissimilarities between the selected patient and the current patient. The AI system should be trained with expert decision-making patterns by modelling

the diagnostic approaches and visualizing the variables by estimating the missing data [Berchuck et al., 2019] backward in time to give them an idea of how history would look. These two features would minimize the limited data availability challenge and allow novices to make more accurate and consistent decisions in glaucoma diagnosis.

3.5.4.2 Evolving AI system for Initial and Subsequent Visits

Our result shows that in the initial visit, the expert optometrists adopt a comprehensive approach to assess all factors, except they don't carry analytical analysis (any progression and change analysis, and structure and function correlation analysis). As they move to subsequent visits, they still follow a comprehensive approach but with limited focus on patient historical information, instead, their focus shifts to analytical assessment. We accordingly suggest that an AI system should behave in a similar way and evolve along with the visits. In the initial visits, the AI system should prompt a novice to adopt a comprehensive approach and highlight important factors such as family history, age etc. In the subsequent visits, the AI system should highlight any analytical changes. It would be better that the AI system present a hierarchy of features, by demonstrating which features require more or less importance to assess during both the initial visit and subsequent follow-up visits. These features of AI system will help novices in understanding the optimal time for any analysis.

3.5.4.3 Follow-up Time Suggestion

The follow-up time analysis show variation among optometrists. We suggest that for better patient management the AI system should provide an optimal time frame for follow-up visits [Bian et al., 2020] which should be based on patient data and clinical indicators.

Along with follow-up interval, the system should also display the variables that influence this decision, such as disease progression, patient history, and symptoms. With this feature we can bring more consistency in follow up time suggestion by optometrists.

3.5.4.4 AI System Features and Behavior

Most of the AI systems developed by researchers for glaucoma diagnosis focus on detection of glaucoma with binary outputs [Hemelings et al., 2020, Saxena et al., 2020, Lin et al., 2022]. However, to gain optometrists trust the AI system developed for glaucoma should also include a confidence score, a detailed explanation of how it arrived at that specific decision and should provide a visualization of the risk factors in a hierarchical manner [Kamal et al., 2022a, Chayan et al., 2022]. This will ensure that the optometrist does not overlook any important details. We further suggest that in the context of glaucoma, the system should not rely on only one source of data, but that it should combine multiple sources of data as used by optometrists. The optometrist will have more trust if the system uses a similar modal data for decision-making. Additionally, the AI system should have an interactive interface that offers an adjustable variables feature [Beauxis-Aussalet et al., 2021] that allows the optometrist to modify the values of any factor and see how different scenarios would affect the outcomes, such as diagnosis, severity, and follow-up time. For example, if the optometrist changes a variable to reflect a worsening condition, the AI system will adjust the follow-up recommendation accordingly. Such an AI system design for glaucoma will ensure more consistent decision making regardless of expertise level.

3.6 Conclusion

This study investigated the diagnostic decision-making differences between novice and expert optometrists in glaucoma diagnosis, focusing on how these variations influence clinical judgment and patient outcomes. Through in-depth interviews, we identified significant disparities in diagnostic approaches. Experts demonstrated higher concordance rates with limited data, emphasizing their cognitive efficiency and intuitive decision-making, while novices relied more on structured methods and external references. Additionally, there were significant variations between experts and novices in the analysis of different exams and prioritization factors.

Based on our findings we proposed guidelines for a human-AI glaucoma diagnosis system that are aimed to minimize diagnostic variations. Incorporating the recommended features in an AI system, and utilizing a human-AI collaborative model, will not only minimize decision inconsistency, but will also improve optometrists' accuracy and enhance patient care by accelerating the diagnostic process and improving decision-making. In future work, we recommend a phased approach, starting by refining these guidelines to meet local requirements, followed by development of an AI system based on the refined guidelines, and finally simulating and validating it through real-world data and pilot testing in clinical environments.

3.7 Acknowledgment

This research benefitted from collaborations with the School of Optometry and Vision Sciences (thanks to Dr. Stanley Woo, Director) and was supported through funding from the Natural Sciences and Engineering Research Council of Canada (NSERC-CRD Grant)

(independent of study design/analysis). We are also thankful to our participants who gave us their precious time. FG's research was funded by the University of Waterloo (Graduate Research Studentship and International Doctoral Student Award). We also acknowledge the use of generative AI tools, specifically Grammarly and ChatGPT-4o, for grammar check and flow of writing. Generative AI tools are not used for any type of data generation or content generation in this paper.

3.8 Data Availability

The clinical data of patient used in the study is obtained from School of Optometry and Vision Sciences, University of Waterloo. The permission was granted to use the data for this study. We have stored all the anonymized data online in a University of Waterloo Borealis repository [Ghaffar, 2024] with restricted access. For research purposes the data access can be granted by emailing to the corresponding author, after taking the necessary steps required.

3.9 Extended Discussion

3.9.1 Participant Classification Criteria

To distinguish between novice and expert participants, we applied a threshold-based classification grounded in glaucoma-specific experience rather than general clinical exposure. Participants were categorized as novices if they had fewer than four years of clinical experience and saw fewer than three glaucoma patients per week. Those exceeding both thresholds were considered experts. These clusters are shown in Figure 3.6. Importantly,

while participants also reported their overall patient volume, this metric was excluded from the classification criteria. General patient load was deemed unrelated to glaucoma-specific experience and therefore not a valid indicator of relevant diagnostic expertise. A practitioner may see a high number of patients per week while encountering relatively few glaucoma cases, limiting the relevance of total patient volume for this study's objectives.

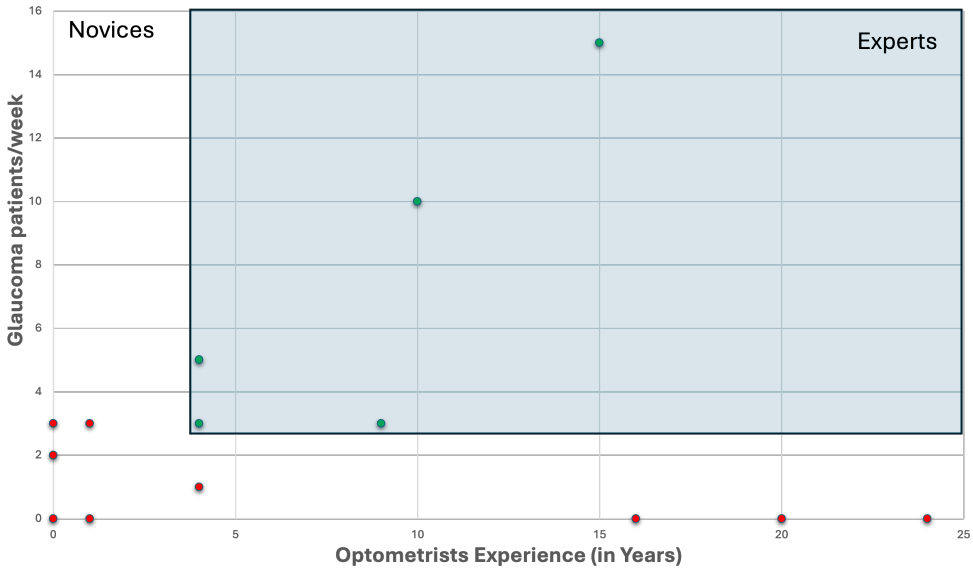


Figure 3.6: The figure shows classification of experts and novice optometrists

This classification approach was informed by multiple considerations. First, it aligns with Benner's Stages of Clinical Competence [Benner, 1982], which is based on Dreyfus and Dryfus model [Dreyfus and Dreyfus, 1980]. It describe a transition from rule-based to experience-based reasoning typically occurring by the third year of practice. Our choice of a four-year cutoff builds on this model, allowing for a conservative boundary where more integrated, context-sensitive clinical judgment is expected to emerge.

Second, this decision is consistent with prior empirical research. For instance, Dunford

et al. [Dunford et al., 2011] categorized physiotherapists as novices or experts based on years of on-call experience, using a five-year threshold. This precedent supports the use of duration-based thresholds to distinguish clinical proficiency levels in domain-specific tasks.

Third, the classification was validated through informal consultation with practicing optometrists, who affirmed that the selected thresholds accurately reflect meaningful divisions in glaucoma diagnostic expertise.

Although clinical expertise exists on a continuum, a binary classification was adopted to enable clear and robust group-level comparisons. Unsupervised clustering methods such as K-means and DBSCAN were explored using multiple experience-related variables, but the limited sample size (n = 14) and sparse data distribution led to unstable cluster assignments. In contrast, the threshold-based approach provided greater interpretability and domain validity within the context of this study.

Future work with larger samples may build on this framework to develop more granular models of clinical expertise, potentially incorporating continuous or tiered classification schemes.

3.9.2 Clarification on the Number of Cases

In this study, four real glaucoma patient cases were selected, each structured across three longitudinal visits, yielding a total of 137 diagnostic instances after the removal of incomplete data. While the number of unique patient cases was limited, the study design intentionally prioritized clinical depth, temporal continuity, and diagnostic realism. Each diagnostic instance included a complete clinical dataset—fundus photographs, OCT scans, visual field reports, and patient history, closely replicating real-world diagnostic conditions. Participants typically spent an average of ten minutes per instance, amounting to

approximately two hours of diagnostic activity per session, including concurrent think-aloud protocols.

This number of cases was chosen with careful consideration of the study's objectives and practical constraints. The aim was not to generalize across a broad range of clinical pathologies, but rather to explore diagnostic variability within and across clinicians using a within-subject, longitudinal structure. This approach enabled a focused examination of how diagnostic judgments evolve over time when evaluating the same patients at multiple visits. Expanding the number of cases would have substantially increased cognitive load and introduced the risk of participant fatigue, which could compromise both diagnostic performance and the quality of verbal data. Furthermore, all cases were derived from electronic health records and refined in collaboration with practicing optometrists to ensure clinical plausibility and diagnostic ambiguity. This allowed for rich cognitive engagement, ensuring that participants encountered realistically complex diagnostic scenarios.

Nonetheless, the limited number of unique patient cases remains a notable limitation of the study. It may constrain the generalizability of findings to broader clinical populations or to a wider spectrum of glaucoma presentations. Future research with expanded case sets and diverse diagnostic contexts would strengthen the external validity of the findings. However, for the specific aim of analyzing decision-making behavior under controlled, high-fidelity conditions, the selected structure provided a methodologically sound and ethically appropriate framework.

3.9.3 Clarification on Sample Size

This study included a total of 14 participants. While the sample size was limited, several methodological factors support its appropriateness and adequacy for the study's objectives.

Most importantly, conceptual saturation was reached during data collection. Recurring behavioral patterns were consistently observed across participants, for instance, novices tended to concentrate on direct assessment tasks such as reviewing OCT outputs in isolation, whereas experts often demonstrated a more longitudinal and comparative approach, focusing on disease progression across visits. These patterns stabilized early, with no substantially new insights emerging in the latter stages of data collection. This stabilization indicates saturation, a widely accepted threshold in qualitative and user-centered research, and suggests that additional participants would have added limited value.

In addition to conceptual saturation, the study's within-subjects design contributed significantly to data richness. Each participant reviewed up to 11 patient visits, resulting in 137 diagnostic episodes overall. By collecting multiple diagnostic instances from each participant, the design minimized the number of participants required.

The sample size is also consistent with precedent in the human factors and medical AI literature. Prior expert-novice comparison studies have used comparable sample sizes, such as the 18-participant study on anesthetist decision-making [St. Pierre and Nyce, 2020], 12 particpants in the study of differences in the cognitive processes of novice and expert public health nurses [Sarsfield, 2014]. Similarly, studies in high-cognitive-load domains, such as ecological interface design [Uereten, 2023] and clinical decision tools [Faisal et al., 2025] have drawn meaningful insights with samples ranging from 13 to 19 participants. Moreover, [Hennink and Kaiser, 2022] influential review suggests that 9–17 participants are typically sufficient for reaching thematic saturation in qualitative studies. In the context of usability testing, the [Faulkner, 2003] and FDA's Human Factors Guidance notes that a minimum of 15 users is adequate to detect the vast majority of usability issues.

Nonetheless, a greater sample size can always lead to findings of more themes in a qualitative study. Moreover, as all participants were recruited from a single geographic and

educational context, and their experiences may not reflect variations present in broader clinical settings. Future studies may benefit from recruiting more diverse samples across institutions and regions to extend the applicability of the findings.

In summary, the sample size used in this study was methodologically justified and empirically productive. It allowed for both rich behavioral analysis and meaningful statistical comparisons between expert and novice clinicians, while adhering to established norms in cognitive and clinical research.

Chapter 4

Different Thoughts to the Same Decision: A Qualitative and Quantitative Investigation of Experience-Based Clinical Reasoning

This chapter is based on the submitted article:

Ghaffar F, Furtado N, Burns C. Different Thoughts to the Same Decision: A Qualitative and Quantitative Investigation of Experience-Based Clinical Reasoning. *Health Policy and Technology.*

4.1 Abstract

Background

Diagnostic decision-making in glaucoma varies significantly between clinicians, often influenced by experience levels. This variability can result in inconsistent diagnoses and delayed treatment, leading to irreversible vision loss. Understanding these differences is essential for improving diagnostic accuracy and designing AI systems that support human-AI collaboration in clinical workflows.

Objective

This study examines differences in diagnostic approaches, the factors influencing decisions, and how novices and experts balance positive (in favor of glaucoma) and negative (against glaucoma) evidence. It also explores variations in glaucoma staging and the potential for AI to address these gaps and enhance clinical decision-making.

Method

Using a within-subject design, optometrists with a diverse range of experience were presented with real patient case data for glaucoma diagnosis. Interviews explored their diagnostic reasoning, and responses were systematically coded and analyzed using both qualitative and quantitative methods. This approach allowed for a detailed comparison of decision-making patterns and the identification of areas where AI could enhance clinical decision-making and support.

Results

Experts integrated a broader range of factors, focused more on progression analysis, and dynamically balanced positive and negative decision factors according to disease severity in staging. They also demonstrated efficient filtration processes, discarding extraneous information and prioritizing critical data points. Novices, on the other hand, relied on fewer explicit indicators, overemphasized these factors, and struggled to balance positive and negative factors in staging, showing less alignment with disease severity.

Conclusion

This study highlights significant differences in decision-making factors between novice and expert optometrists, emphasizing the need for designing tailored AI systems to address these differences and bridge the expertise gap by supporting novice learning, enhancing diagnostic consistency, and complementing expert decision-making.

4.2 Introduction

Diagnostic decision-making in healthcare is as much art as science. Even seasoned professionals, with their varied experiences and knowledge, can arrive at different conclusions when interpreting the same clinical evidence [Croskerry et al., 2023]. This variability is particularly pronounced in conditions like glaucoma, a thief of sight that often progresses undetected until irreversible damage has occurred [Topouzis and Anastasopoulos, 2007]. Glaucoma's subtle early signs, often reliant on subjective clinical assessments, can be easily missed, and leads to different decisions [Azuara-Blanco et al., 2007]. This diagnostic challenge carries significant clinical consequences, as delayed or missed diagnoses can lead

to irreversible vision loss, diminished quality of life, and increased healthcare burden for patients [Doozandeh et al., 2022].

This study addresses a crucial gap in the existing literature by addressing and going deeper into the nuanced clinical reasoning employed by optometrists when diagnosing glaucoma. Specifically, we aim to dissect how expert and novice practitioners weigh various factors, both in favor of (positive) and against (negative) a glaucoma diagnosis, and how these considerations differ between experience levels. While previous research has examined some aspects of diagnostic variability in glaucoma [Bettin and Matteo, 2013, Reus et al., 2010] a key gap remains: no studies have directly compared the decision-making processes of novice and expert optometrists, particularly the underlying reasoning and the factors influencing their diagnostic conclusions.

AI holds the promise of standardizing clinical decision-making, offering a more objective analysis of complex data, and ultimately enhancing diagnostic accuracy. Recent studies have shown AI's remarkable ability to analyze vast amounts of patient data, including retinal images, visual field tests, OCT scans, and medical histories, often outperforming human experts in detecting early signs of glaucoma [Akter et al., 2022, Chaurasia et al., 2022, Mursch-Edlmayr et al., 2020]. Moreover, the integration of AI into clinical practice has the potential to streamline workflows, alleviate the burden on healthcare professionals, and improve patient outcomes [Andersen et al., 2023].

However, to harness the full potential of AI in optometry and bridge the gap between human and machine intelligence, we must first gain a comprehensive understanding of the factors that influence the decision-making processes of both novice and experienced optometrists [Cook, 1993, Lai et al., 2023]. Understanding these decision-making patterns is not only crucial for improving diagnostic accuracy and consistency but also for designing Human-AI systems [Ghaffar et al., 2024b, Hyodo et al., 2023]. This can further inform

educational strategies that can better equip future optometrists with the skills needed to navigate the complexities of glaucoma assessment [Brunyé et al., 2019]. By identifying the specific factors that influence diagnostic decisions, we can pinpoint areas where novices and experts diverge in their assessment, leading to targeted interventions for novice optometrists, educational strategies, and AI systems to improve clinical judgment and bridge the expertise gap for optimal patient care.

Specifically in this research study, we seek to answer the following questions:

1. What are the differences in factors that influence the decisions of expert and novice optometrists in relation to different decisions?
2. How do novices and experts balance positive (in favor of glaucoma) and negative factors (against glaucoma) in staging their decision?
3. What role can AI play in bridging the expertise gap?

To investigate these research questions, we carried out interviews with both novice and expert optometrists, concentrating on their methods for diagnosing glaucoma. After conducting these interviews, we coded the responses, and carried out both qualitative and quantitative analyses. This systematic approach enabled us to thoroughly explore the decision-making patterns, investigate the reasons behind their decisions, and determine the necessary steps for effectively integrating AI into optometric practice.

In conclusion, this research aims to unravel the reasons which result in different decision by optometrists when diagnosing glaucoma. By scrutinizing the factors that influence these decisions, particularly the interplay of positive and negative evidence, we will shed light on the inherent variability in diagnostic practices between experts and novices. This knowledge will pave the way for targeted educational interventions to enhance clinical judgment,

improve patient care and also lay the groundwork for the development of AI tools that can effectively complement and augment the expertise of eye care professionals. Ultimately, our goal is to foster a new era of precision and consistency in glaucoma diagnosis, empowering clinicians with the insights and tools needed to safeguard the precious gift of sight for their patients.

4.3 Methodology

This study follows the same methodology as our earlier research (Faisal Ghaffar et al., 2024).

4.3.1 Data Collection

The data for this study was obtained from the School of Optometry and Vision Sciences at the University of Waterloo, with ethics clearance from the Research Ethics Committee on May 27, 2021 (clearance number 43288) as given in Appendix A.1. The data attributes collected in our earlier study, informed by research from [Leske et al., 2008, Le et al., 2003, de Voogd et al., 2006, Gordon et al., 2002, Group et al., 2007]

Four patient records were selected for in-depth interviews, chosen for their potential to provide diverse diagnostic perspectives. Each optometrist reviewed three patients with three visits each and one patient with two visits, balancing detailed case analysis with practical time constraints. This approach enabled a comprehensive examination of clinical decision-making variations.

We used the same data collected from interviews with 14 optometrists in our earlier study. Those optometrists were all from the Waterloo region, primarily affiliated with

the University of Waterloo's School of Optometry and Vision Science. Each optometrist, averaging 7.54 years of experience (SD = 8.34 years), saw an average of 32.57 patients weekly, participated in a 100–120-minute online interview using the 'think aloud' protocol to review patient records.

4.3.2 Data Coding

The same cleaned data used in our earlier study was utilized. Post-cleaning, the refined transcripts were imported into NVIVO (Version 14.23.2). This time the focus of coding was on answering why participants thought a patient had glaucoma, did not have glaucoma, or was suspected of having glaucoma. Additionally, the coding focused on the balancing of in favor and against reasoning at specific stage of the disease.

The open coding process was led by FG, with multidisciplinary input enriching the analysis. CB, with an engineering background, reviewed the codes for technical and analytical accuracy, while NF, an optometry expert, ensured alignment with optometric principles and practices. This collaborative approach ensured that the coding accurately reflected diverse professional viewpoints. The inductive coding process allowed the data to drive categorization, leading to an organic understanding. Each transcript line was examined, and open codes were assigned to succinctly summarize content while remaining true to the original text. These codes were then grouped into more generalized codes, as detailed in Table —-.

4.3.3 Data Analysis

To carry out the analysis, we first categorized participants into two groups: novices and experts, the same way we divided it in earlier study. Participants collectively reviewed 137

scenarios, with experts evaluating 48 cases and novices examining 89.

We compared 88 visit reviews by novices to 45 visits review by experts. From the initial dataset, four patient records were selected for in-depth interviews based on their potential to provide diverse diagnostic perspectives. Each optometrist reviewed three patients with three visits each and one patient with two visits, balancing detailed case analysis with practical time constraints.

Firstly, we examined the difference in approaches by in-depth analysis of justifications and reasons provided by novices and experts in taking different decisions (having glaucoma, not having glaucoma and glaucoma suspect). We employed quantitative and qualitative methodologies to explore the reasons and justifications behind each diagnostic and staging decision. Through this we identified key themes that distinguish the approaches of novice and expert optometrists.

Second, we highlighted the stark differences in clinical methodologies, the confidence with which both groups take a particular decision and investigating the rationale behind these decisions, our study aimed to identify the key factors driving clinical judgments.

Lastly, we analyzed variations in staging decisions, which offers a window into the depth and breadth of assessments made by novices and experts. We specifically looked into how experts and novice balance their decisions by providing justification both in favor of and against glaucoma. Such insights are valuable for refining clinical assessment protocols and ensuring consistency in the staging of the disease.

Overall, our study offers a comprehensive analysis to explore the factors influencing novices and experts in their diagnostic and staging decisions. These findings are crucial for developing future educational programs, clinical guidelines, and AI systems in glaucoma. By addressing the specific needs and challenges faced by each group of optometrists, we

can ensure these tools are effectively tailored to improve clinical practice.

4.4 Results

4.4.1 Variations in Decision-Making: Approaches and Confidence of Novices Versus Experts and the Reasons Behind Their Decisions

Novices and experts exhibit varied decision-making approaches and confidence levels, with reasons for their choices. We specifically examine the critical factors these professionals consider when making decisions about classifying a patient as having a diagnosis of glaucoma, glaucoma suspect, or not having glaucoma, and explaining their confidence with making these clinical decisions. Additionally, the analysis examines the justifications provided for their decisions, shedding light on the rationale underpinning their diagnostic choices.

Figure 4.1 illustrates the percentage distribution of various assessments conducted by each group, categorized as "having glaucoma", "glaucoma suspect", and "not having glaucoma". The data shows there are distinct patterns in diagnostic assessments. Firstly, direct glaucoma exams assessments (optic nerve exams and optic nerve function exams), alongside the evaluation of risk factors, are fundamental at all stages of diagnosis. Secondly, external factors (family history, medical history, and patient background) predominantly contribute to raising suspicions or ruling out glaucoma, rather than confirming its presence. Lastly, analytical tools (clinical findings, progression and change analysis, and correlation analysis) are more pivotal in affirmatively confirming glaucoma, as opposed to its exclusion or mere suspicion.

Moreover, novices and experts displayed distinct patterns in justifying their decisions. Novices, on average provided 3.02 justifications per decision, whereas experts averaged 2.24 reasons to support their decision. These justifications were subsequently categorized into two types: negative reasons, indicative of the absence of glaucoma, and positive reasons, suggestive of the presence of the disease.

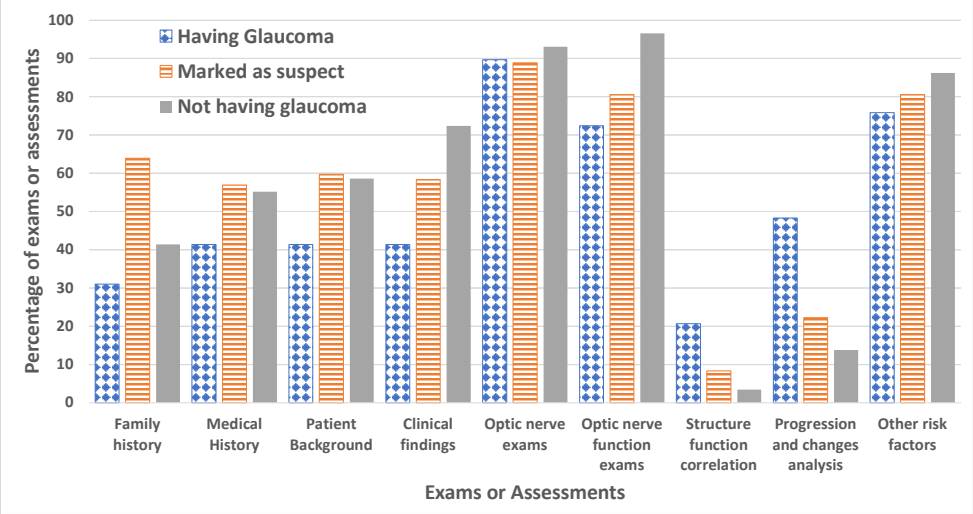


Figure 4.1: Comparative analysis of glaucoma-related eye examinations across different diagnostic decisions.

The novices, on average, provided 1.06 negative reasons and 1.95 positive reasons. In contrast, experts furnished an average of 0.77 negative reasons and 1.46 positive reasons per case. This analysis reveals that novices in glaucoma diagnosis tend to use a more exploratory and comprehensive approach, likely due to less confidence and experience, leading them to provide a greater number of reasons. In contrast, experts demonstrate a streamlined, focused approach, offering fewer reasons overall, which reflects their higher confidence and experience in identifying key indicators of the disease. Another reason for difference is that experts employ a better and efficient information filtration process than

novices. After initial review experts filter out unwanted information for decision making while novices are less efficient in doing that. This contrast highlights the significant impact of experience on diagnostic strategies in medical practice.

Additionally, a comparative analysis of decision confidence between expert and novice optometrists shows a pronounced difference. Employing a Likert scale ranging from 1 to 5 to measure confidence in clinical decision-making, it was found that experts exhibited a high degree of confidence (scores of 4 and 5) in 64% of evaluations, without any instances of the lowest confidence echelon (score of 1). In contrast, novices demonstrated a similar degree of confidence in only 50% of cases and allocated the minimal confidence score (1) in 5% of evaluations. This distinction in confidence levels was substantiated by a Chi-square statistic of 99.241 and an exceedingly low p-value of 0.000172, firmly rejecting the null hypothesis of no difference in confidence levels. This behaviour is also presented in another study [Inoki et al., 2020] where they explored the diagnostic performances of novice and expert endoscopists and their confidence. Subsequent analysis of confidence score revealed a dichotomy in confidence distribution, with experts displaying a propensity for higher confidence in confirmed cases of glaucoma, whereas novices were more assured in their evaluations of suspected glaucoma cases and those without the condition.

In the sub-sections, our investigation extends to examining variations in the approach to their decisions and the specific reasons provided by each group post-case review to understand the rationale behind their diagnostic choices.

4.4.1.1 Diagnostic Approach and Decision Reasons Divergence in Glaucoma Cases

In the diagnosed cases of glaucoma, the approaches and the reasons provided by novice and expert optometrists show distinct patterns. The main themes we extracted are discussed below.

Experts carry out comprehensive analysis vs novices focus on specific indicators. We analyzed the exams and related patient data reviewed by the participants. It was observed that before classifying a patient as having glaucoma, experts look into most of the available data following a comprehensive approach while novices don't follow this trend. We conducted a Welsh test [Welch, 1947] (the data was uniformly distributed with $P<0.001$) to verify this variation and found that the difference is statistically significant with a p-value of 0.025. Table 4.1, which represents both novice and expert assessments, further provides details on the approach difference. Experts consider a broad spectrum of factors, including direct glaucoma exams assessment, external factors, analytical analyses, and risk factors. while novices tend to focus more narrowly on direct glaucoma exams assessment. Though novices do consider a variety of factors, their emphasis is less extensive compared to experts. This difference highlights that novices generally depend more on direct clinical indicators for diagnosing glaucoma, while experts engage in a more comprehensive evaluation to confirm their diagnosis.

Novices focus on external factors while experts prioritize disease progression. When asked for reasons to support their decision, optometrist provided both positive and negative reasons. Figure 4.2 showcases the positive reasons provided by novices and experts when justifying a diagnosis of glaucoma. Novices give significant attention to optic nerve structure and function, and they frequently incorporate external factors and other

Table 4.1: Percentages of decisive factors used by optometrists in their evaluation

Assessment Categories	Glaucoma		Suspect		No Glaucoma	
	Novices (N=29)	Experts (N=9)	Novices (N=51)	Experts (N=21)	Novices (N=16)	Experts (N=13)
Family History	25	44.44	66.67	57.14	37.5	46.15
Medical History	35	55.56	58.82	52.38	56.25	53.85
Patient Background	30	66.67	56.86	66.67	50	69.23
Clinical Findings	30	66.67	50.98	76.19	62.5	84.62
Optic Nerve Exams	90	88.89	90.2	85.71	87.5	100
Optic Nerve Function Exams	75	66.67	82.35	76.19	100	92.31
Structure-Function Corr.	15	33.33	5.88	14.29	0	7.69
Progression & Change Analysis	35	77.78	17.65	33.33	18.75	7.69
Other Risk Factors	75	77.78	86.27	66.67	81.25	92.31

risk factors into their assessments, albeit with scant focus on the progression of the disease or the correlation between structure and function. Experts, however, prioritize optic nerve structure and function and place a strong emphasis on progression and change analysis. Remarkably, their references to external factors are half as frequent as those of novices, and risk factors are not mentioned at all. It is interesting to note that the attention given to correlation analyses remains consistent across both groups.

This pattern implies that while novices may rely on a wider range of diagnostic signals, experts focus on the evolution of the disease, indicating a more targeted and perhaps experience-based approach in their diagnostic process.

This is also evident from a representative case, where a novice stated:

"There is documented thinning of the rim tissue of the RNFL layer. There is the presence of Glaucoma damage that can be quantified... "

While experts stated that:

"Given the fact that there are changes on both OCT and visual field even though maybe they're still in the early stages because of the changes in the visual field and OCT changes are the main two things..."

This integrative method is further validated by the higher percentages indicating a broader scope of diagnostic evaluation. In another case, the preference for external factors

by novices was quantitatively substantiated, with 50.00% of novices highlighting these elements, as one stated:

"I would make the diagnosis at this point so it would be primary open-angle glaucoma because there are no secondary glaucoma causes on gonioscopy or slit lamp and it would be normal tension glaucoma because the pressure is not exceeding 21, 22 IOP. I'm being a bit more conservative on this case taking into account the family history."

In contrast, experts gave precedence to the disease's progression, with 44.44% accounting for progression and change analysis. An expert's comment affirmed this focus:

"Based on the progression of RNFL, I think this patient has Glaucoma because Glaucoma is defined as a progressive disease and based on his RNFL trend report there is progressive thinning of the RNFL layer in both of his eyes and it's significant enough that some of them are in the double digits."

The analysis revealed a marked contrast in the decision-making processes of novice and expert optometrists.

Experts optimize information filtering for decision-making. Experts began their diagnostic evaluation by considering a comprehensive array of variables shown in Table 4.1, indicative of a broad and integrative understanding of glaucoma. Despite this wide initial consideration, experts demonstrated a propensity to distill their focus to a select group of factors deemed most crucial for the diagnosis (e.g. Removing the risk factors) shown in Figure 4.2. This selective approach suggests that through their experience, experts have developed the ability to prioritize specific indicators that are most telling for each case. In contrast, novices engaged with a more limited set of variables from the outset. The data suggest that this narrower scope is reflective of their developmental stage in clinical expertise, which may not prompt an extensive consideration of potential diagnostic factors.

Unlike experts, novices appeared to utilize nearly all the information at their disposal in the decision-making process, treating each variable with equal importance. The absence of prioritization in the novices' approach may stem from a relative uncertainty about which indicators are most significant in confirming the presence of glaucoma.

Figure 4.2. Comparison of justifications provided in favor of Glaucoma Diagnosis: Expert and Novice Optometrists

This optimized filtration process is further supported by the differences in justifications between novice and expert optometrists are notable. On average, novices provide 0.5 reasons against a glaucoma diagnosis and 2.35 reasons in favor, while experts offer 0.27 reasons against and 2.56 reasons in favor. Experts provide significantly fewer negative justifications compared to novices and more reasons supporting a glaucoma diagnosis. This disparity suggests that experts are more confident in their decisions than novices.

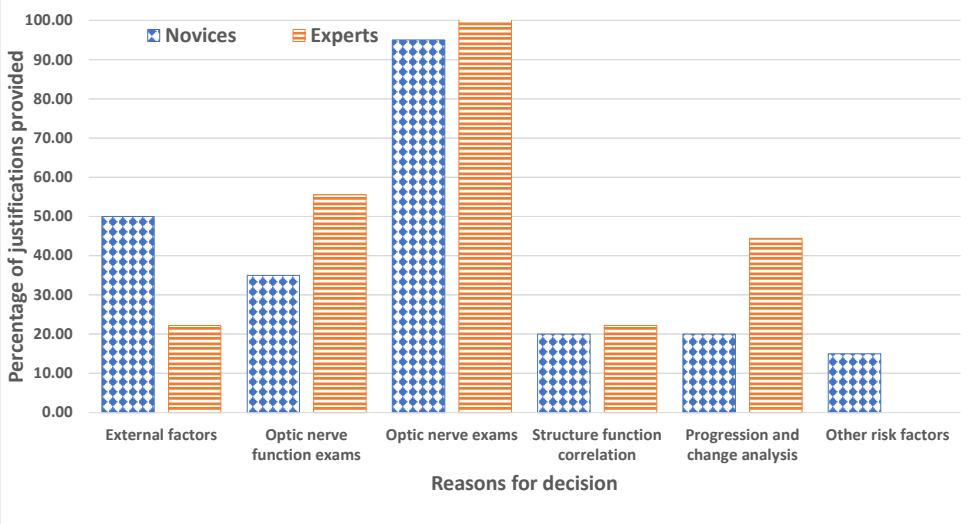


Figure 4.2: Comparison of justifications provided in favor of Glaucoma Diagnosis: Expert and Novice Optometrists

4.4.1.2 Diagnostic Approaches and Decision-Reasons Divergence in Suspected Cases

In the evaluation of suspected glaucoma cases, the diagnostic strategies and justifications offered by novice and expert optometrists exhibit marked differences. The sections below examine these divergent approaches, elucidating the primary themes identified from our analysis.

Novices' adaptability to a holistic approach and experts' steady commitment to holistic evaluations. In cases of suspected glaucoma, as indicated in Table 4.1, we observed novices shift towards a more holistic diagnostic approach, diverging from their previously narrow focus evident in confirmed cases of the disease. Interestingly, this adaptation appears to converge toward the evaluation strategy employed by experts, who consistently apply a comprehensive approach to both confirmed and suspected cases. Our analysis underscores this alignment, revealing no statistically significant variation (p¿0.05) in the methodologies adopted by novices and experts in the context of suspected glaucoma. This suggests a harmonization of diagnostic strategies under uncertainty, with novices expanding their assessment criteria to encompass a broader range of factors, presumably to navigate the ambiguity inherent in these cases. Also, novices might tend to adopt a cautious strategy, integrating various factors due to the uncertain nature of these cases.

Expert consistency in carrying disease progression and change analysis. Despite the strategic comprehensive approach alignment, a clear distinction (with $p < 0.01$ on the Mann-Whitney test) remains in the disease progression and change among the two groups. Novices, although broadening their evaluative scope, still engage in analytical analysis to a lesser extent (17.65%) compared to experts (33.33%) as shown in Table 4.1. Experts, on the other hand, continue to focus on the dynamics of disease progression,

tracking any changes and assessing the structure-function correlation. This suggests expert rely on the disease progression and interrelationship of various factors to inform and support their diagnostic suspicions.

Experts optimize information filtering for decision-making. Experts and novices differentiate their approach through the quantity and nature of reasons provided for or against a glaucoma diagnosis. Experts averaged 2.14 in favour and 0.82 against the diagnosis while novices tend to list more reasons, with averages of 3.13 for and 1.05 against, indicating a less selective information processing approach. Further analysis, supported by statistically significant findings from the Mann-Whitney test ($p < 0.001$), confirms that both novices and experts engage in an information-filtering process, albeit with varying degrees of effectiveness. The calculated effect sizes r=6.633 for experts and r=10.999 for novices reveal that experts possess a more sophisticated information filtration system. This suggests experts' reliance on key diagnostic indicators, honed by experience, whereas novices consider a broader array of signs, reflecting their developing diagnostic skills.

Divergent criteria for excluding glaucoma. In evaluating negative reasons in suspected cases, Figure 4.3. shows that novices prioritized the absence, inaccuracy, or unreliability of data (13.73%) over experts (8.70%), while experts focused more on the lack of disease progression (17.39%) and the absence of certain risk factors (19%) and structure-function correlation (4.35%). Conversely, novices emphasized direct glaucoma assessments like optic nerve exams more than experts. This is also evident from one representative case when a novice was marking it as a suspect and justified their decision with poor reliability of the available data:

"...because there are abnormal findings on the OCT and on the visual field as well. but I think both of the tests have poor reliability so it makes me not want to say for sure that she has glaucoma...".

For the same case, experts marked it as suspect and justified it with the patient not having progression:

"...looking at the OCT and all the thinning around the RNFL. I wouldn't say that they have glaucoma yet because I feel like they need some progression to say that they have glaucoma..."

This suggests experts tend to emphasize the absence of progression along with optic nerve assessments, while novices' decision is affected by the unreliability of data and focus more on direct optic nerve assessments. This reflects a divergence in ruling out definite glaucoma in suspected cases between the experience levels.

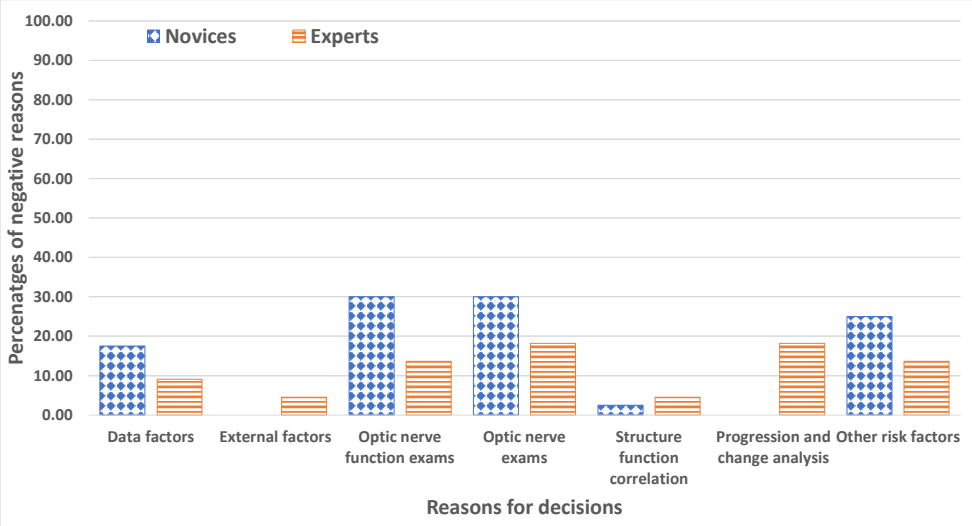


Figure 4.3: Negative indicators against glaucoma confirmation in suspected cases: Expert vs. Novice optometrists'

Divergence in determinative reasons. In analyzing the positive reasons in favour of glaucoma, Figure 4.4. illustrates that direct glaucoma exam assessments are central to the decision-making process for both novice and expert clinicians, with a slightly higher reliance

noted among novices. The figure also points out a substantial discrepancy in the consideration of external factors and other risk factors, with novices attributing significantly higher percentages to these elements (112.05% and 27.50%, respectively) as opposed to experts (36.36% and 9.09%, respectively). Additionally, while structure-function correlation and progression and change are factored into the assessments by both groups, experts place a marginally higher emphasis on structure-function correlation than novices, yet these factors are less predominant than in confirmed cases of glaucoma. This quantitative analysis is further supported by comments from a case study.

A novice clinician remarked on the utility of visual fields and the relevance of family history, stating:

"I think the visual fields are pretty reliable and there do seem to be some defects, especially in the left eye and then given the strong family history, that would be a little bit of concern and also the appearance of the optic nerve on that left side would make me a little bit concerned."

In contrast, an expert clinician's comment revealed a more reserved and data-focused approach:

"so yeah, even though there are all these changes on OCT but there doesn't seem to be much change to the visual field even though it's been nine months since her last exam. there still doesn't seem to be any significant visual field defect so I'm still concerned about their OCT but I'm not necessarily confident that they would say they have glaucoma yet."

The expert's consideration of the lack of visual field changes despite OCT abnormalities demonstrates a reliance on correlating structural and functional change.

The integration of quantitative analysis with these qualitative comments suggest that while novices tend to consider a wider array of factors including external ones such as family

history, experts prioritize the correlation between clinical assessments and diagnostic tests. This distinction showcases the impact of experience on the emphasis and interpretation of diagnostic indicators in glaucoma suspicion.

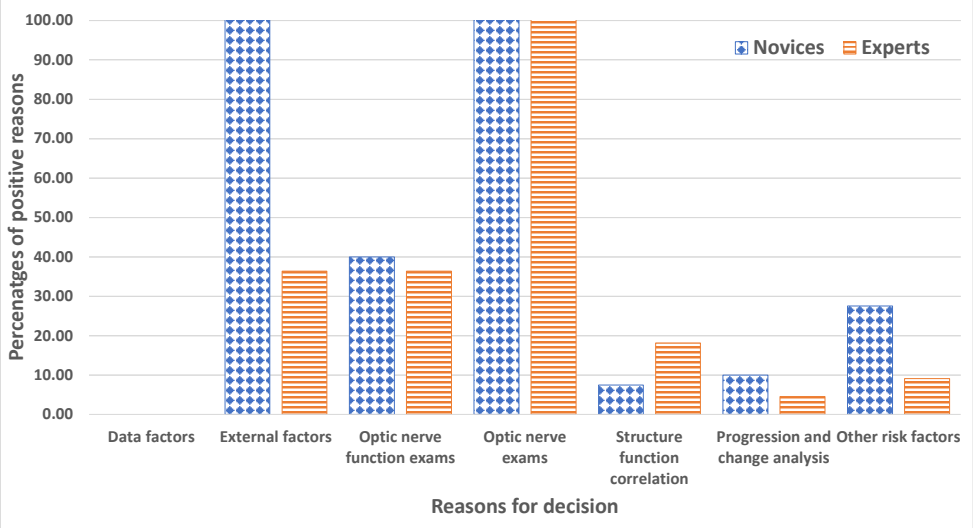


Figure 4.4: Positive indicators in favour of glaucoma confirmation in suspected cases: Expert vs. Novice optometrists'

4.4.1.3 Diagnostic Approaches and Decision-Reasons Divergence in Glaucoma Exclusion

In the evaluation of cases where patients were diagnosed as not having glaucoma, the diagnostic approaches and the reasons provided to justify glaucoma exclusion had visible differences. The themes below extracted from our analysis explain the differences.

Experts' comprehensive evaluations vs novices' focused assessments. In scenarios where glaucoma was not diagnosed, the data delineates distinct patterns in the frequency of assessments carried out by novice and expert practitioners shown in Table 4.1.

Experts demonstrated thoroughness in their evaluations by reviewing most of the available data. Such comprehensive analysis suggests that experts possess a more extensive consideration of the potential indicators that may negate a glaucoma diagnosis. Novices, while showing a marked focus on optic nerve function and structure exams, referenced other assessments less frequently compared to their expert counterparts. Both novices and experts registered low percentages in the assessment of structure-function correlation and progression analysis, hinting at a possible shared perspective regarding the limited relevance of these evaluations in non-glaucoma diagnoses.

Divergence in decision reasons for glaucoma exclusion. When prompted to explain the reasons for their decisions, both groups cite the optic nerve exam, optic nerve function exam and absence of risk factors as key considerations. However, there was a modest inclination among experts to consider data factors (8.33%) and the absence of structure-function correlation (16.33% each) as shown in Figure 4.5. Though these figures are small, they reflect a nuanced approach adopted by experts in excluding a glaucoma diagnosis.

In a representative case, an expert backed their decision solely on no correlation between vision and anatomy. Although they did identify several other risk factors and even found out in the patient medication record that they were on glaucoma medications but still classified them as not having glaucoma. The expert commented that:

"Because. Well, I think this last visual field isn't very good. so does not have glaucoma because the visual field doesn't match well. No. What am I trying to say? Does not have glaucoma because it's more like. It's a loss in the tissue. I think loss in the tissue. Well, I'm looking at the macular asymmetry, and it seems like there's quite a bit of change and loss in that inferior temporal area of the nerve fiber layer. But that's probably more because of the drance heme and not because of a sort of glaucoma does cause I don't

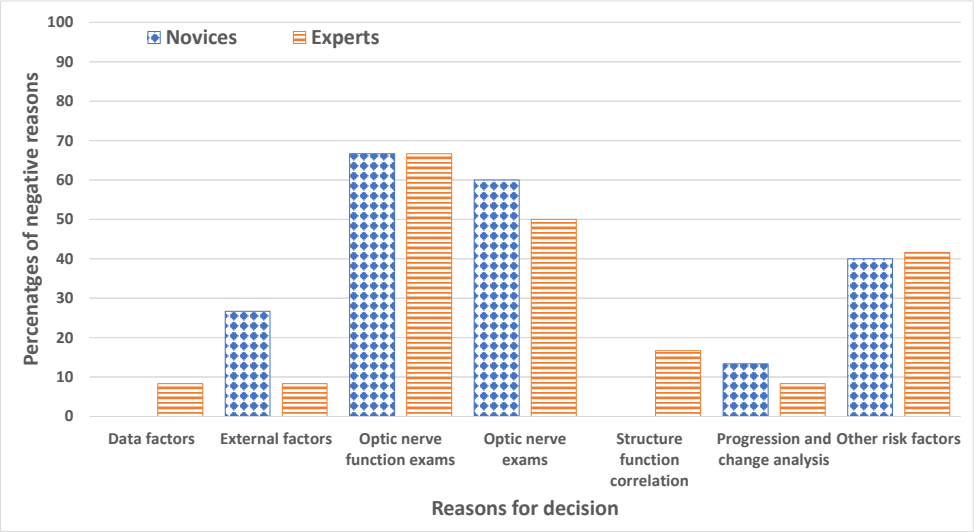


Figure 4.5: Negative indicators against glaucoma in no glaucoma cases: Expert vs. Novice optometrists'

know if that makes sense. but she is taking Latanoprost. I think I just saw that. hmm. But three times. three times meaning like drance hemorrhages three times. Yeah. I don't know this is a bit puzzling. I didn't read the part where she was already taking latanoprost and Azopt [Brinzolamide], so if I looks at it here, I would say. Hmm so. She has a characteristic inferior temporal nerve fibre layer loss, which kind of correlates to the inferior. the macular asymmetry as well, thinning that we're seeing now. That would typically Indicate glaucoma, but her visual fields don't necessarily show a superior nasal loss in the field, so there's no correlation between vision and anatomy. She does have a hemorrhage, but that may not be progressive due to glaucoma. No, still no."

In contrast, when novices evaluated the same case, they classified it as a glaucoma suspect, reflecting a more conservative approach. A novice clinician pointed to the recurrence of drance hemorrhages and mild OCT progression, alongside the patient's hypertension, as factors supporting a higher glaucoma suspect classification. The participant commented

that:

"There are still like recurrences of the drance hemorrhages. There's the hypertension is still not significant, although they are being treated, but I don't think it would be sufficient to cause the hemorrhages. And there are drance hemorrhages. And there's been some mild progression with the OCT so yeah, I'm, I'm putting him in the higher glaucoma suspect range."

This suggests distinct diagnostic methodologies between novice and expert clinicians in ruling out glaucoma. Experts apply a more detailed and interpretive approach, emphasizing the congruence between visual function and anatomical evidence. Novices, on the other hand, tend to lean towards suspecting glaucoma in the presence of certain risk factors and clinical indicators, even though there is no structural and functional correlation. These differences highlight the varying levels of clinical experience and interpretative complexity between the two groups.

4.4.2 Variation in Glaucoma Staging Decision

The participants were asked to stage their decision in different stages. The suspected decision stages are low, medium and high suspicion while the glaucoma decision stages are early, mild, moderate and advanced stages.

The distribution in Figure 4.6 and Figure 4.7 shows that in both glaucoma-diagnosed cases and suspected cases novices' distribution follows a bell-shaped curve drawn with second-order polynomial regression. This is typical for less experience. In both cases, we see that experts' responses are more tightly clustered than novices. Moreover, in the glaucoma stages, we see a higher dispersion of decisions of novices which shows a higher inter-group variation of novices than experts.

We further analyzed the reasons provided by optometrists in each stage of glaucoma. Our findings highlight that experts generally offer fewer negative reasons against diagnosing glaucoma (averaging 0.34) compared to novices (0.70), suggesting a more decisive approach in their evaluations. Interestingly, experts also give slightly fewer positive reasons (1.11) than novices (1.13), indicating a more refined and possibly cautious diagnostic process.

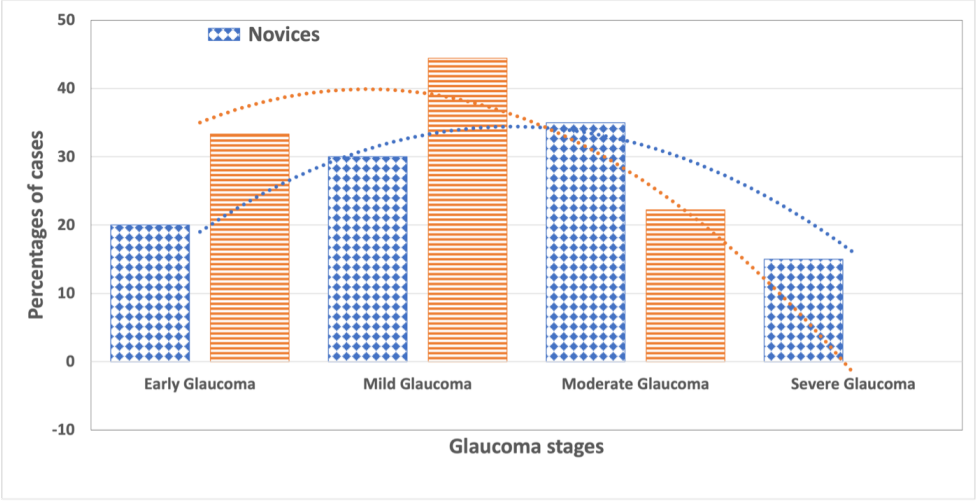


Figure 4.6: Distribution of confirmed glaucoma cases stages by experts and novices.

Figure 4.8 shows in-depth details of similar and distinct trends in both groups. There is a consistent trend where both novice and expert optometrists are more inclined to provide positive justifications, especially in the later stages of glaucoma, reflecting a common understanding of the disease's progression. Experts, however, demonstrate a more nuanced approach by maintaining a balance between positive and negative reasons across different stages, with a marked preference for positive reasons in the moderate stage of glaucoma. This contrasts with novices, who often interpret high suspect stages as definitive of glaucoma, relying predominantly on positive reasons. Utilizing second-order polynomial regression to draw trend lines, we unraveled the intricate relationship between glaucoma stages,

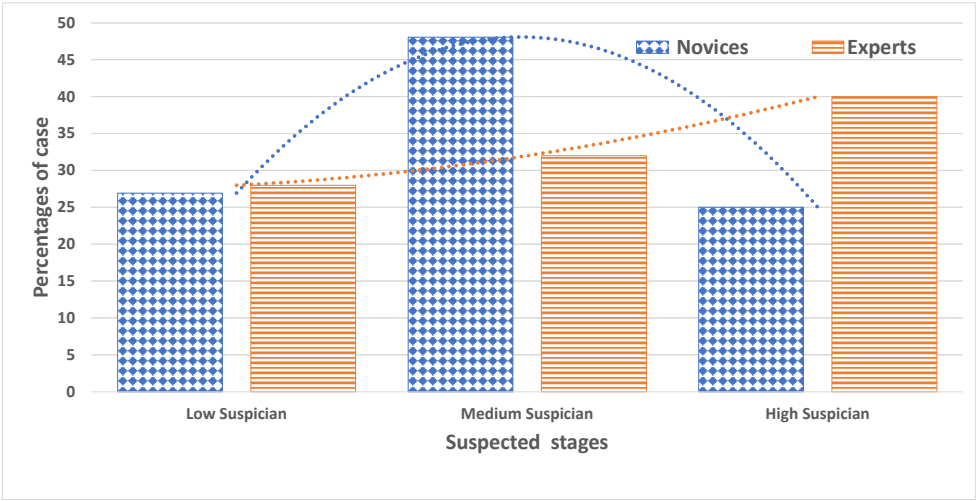


Figure 4.7: Distribution of suspected cases stages by experts and novices.

optometrist expertise, and diagnostic reasoning. This method reveals that experts exhibit a decrease in negative reasons from low to moderate stages, signaling growing confidence in their diagnoses, while novices show a plateau in suspected stages. The positive trend lines for experts shows a significant increase in supportive reasons, emphasizing the complexity and depth of expert reasoning in glaucoma assessment. This comprehensive analysis sheds light on the evolving nature of diagnostic reasoning as it relates to optometrist experience and the progression of glaucoma, offering valuable insights into the decision-making process in clinical settings.

4.4.3 Limitation and Future Work

Our study has certain limitations that need to be addressed in future research. Firstly, the sample size of 14 participants is limited. A larger sample size would provide more robust data and increase the generalizability of the study findings. Secondly, the staging of the

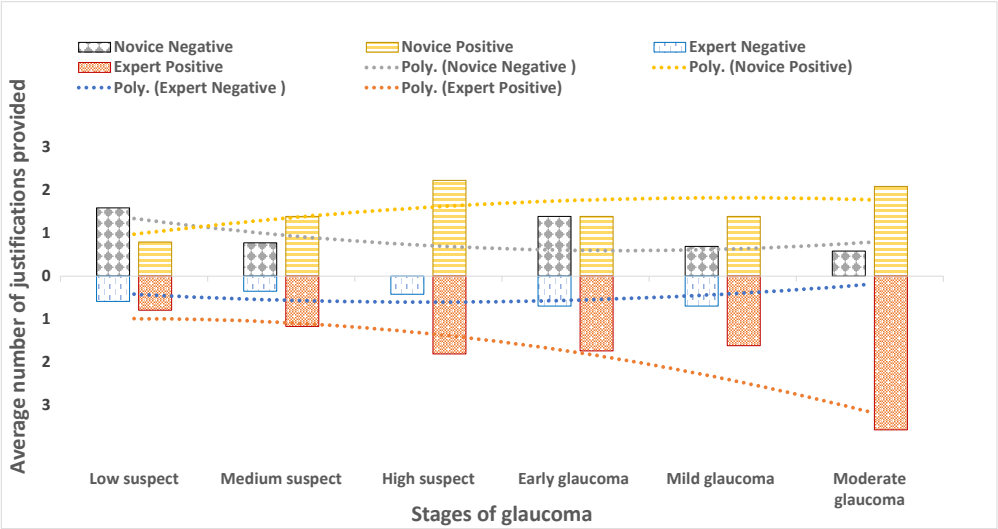


Figure 4.8: Comparison and trends of positive and negative reasons in staging: Experts vs Novices

disease needs to be investigated in more detail to identify the factors that influence the staging decision. Thirdly, our study does not focus on the trust and user acceptance of the AI system. Understanding how optometrists perceive and accept the system is crucial for its successful implementation and utilization in clinical practice.

Future research should prioritize expanding the sample size to offer more comprehensive insights and enhance the validity of the findings, capturing a broader range of user experiences and system performance metrics. Additionally, a detailed investigation into the staging of glaucoma is necessary to refine the system's diagnostic capabilities by identifying the key factors influencing staging decisions. Furthermore, evaluating the trust and acceptance of the AI system among optometrists will help in designing strategies to improve user engagement and adoption.

Chapter 5

From Diagnostic Variation to AI System Design

5.1 Introduction

The study in chapters 3 and 4 examined how optometrists (novice and expert) approach glaucoma diagnosis, revealing critical differences in diagnostic reasoning. These studies highlighted key patterns of variability in decision-making strategies. This variability led to identification of several factors that guide the human-centered AI design for optometric practices.

This chapter explains the process of the mockup AI systems and provides a connection with the study discussed in chapter 6 and 7. Additionally, it outlines the design of the studies in these upcoming chapters, including participant groups, conditions, and evaluation metrics. This chapter serves to demonstrate the conceptual and methodological continuity that links the two studies as part of a unified research design. Some of the information in

this chapter will also be repeated in the next two chapters.

5.2 What Kind of AI System Do Optometrists Want?

This section outlines key design recommendations for AI systems in optometry, based on insights from the variation study. It is divided into two subsections: the first presents suggestions derived directly from observed decision-making variations, while the second captures additional insights gathered through open-ended discussions at the end of the interviews. Together, these recommendations aim to guide the development of human-centered AI tools that support the needs of both novice and expert optometrists.

5.2.1 AI System Design Suggestions Derived from the Observed Decision-Making Variation

5.2.1.1 Addressing Data Availability Limitations

The findings in the previous two chapters highlight the need for AI systems to include a module that presents detailed patient histories. This would help overcome limitations due to missing or unavailable data, ensuring optometrists, especially novices, can make more informed decisions.

5.2.1.2 Adaptive Support Based on Visit Type

During initial visits, the AI system should prompt optometrists to adopt a comprehensive approach by emphasizing key factors such as family history, age, and ethnicity. For follow-

up visits, the AI should shift focus to highlight analytical changes, helping clinicians track disease progression or stability efficiently.

5.2.1.3 Standardizing Follow-Up Time Frames

Our analysis revealed variation in recommended follow-up times among optometrists. To support more consistent patient management, the AI system should suggest an optimal time frame for follow-up visits based on clinical findings and best practice guidelines.

5.2.1.4 Prioritizing Relevant Clinical Information

To mimic expert behavior, the AI should be designed to filter and prioritize critical diagnostic factors. This includes reducing cognitive load by discarding irrelevant data and directing attention to the most essential clinical indicators.

5.2.1.5 Optic Nerve Evaluation and Risk Factor Integration

The AI should support diagnostic practices by analyzing optic nerve structure and function. This includes highlighting key features in data and integrating relevant risk factors such as intraocular pressure, family history, and systemic conditions to present a concise, clinically meaningful summary.

5.2.1.6 Supporting Progression Analysis

AI should enable longitudinal analysis by comparing historical and current data, highlighting disease progression. It should detect significant changes in imaging and functional data

over time, reducing manual comparison efforts and enhancing early detection of deterioration.

5.2.1.7 Output Interpretability Beyond Binary Decisions

The AI system should also provide a confidence score, disease severity along with a transparent explanation of how the decision was made, including which features contributed most to the outcome.

5.2.2 Additional Key Insights from Participant Discussion to Inform AI Design

To guide the development of an AI system that meaningfully supports glaucoma diagnosis, a series of exploratory questions were posed to both novice and expert optometrists. These questions aimed to explore how clinicians prioritize diagnostic data, interpret system outputs, engage with data presentation formats, and evaluate the role of AI in their workflow. Rather than focusing on diagnostic outcomes, the intent was to derive insights into clinician expectations and cognitive patterns. The findings add to the conceptual basis for the AI system design.

5.2.2.1 Data Prioritization and Modality Preferences

Participants were first asked what types of clinical data should be used in glaucoma diagnosis. Both experts and novices consistently identified OCT and visual field data as core modalities. However, there were notable differences in how each group prioritized additional information. Experts frequently emphasized the importance of context, with

three out of five specifically highlighting the role of medical history in guiding diagnostic interpretation. Fundus imagery was mentioned by only two experts, and CD ratio was referenced by just one. similarly, 7/9 novices mentioned fundus photographs, visual field, and OCT but fewer considered medical history to be relevant. This suggests a more limited integration of contextual or historical patient information by novices into their diagnostic process.

5.2.2.2 Optometrists Expectations for AI Output and Decision Support

When asked about the type of output an AI system should provide, both groups agreed that diagnostic conclusions are important, but their preferences diverged in terms of format and depth. Experts expressed a desire not only for a definitive diagnosis but also for additional information such as disease staging, progression trends, risk factor assessment, and potential treatment options. This reflects a holistic approach to diagnosis that aligns with their broader clinical experience and decision-making needs.

Novices, on the other hand, displayed a preference for more guided and simplified outputs. Some preferred a probabilistic likelihood of glaucoma (e.g., a percentage score), while others wanted binary answers or viewed the system as a screening aid rather than a decision-maker. Several novices also expressed interest in knowing the stage or speed of disease progression, which suggests a desire to develop a better understanding of the condition itself.

5.2.2.3 Preferences for Data Visualization and Presentation

In addition to content, the way information is presented also emerged as a significant factor. All experts agreed that visualizing clinical trends over time, such as RNFL progression

or changes in visual field performance, was essential for meaningful interpretation. This reflects a diagnostic strategy that emphasizes pattern recognition and temporal reasoning.

Novices, however, stated a strong interest in comparing patient data to normative values, finding this type of reference point useful for contextualizing individual results. While some appreciated visualizations, others preferred text-based summaries or minimal visual detail. One participant explicitly noted a preference for plain-text comparisons over graphical outputs.

5.2.2.4 Explainability and Decision Transparency

Another critical area of interest was that where AI systems should explain their decision-making processes. All participants, both experts and novices, agreed that some form of explanation is essential. Among experts, two out of three stated that it was important not just to receive a diagnosis, but to understand how the AI reached that conclusion.

Novices similarly valued transparency, though their preferences reflected different cognitive needs. One preferred a simple Boolean explanation (e.g., yes/no rationale), while others explicitly wanted a clear, step-by-step breakdown of how the decision was made. These responses suggest that explainability serves different functions depending on the user: it supports learning and confidence-building for novices, while enabling verification and workflow alignment for experts.

5.2.2.5 Trust in Unimodal AI Systems

Finally, participants were asked whether they would use an AI system that relied solely on fundus photographs. The overwhelming consensus across both groups was that this approach is insufficient. Among experts, three out of five rejected the idea outright, noting

that fundus data alone does not provide enough information to make a reliable diagnosis. One expert indicated conditional support, using it only as a supplementary tool and two stated they would only use it if the system were exceptionally accurate, a claim they regarded with skepticism.

Novices were similarly critical. Eight out of nine stated they would not use a fundus-only system, citing lack of information and diagnostic uncertainty. A few acknowledged that such a tool might serve as a supplementary aid, but not as a standalone diagnostic solution. One novice said they would accept it only if it were highly accurate, which they believed was unlikely.

5.3 Mockup AI System Design

Based on the insights from our initial study, we developed mockup AI systems to further investigate design considerations. However during our analysis of first study, we found participants expressed strong reservations about AI systems that rely solely on fundus images. This feedback revealed a crucial gap in our original approach: even if we developed a system based only on fundus imaging, we would fail to explore how the type of input data, also known as input modality, influences users' trust and acceptance of AI systems.

This is a significant concern, as the kind of data an AI system uses can greatly impact how reliable or trustworthy it appears to healthcare professionals. For example, an AI that integrates multiple types of diagnostic information may be perceived as more robust and accurate than one that uses a single source. Without considering this, we risk overlooking a key factor that shapes user trust and acceptance.

To address this, we designed a second study focusing on how different input modalities

affect optometrists' trust in and acceptance of AI systems. Instead of creating just one AI prototype, we developed two types of mockup systems: a unimodal system and a multimodal system. While both versions share the same user interface, designed to support both novices and experts, they differ in the types of data they process.

The unimodal system uses only a single data type: fundus images.

The multimodal system integrates multiple data sources, including fundus images, optical coherence tomography (OCT), visual field test results, patient background information, and medical history.

This section outlines the design of both the unimodal and multimodal AI mockup systems. In the following section, we explain how these designs reflect and respond to the findings from our first study.

The general interface is universal and can accommodate both unimodal and multimodal AI system as shown in Figure 5.1.

5.3.1 AI System Interface for Unimodal Data

The user interface of the unimodal glaucoma screening system was designed to provide clinicians with a clear, interpretable, and actionable assessment based solely on fundus images as shown in Figure 5.2. The layout is divided into two main panels: a summary panel on the left and a detailed diagnostic view on the right, ensuring that both a high-level overview and in-depth analysis are readily accessible.

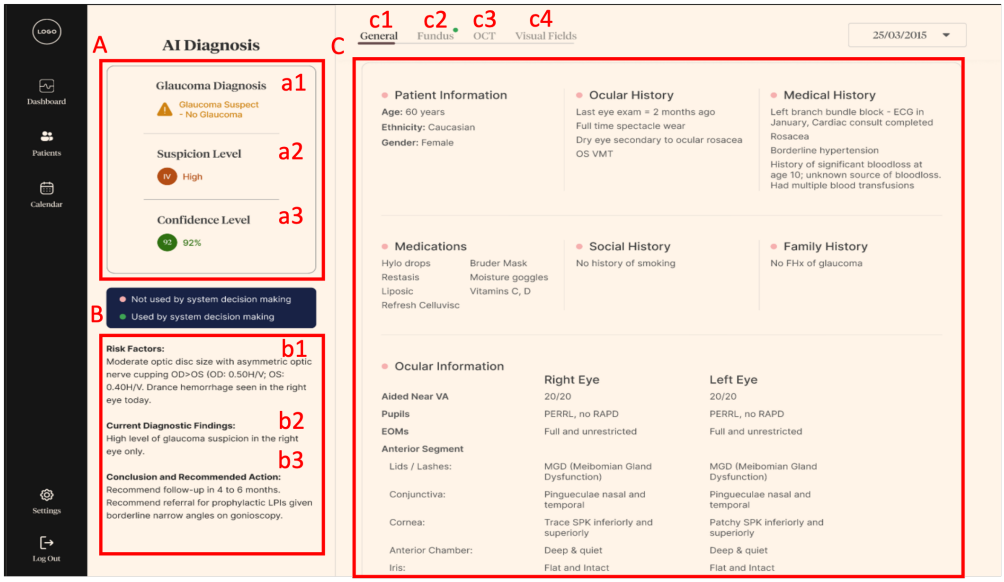


Figure 5.1: AI mock-up system interfaces (Left-hand side shows the AI global output and the right-hand side includes all the patient information)

5.3.1.1 AI Diagnostic Summary Panel

On the left hand side, the system provides an immediate summary of the AI generated diagnosis. At the top, the AI Diagnosis heading introduces the primary result. Below this, the Glaucoma Diagnosis section displays a clear label indicating the diagnostic outcome "glaucoma suspect, no glaucoma, or has glaucoma" In our study, this diagnosis was not generated by the AI system itself but rather taken directly from the clinical diagnosis provided by expert ophthalmologists.

Beneath the diagnosis, the system displays two key metrics: the Suspicion Level and the Confidence Level. The Suspicion Level is shown as "high, medium or low" which reflects the AI model's interpretation of the degree of abnormality in the optic nerve head structure. Since the clinical system did not originally provide a suspicion score, we derived this value

from our initial participant study. Specifically, it was calculated based on a majority vote among expert annotations; if most clinicians flagged the case as highly suspicious, the suspicion level was labeled as "High."

The Confidence Level is displayed as a percentage and represents how certain the system is in its classification. In our study, this was approximated by calculating the percentage of participants whose diagnosis matched the clinical ground truth.

The left panel also includes a legend that uses color coded indicators to explain which features were used or ignored by the algorithm. Green icons indicate features that were factored into the AI's analysis, whereas blue indicates those not used. This enhances interpretability and encourages clinician trust by revealing the model's decision-making components.

Further down, a brief section titled Risk Factors summarizes the specific anatomical features extracted from the fundus images. These include the optic disc size and the vertical and horizontal cup-to-disc ratios for both eyes. These values, along with the observation of moderate optic nerve cupping, are well established indicators in glaucoma screening and form the core basis of the system's suspicion assessment. The concluding section, titled Conclusion and Recommended Action, recommends next action for patient, and advises a follow-up time. Both of these content areas, Risk Factors and Conclusion and Recommended Action were derived from the findings of our first study cohort. The terminology and presentation were further reviewed and standardized in collaboration with practicing optometrists, ensuring that the output aligns with professional language and decision-making conventions used in clinical optometry

5.3.1.2 Information Specific to Fundus Imaging

The right panel of the interface provides a detailed breakdown of the system's assessment, focusing on the analysis of structural features extracted from fundus images when fundus tab is selected. Under the section titled System Diagnosis, the system evaluates each eye individually, OD (Right Eye) and OS (Left Eye), with a specific emphasis on the optic nerve head morphology. For each eye, the system presents detailed observations, such as the extent of optic nerve cupping, the presence or absence of hemorrhages, and the cup-to-disc ratio measurements. This label is displayed in green text if there is no abnormality otherwise shown in red for clarity and visual emphasis within the interface. The content presented in this section, both in terms of clinical descriptors and the diagnostic interpretation, was derived from our initial study dataset. The mockup AI-generated summaries were designed to simulate realistic system outputs and were reviewed and refined in collaboration with professional optometrists.

Importantly, a disclaimer is provided at the bottom, noting that the results are intended for screening purposes only. It emphasizes the need for clinical follow-up and acknowledges that the system's accuracy is subject to ongoing validation and improvement.

Finally, the interface includes a visual representation of the fundus images used in the analysis. These high resolution retinal photographs, one for each eye, are presented side by side under the heading "Fundus Image." Each image can be expanded for closer inspection. These images provide the visual evidence behind the AI's interpretation, allowing clinicians to validate the assessment manually if needed.

The system also includes a date selector in the top right corner, allowing users to view historical assessments and track changes in optic nerve health over time. Navigation tabs at the top labeled "General," "Fundus," "OCT," and "Visual Fields" provide these

information but not involved in the AI diagnostic process.

5.3.2 The Multimodal AI Mockup System Interface

The multimodal AI system integrates diagnostic information from multiple clinical sources. This version of the interface brings together outputs from fundus photography, OCT, visual field testing, and patient historical data. By combining structural and functional inputs, the system aims to replicate a clinically realistic screening workflow.

The right part of the multimodal AI interface focuses on detailed findings from either Fundus, OCT or visual field, based on the tab clicked. The fundus is already explained in Section 5.3.1.2.

5.3.2.1 OCT Tab

The OCT tab provides separate mockup AI diagnostic assessments for the right eye (OD) and left eye (OS), specifically analyzing the optic nerve rim, macula and progression analysis. This data is taken from our earlier study and then standardized by optometrists for familiar clinical language to make the output easy for clinicians to interpret. The system also includes a status label, such as "Uncertain – first visit," in cases where there is no prior data for comparison. An overall assessment is shown beneath the individual eye findings, summarizing the AI's interpretation across both eyes. This summary helps streamline the decision making process by giving clinicians a quick reference point from OCT.

Below the AI diagnostic summary of the OCT, the interface presents OCT data, including Retinal nerve fiber layer analysis and macular asymmetry analysis. These visual tools help clinicians confirm the AI system's findings and detect subtle changes in nerve fiber thickness and symmetry.

5.3.2.2 Visual Field

The Visual Fields tab displays AI generated assessments for the right eye and left eye, reporting the findings, progression analysis, and status for each eye individually. These elements help identify functional vision loss patterns and evaluate whether any significant changes have occurred over time.

Following the individual eye assessments, the interface provides an overall visual field assessment, summarizing whether there are any definitive signs of glaucomatous progression across both eyes.

Below the AI diagnostic summary, the interface presents detailed visual field data, including Single field analysis for both OD and OS, as well as glaucoma progression analysis baseline results and any available glaucoma progression analysis. These detailed metrics allow optometrists to confirm the AI system's findings and support clinical decision making with objective functional data.

5.3.2.3 Patient Background and History Tab

The General tab provides a summary of key background information relevant to glaucoma assessment. It includes the patient's medical history, current medications, social history, and family history, with a particular focus on any known familial risk of glaucoma. In addition to systemic details, the tab also presents findings from the ocular examination for each eye. This covers visual acuity, pupil responses, extraocular motility, and a detailed anterior segment evaluation, including observations related to the lids and lashes, conjunctiva, cornea, anterior chamber, iris, and lens. Key measurements such as intraocular pressure, assessed through Goldman tonometry, and central corneal thickness are also provided.

5.3.3 Summary Panel

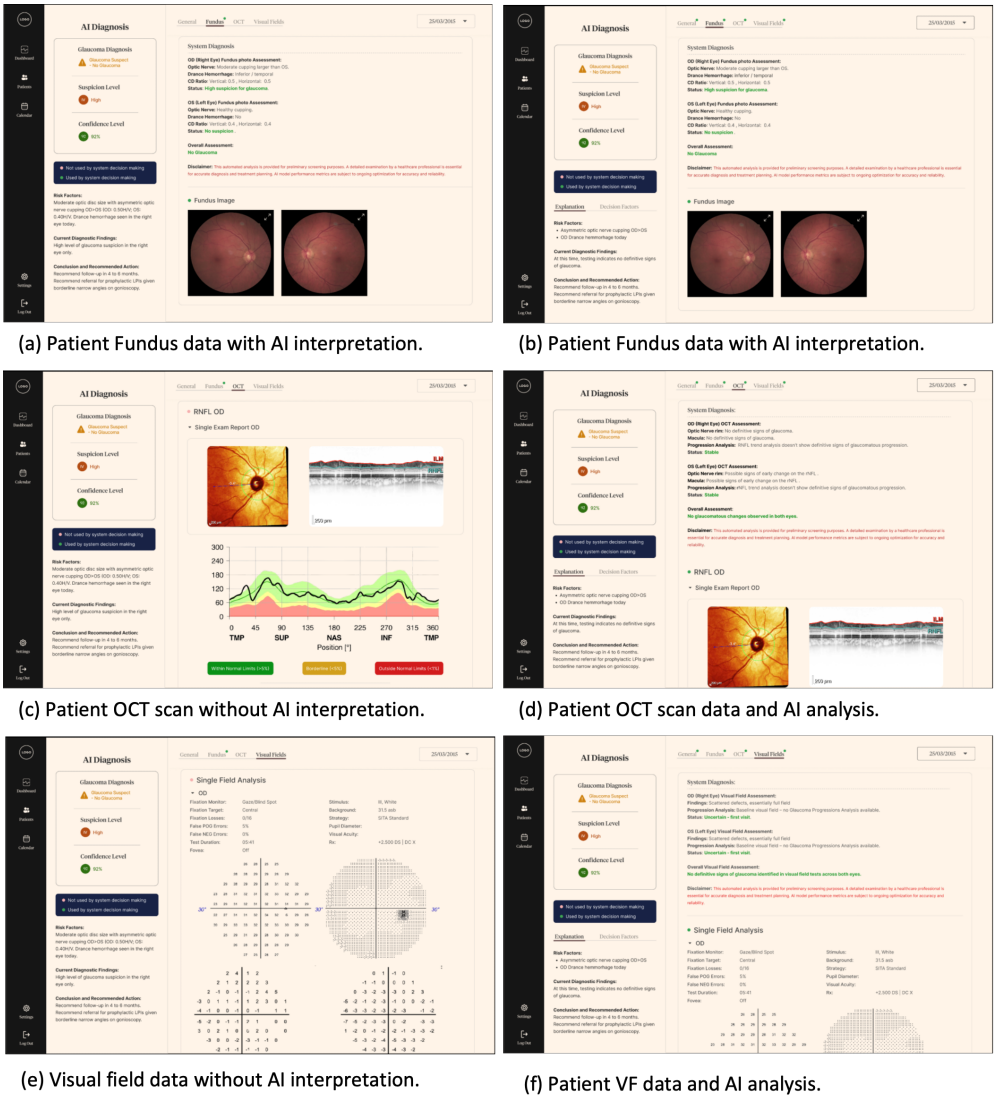
The left hand summary panel in the multimodal system retains the same structure and components as described in section 5.3.1.1. This includes the AI diagnosis, suspicion level, confidence level, and the feature use legend. In the multimodal one these decisions are taken based on all the available data. The risk factors includes factors from fundus, OCT, visual field and patient background and history as opposed to unimodal where the risk factors were only from fundus imaging. As previously explained, these elements were derived from clinical diagnoses and expert annotations from our initial study, and their presentation was standardized in collaboration with practicing optometrists.

5.4 Alignment of AI Systems with Optometrist Needs

The development of the AI interface was shaped by the exploratory insights gathered from both novice and expert optometrists. Rather than aiming to replicate any single diagnostic workflow, the interface was designed to reflect the diverse priorities, expectations, and reasoning styles observed across participants. The goal was to create a system structure that could flexibly support a range of clinical approaches while maintaining alignment with real-world diagnostic practice.

5.4.1 Adaptive Support According to Visits

To address adaptive support based on visit type, the OCT AI system module is designed to recognize the context of each visit. During initial visits, the system does not provide progression analysis, since prior data is unavailable but explicitly notifies the user that progression insights cannot be generated due to it being the first encounter. This avoids



(a) Patient Fundus data with AI interpretation.

(b) Patient Fundus data with AI interpretation.

(c) Patient OCT scan without AI interpretation.

(d) Patient OCT scan data and AI analysis.

(e) Visual field data without AI interpretation.

(f) Patient VF data and AI analysis.

Figure 5.2: Unimodal and multimodal AI mock-up system, side-by-side comparison of all sections. The left-hand side shows a unimodal system and the right-hand side shows a multimodal system

confusion and sets accurate expectations. In subsequent visits, the same module automatically presents progression analysis, highlighting relevant changes over time in OCT data to support clinical decision making regarding disease stability or advancement.

5.4.2 Follow-Up Time Suggestion

To support the standardization of follow-up time recommendations, the conclusion and recommendation includes personalized follow-up time suggestions based on the patient's condition, as illustrated in Figure 3.1.

5.4.3 Prioritization of Risk Factors and Relevant Clinical Information

For prioritizing relevant information, in the AI summary section in the left hand side of the interface the system present only high critical factors that drive the decision instead of listing all the factors as shown in Figure 3.1. This gives the optometrist an idea of what factors are more important to look into. This behavior also mimics expert optometrists, the AI is designed to filter and prioritize critical diagnostic factors. This reduces cognitive load by discarding irrelevant data and directing attention to the most essential clinical indicators.

5.4.3.1 Providing Additional Information Beyond Binary Output

To address the shared emphasis on key diagnostic modalities, the system integrates OCT and visual field data, identified by nearly all participants as core to glaucoma assessment. Additionally, contextual information such as medical history, medications, and anterior segment findings is included to reflect expert users' preference for incorporating longitudinal and systemic factors into diagnostic reasoning and so that they have access to all data. Although fewer novices prioritized historical data, its inclusion allows for a more

comprehensive and clinically realistic foundation, particularly for those aiming to build confidence or deepen their diagnostic understanding.

5.4.4 AI System Output Format

In terms of output formatting, the interface offers a diagnostic label accompanied by a suspicion level, confidence score, and brief interpretive summary. This layered presentation aims to support a spectrum of user needs: from novices who prefer simplified, guided outputs, to experts who seek more nuanced assessments such as trend analysis or progression status. Importantly, the system avoids binary conclusions and instead presents results in a way that encourages interpretation within clinical context.

5.4.5 Data Presentation

The presentation of information was also shaped by participants' preferences for visualization. The system includes both text summaries from the AI system as well as the actual visual data (e.g., OCT cross-sections, RNFL progression graphs, GPA plots) to allow users to engage with data in the format they find most intuitive.

5.4.6 Decision Transparency

Recognizing the importance of decision transparency, the interface includes a feature-use legend that highlights which inputs were utilized by the AI during its diagnostic process. This is intended to support trust and interpretability across user levels, providing decision-traceability.

While this AI system was informed by optometrist feedback, further investigation is needed to evaluate how well these different AI system types, particularly unimodal versus multimodal, affect their trust and acceptance.

5.5 Technical Feasibility of the AI Mock-ups

The AI systems presented in this study were implemented as mock-ups to simulate realistic decision-support behavior in both unimodal and multimodal conditions. Unlike purely conceptual designs, these prototypes were grounded in established and technically feasible approaches that align with the current state of AI in ophthalmology and clinical decision support.

Specifically, the multimodal system integrated data streams and analyses that are already well-supported by existing AI architectures. Fundus images can be analyzed with Convolutional Neural Networks (CNNs) or Vision Transformers (e.g., ResNet, EfficientNet) to extract clinically relevant features such as optic cup and disc segmentation, cup-to-disc ratio , neuroretinal rim thinning (ISNT rule), retinal nerve fiber layer defects, peripapillary atrophy , vessel displacement and density changes, and optic disc morphology [Coan et al., 2023, Kumar et al., 2023]. Optical coherence tomography data can be processed using 2D CNNs for individual slices or 3D CNNs and Transformer-based recurrent architectures (e.g., GRU-enhanced models) [Ashtari-Majlan and Masip, 2025] for volumetric analysis of RNFL thickness maps, ganglion cell complex profiles, lamina cribrosa depth, macular layers, and progression patterns over time. Visual field data can either be handled in tabular form or transformed into grayscale images for CNN-based analysis of mean deviation, pattern standard deviation , visual field index , hemifield asymmetry, point-wise sensitivity loss, and probability maps [Akter et al., 2025]. Finally, structured clinical data

such as intraocular pressure, central corneal thickness, age, refractive error, family history, and glaucoma staging are readily modeled through dense feedforward networks after appropriate standardization. While the explanation and interpretation proposed can be achived with SHAP and LIME explainable AI models [Kamal et al., 2022b] or with current advancement of LLM's [Raja et al., 2025]. .

By leveraging these established AI models and studies, the mock-ups in this study represent plausible and technically achievable designs rather than speculative or futuristic concepts. They were purposefully assembled to reflect current best practices in AI for glaucoma care while facilitating controlled evaluation of user trust and acceptance.

5.6 Study Design: Explaining the Participants, Conditions, and Measures

To evaluate trust and acceptance of AI-assisted diagnostic tools in optometric practice, a user study was conducted. The primary objective was to assess the impact of AI modality, unimodal versus multimodal, on clinician trust and acceptance. The study also explored whether user specific factors (such as gender and clinical experience) and system-specific factors (such as the presence of errors) influenced perceptions of the system's reliability, clarity, and adoption potential.

5.6.1 Participants

A total of 20 licensed optometrists participated in the study. All participants were actively engaged in clinical care and had experience in glaucoma screening and diagnosis. Recruitment was conducted through professional networks, in Waterloo Region. Informed consent

was obtained from all participants, who were also compensated for their time. Details regarding ethics approval are provided in Appendix B.1.

Although the primary focus of the study was to examine the effect of interface modality on trust and acceptance, additional demographic information, including participants' gender, years of clinical experience, and general acceptance of AI system, was also collected. These variables were included to explore whether individual differences influenced how participants perceived and interacted with the AI system. Specifically, the study aimed to determine whether certain user characteristics were associated with higher or lower levels of trust and willingness to adopt the system in clinical practice.

5.6.2 Study Conditions

Each participant reviewed two simulated patient cases, with each case including three longitudinal visits to simulate real-world glaucoma monitoring and progression assessment.

Participants used two versions of the AI interface across these cases:

Unimodal Interface: Participants had access to the complete clinical dataset for each patient including fundus photographs, OCT scans, visual field test results, and patient history. However, the AI system provided diagnostic output only for fundus images. This output consisted of a diagnostic label and a confidence score. No AI generated analysis or interpretation was shown for OCT, visual fields, or historical data.

Multimodal Interface: This version also presented the same full set of clinical information. However, the AI system provided integrated diagnostic outputs across all available modalities. These included disease progression insights, from OCT and visual field, and patient background.

Importantly, the participants were informed both systems are powered by the same AI model, having same accuracy and produced identical diagnostic conclusions for each case and visit. This ensured that any differences in user perception or trust were due to the modality and scope of AI output, not differences in diagnostic accuracy.

To reduce potential bias, participants first reviewed both patient cases using the unimodal interface, followed by the same cases with the multimodal interface. This fixed order ensured that participants did not develop a preference for the multimodal system based on early exposure, allowing for a more balanced and controlled comparison between these systems.

5.6.3 Study Procedure

Each participant followed a structured, repeatable workflow for both AI systems conditions:

Consent and demographics: Participants were first asked to provide informed consent as given in Appendix B.2. Following this, their gender, years of clinical experience, and familiarity with AI were recorded to analyze potential influences on the outcome measures across interface conditions as given in Appendix A.3.

Pre-attitudes toward AI: Before interacting with the AI systems, participants completed a questionnaire assessing their general acceptance of AI in clinical practice (given in Appendix B.3). This baseline measure was later compared with their responses following interaction with the unimodal and multimodal interfaces.

Patient Data Review: At each visit, participants first reviewed the available clinical data (e.g., imaging, history, test results).

AI output consultation: After reviewing the clinical data, participants accessed the AI output relevant to the interface condition. In the unimodal interface, this output

included a diagnostic label, suspicion level, confidence score, and information about risk factors and recommendations, based solely on fundus images. Participants also reviewed the AI's detailed output for both the right (OD) and left (OS) eyes.

In the multimodal interface, in addition to the same information provided in the unimodal condition, the participants consulted the AI output individual findings from OCT and visual field data for both OD and OS, and also the integrated findings.

Diagnostic decision: Participants were prompted to make a final diagnostic judgment while verbalizing their thought process using a think-aloud protocol. This method captured their reasoning, trust in the AI, and engagement with the data.

Post visit questions: After each visit, participants responded to a short set of questions addressing their understanding of the case, trust in the AI's output, and the degree to which it influenced their decision-making. Those questions are provided in appendix B.3

Post case questionnaires: After completing all three visits for a single patient case, participants filled out a structured questionnaire assessing trust (provided in appendix B.3) and acceptance (provided in appendix B.3) of the system just used.

Repeat for second interface: The same procedure was then followed using the second patient case.

5.6.4 Measures

To evaluate participant responses and experiences, both quantitative and qualitative measures were used:

Trust was measured using a carefully designed questionnaire to capture input data

modality, information quality, technical accuracy, output explainability and system quality.

Acceptance was measured through a modified version of Unified Theory of Acceptance and Use of Technology (UTAUT) with an additional dimension of input preference.

5.7 Methodological Choice

5.7.1 Mock-Up vs. Real AI

We did not develop an actual AI system because our study's purpose and scope were not to evaluate algorithmic performance, but rather to understand how clinicians respond to differences in data modality. To achieve this, we deliberately employed a mock-up AI system, which was a methodologically sound choice. Using a mock-up allowed us to maintain experimental control by providing consistent outputs across all participants and removing variability introduced by backend algorithms. This also helped us avoid the confounds of a black-box system. Real AI models can be opaque and unpredictable, which could shape participants' trust responses based on algorithmic opacity rather than the interface or data modality. By eliminating this ambiguity, the mock-up enabled us to isolate and cleanly test our theoretical constructs. Furthermore, ecological validity was maintained by co-developing the system design with optometrists to reflect realistic diagnostic workflows. Participants perceived the system as authentic, with some even remarking that it resembled a tool they would want to use in practice. To avoid bias, we did not disclose during the sessions that the system was a mock-up, but we debriefed participants afterward to ensure ethical transparency.

5.7.2 Use of Fundus images in unimodal AI system

We chose fundus photographs as the input modality for our unimodal AI glaucoma model due to their unmatched accessibility, cost-effectiveness, and proven efficacy in detecting glaucomatous changes with high accuracy. Fundus cameras are significantly more affordable and widely available than alternative imaging modalities, such as OCT, making them particularly suitable for large-scale implementation in primary care and resource-limited settings [Coan et al., 2023]. The advent of portable, low-cost fundus cameras has further extended their utility, enabling efficient population-based screening in diverse and underserved communities [Das et al., 2023].

AI models trained on fundus photographs have consistently demonstrated strong diagnostic performance in identifying characteristic structural abnormalities of the optic nerve head, validating their use as a robust and practical tool for high-throughput screening [Hemelings et al., 2020]. While OCT provides more detailed and quantitative structural information, its higher cost and reliance on specialized equipment limit its feasibility for widespread deployment. Consequently, fundus-based AI systems are commonly adopted for glaucoma detection, and we followed this established approach in developing our unimodal AI system.

In this study, we focused solely on fundus photographs for the unimodal AI system and incorporated all available data sources in the multimodal AI system. This focus remains a limitation, as it does not explore the potential of other data types or combinations. Investigating how different unimodal and multimodal configurations affect user trust and acceptance of AI systems represents an important direction for future research.

5.7.3 Potential Biases and Mitigation Strategies

This study employed a controlled experimental design to examine the impact of AI system modality on user trust and acceptance. While some biases are inherent to the experimental context, the design incorporated deliberate strategies to mitigate or systematically account for their effects. Several potential biases inherent to the experiments were identified and addressed as follows.

Ordering Bias (Sequence Effect): The multimodal system was intentionally presented after the unimodal system because it provided richer information. Presenting the multimodal system first might have biased participants to prefer it and undervalue the unimodal system that followed. By showing the unimodal system first, participants experienced the simpler condition before being exposed to the richer, more advanced system, allowing us to more accurately assess how increasing information richness influences trust and acceptance.

To further mitigate retrospective bias, where participants' impressions of earlier conditions might be influenced by later experiences, measurements were collected immediately after each phase. This ensured that evaluations of each system reflected participants' immediate experience, rather than being colored by later comparisons.

Framing Effect: The study intentionally leveraged framing effects by varying the source and richness of AI explanations (fundus-only in unimodal versus all modalities in multimodal) while keeping diagnostic accuracy and patient data constant. This was not viewed as a bias to eliminate but as a deliberate experimental manipulation to examine how explanation framing influences user trust and acceptance.

Automation Bias and Complacency: Human-automation interaction is prone to over-reliance (automation bias), reduced vigilance (complacency), and deference to per-

ceived expertise (authority bias). Rather than attempting to eliminate these effects, the study allowed them to emerge naturally by not disclosing to participants whether they were interacting with a real AI system or a mock-up. This approach preserved ecological validity and enabled observation of authentic user responses to automation across modalities.

Between-Subject Variability: A within-subject design was employed to eliminate confounding due to individual differences across participant groups. Since each participant experienced both the unimodal and multimodal conditions, their own baseline skills, preferences, and trust tendencies remained constant, enabling more direct attribution of observed differences to the AI system modality itself. This reduced bias from inter-individual variability and strengthened the internal validity of the comparisons.

5.8 Next Chapters

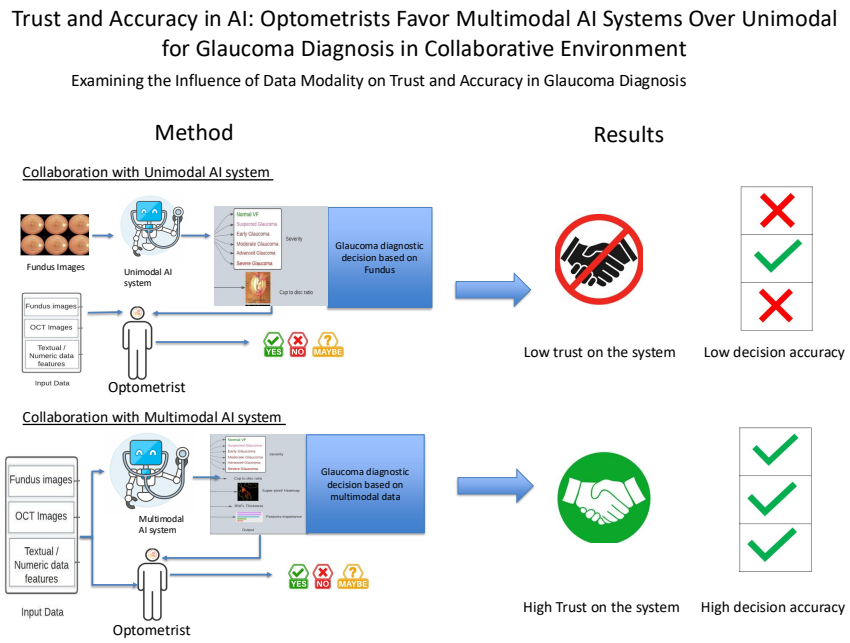
The next chapters 6 7 presents effects of unimodal and multimodal interfaces on trust and user acceptance. By comparing user engagement with unimodal and multimodal systems, these studies aim to identify which system most effectively support clinical reasoning, inspire confidence, and align with real-world diagnostic workflows. This evaluation will offer a deeper understanding of the AI systems that can be used in collaboration with optometrists.

Chapter 6

Trust and Accuracy in AI: Optometrists Favor Multimodal AI Systems Over Unimodal for Glaucoma Diagnosis

This chapter is based on the submitted article:

Ghaffar F, Islam Y, Furtado N, Burns C. Trust and Accuracy in AI: Optometrists Favor Multimodal AI Systems Over Unimodal for Glaucoma Diagnosis. *Computer in Biology and Medicine.*



Graphical Abstract

6.1 Abstract

Background

User trust in Artificial Intelligence (AI) is crucial for the successful collaboration of human and AI systems. This becomes even more significant if AI systems are to be used in clinical practice for diagnosing diseases, such as detecting glaucoma. However, user trust and accuracy are influenced by the modality of data used by AI systems, thus ultimately affecting human-AI collaboration.

Objective

The objective of this study was to discover the modality of an AI system that aligns best with an optometrists' mental models. This was achieved by comparing trust levels between unimodal and multimodal AI systems used for glaucoma diagnosis and further assessing how different dimensions of trust are impacted by these modalities.

Methods

We conducted a study with 20 optometrists who were presented with both unimodal and multimodal AI mock-up systems that were specifically designed for glaucoma diagnosis. We measured trust across five key dimensions using Likert scale questionnaires. For the data analysis, we employed statistical analysis to assess trust in the unimodal and multimodal AI systems. Optometrists' performance was evaluated based on the alignment of their decisions with those of the unimodal and multimodal AI systems.

Results

The results showed that the multimodal system had a higher average trust rating of 3.98 on a Likert scale, indicating greater trust compared to the unimodal system, which had an average trust rating of 3.00. This difference was statistically significant ($p<.001$), with further analysis revealing significant variation across all trust targets ($p<.001$). Additionally, optometrists demonstrated higher F1 scores with the multimodal system (43.1%) compared to the unimodal system (37.3%), while accuracy remained comparable between the two systems (63.0% for multimodal and 63.3% for unimodal). However, major differences across these metrics were observed in relation to optometrists' expertise.

Conclusions

Optometrists have high trust and decision accuracy when working with multimodal AI systems compared to AI systems that use unimodal ones. Their decisions are more aligned with multimodal systems, as both use the same data, leading to a higher level of trust. To have effective human-AI collaboration, AI systems in healthcare should use the same multimodal data as that of the user's mental model.

6.2 Introduction

The rapid integration of artificial intelligence (AI) into fields like healthcare, design, and finance has sparked intense curiosity about how people interact with and trust these systems [Dietvorst et al., 2015, Glikson and Woolley, 2020, Shevtsova et al., 2024]. Trust is a deciding factor in determining the success of AI technologies [Taddeo and Floridi, 2018] as it affects the AI system adoption [Stevens and Stetson, 2023]. People are more likely to work with AI when they are confident in its reliability and ability to support their decision-making. As AI systems evolve and become more accurate, it is becoming increasingly important to discover how users' trust in a system varies depending on the modalities of data presented, especially in the healthcare domain, where AI systems can be used to assist in clinical decision-making. Understanding these systems is crucial for situations where AI is supposed to team up and collaborate with human experts to make decisions.

In healthcare, particularly in ophthalmology, the difference between unimodal and multimodal AI systems is very relevant. AI has shown significant potential in the diagnosis and management of complex eye diseases [Ting et al., 2019] such as glaucoma [Xue et al., 2022, Ashtari-Majlan et al., 2024], which is a leading cause of irreversible blind-

ness all over the world [Sun et al., 2022]. These AI systems are developed to support optometrists and give a second opinion on their diagnostic decisions. Some systems use a single type of diagnostic data, such as fundus imaging [Fu et al., 2018], while others combine multiple data types, including fundus imaging, visual field data, optical coherence tomography (OCT), and patient histories [Lim et al., 2022] and make ensemble models to detect glaucoma. The purpose of using multimodal data was to enhance the accuracy of the systems [Sun et al., 2023], and present them as counterparts or collaborators of human doctors. However, the effectiveness of these systems is not determined solely by their technical accuracy, but also by how well they align with human mental models and how much trust they inspire in the users.

Trust is a well-studied subject, and researchers deemed it essential for the successful integration of AI systems. Trust is defined as the willingness to be vulnerable to another's actions based on expectations of reliability, accuracy, and integrity [Lee and See, 2004, Mayer et al., 1995a, Rousseau et al., 1998]. It is not a static quality, but it is shaped by several factors. Research shows that user trust in AI systems is strongly influenced by the data it uses, the information and system quality, the accuracy of the AI system and the reliability, transparency, and clarity with which they explain their decisions [Kaplan et al., 2023, Lukyanenko et al., 2022, Cai et al., 2019, Miller, 2019a]. The ethical dimensions of trust play a crucial role, showing that AI systems adhering to ethical principles can significantly enhance user trust [Yang et al., 2023]. However, this comes with the caveat that the user's trust in the system needs to be correctly balanced. Too little trust in an AI system can risk underutilization or complete disregard, while too much trust risks a danger of over-reliance, resulting in potential misuse or other adverse outcomes [Kostick-Quenet et al., 2024, Parasuraman and Riley, 1997]. Furthermore, trust in AI evolves in response to real-time interactions, as emphasized in [Textor et al., 2022],

which highlights the importance of adaptive trust calibration over time. Therefore, the right balance of trust is essential for the safe and effective use of AI.

Trust in AI systems is particularly crucial, as it directly influences the quality of collaboration between users and AI technologies [Knop et al., 2022], specifically in healthcare, where mistakes can have serious consequences. It is shaped by several factors such as transparency, explainability, usability, and users' prior experience with AI[Tucci et al., 2022b]. However, one important and underexplored determinant of trust is the modality of input data used by AI systems. Clinicians typically employ data from multiple data sources [Ghaffar, 2024] when making decisions, while AI systems may rely on unimodal or multimodal inputs, often achieving comparable predictive accuracy. The choice of data modality shapes how information is represented, interpreted, and acted upon. Each modality emphasizes different aspects of a clinical problem, leading to distinct inferential pathways and cognitive demands [Cao et al., 2009]. Multimodal systems can offer richer, more contextually grounded representations that align closely with human reasoning, yet they may also introduce greater cognitive load and integration complexity. In contrast, unimodal systems may be cognitively simpler and easier to interpret, but risk oversimplifying complex clinical contexts and omitting critical information. These cognitive differences can affect users' perceptions of trustworthiness [Kalatzis et al., 2023].

However, despite this conceptual relevance, a clear gap exists in empirical research. This gap gives rise to several challenges. First, it is necessary to investigate how unimodal versus multimodal data inputs affect clinicians' trust in AI systems. Second, research must examine how modality influences diagnostic accuracy during AI-assisted decision-making. Moreover, it is important to explore the variation of trust across modality types in relation to changes in system-level factors (e.g., accuracy) and user-level factors (e.g., expertise.) Addressing these challenges is essential to developing AI systems that are not only accurate

but also perceived as trustworthy and well-aligned with clinical workflows and reasoning practices. Solving these challenges is crucial for guiding the design of AI systems that are not only accurate but also trusted, interpretable, and aligned with clinicians' cognitive processes. Insights into how data modality influences trust and performance can help tailor AI tools to different clinical settings and users's needs, supporting safer and more effective human-AI collaboration.

To address these challenges, our study makes a substantial contribution by developing and rigorously evaluating both unimodal and multimodal AI mock-up systems specifically designed for glaucoma diagnosis using real patient data. We developed this system in close collaboration with experts in AI, human factors, and optometry to ensure they align with clinical needs and reflect real-world decision-making processes. After that, we conducted in-depth interviews with optometrists to gather insights into their trust, experiences, and perceptions of these two systems. The optometrists' trust was evaluated across five key targets: input data modality (IDM), information quality (IQ), technical accuracy (TA), output explainability (OE), and system quality (SQ), and measured using a 5-point Likert scale questionnaire. These targets were informed by established trust frameworks, particularly the Performance–Process–Purpose (PPP) model [Lee and See, 2004]. Based on our conceptual rationale, IDM and TA are primarily associated with Performance, OE and IE with Process, and SQ with Purpose, while acknowledging that these associations are not mutually exclusive.

While our primary focus is on how data modality influences trust, prior research indicates that individual user characteristics, such as gender and professional expertise, also shape trust in technology. Psychological studies have reported gender based differences in trust behavior [Haselhuhn et al., 2015, Rau, 2011], and recent work in AI contexts has identified gender specific variations in perception and trust [Armutat et al., 2024,

Gallimore et al., 2019]. Expertise also influences interactions with AI by shaping how users assess and calibrate trust [Bayer et al., 2022, Nourani et al., 2020]. We therefore included gender and expertise as exploratory variables to assess their potential role in shaping trust in unimodal versus multimodal AI systems, with the aim of informing equitable and inclusive AI design.

This study provides a comprehensive understanding of how differing data modalities influence optometrists' trust in AI systems and their diagnostic performance. Furthermore, the findings offer guidance for improving AI system design and deployment in clinical practice. Our study is guided by the following research questions:

1. How does clinicians' trust in AI systems vary between unimodal and multimodal AI systems when diagnosing glaucoma?
2. How do user factors (demographics and expertise) and system factors (accuracy) influence trust when using unimodal versus multimodal AI systems?
3. How does clinicians' use of unimodal and multimodal AI systems affect their accuracy of glaucoma diagnosis?

Our findings indicate that optometrists exhibited higher trust in the multimodal system (mean = 3.98) compared to the unimodal system (mean = 3.00), with a statistically significant difference ($p<.001$) in trust between the two systems. Across the five targets, the multimodal had a mean rating of (3.9, 3.9, 3.6, 3.9, and 4.7) whereas unimodal mean rating was (2.1, 2.9, 2.6, 3.2, 4.3) with statistically significant difference ($p<.001$) between both systems. There was no significant effect of demographic factors observed. Additionally, trust ratings were more consistent across clinicians for multimodal systems, reflecting their greater reliability and alignment with clinical workflows. Optometrists demonstrated

higher performance across precision, recall, and F1 score when using the multimodal system (72.4%, 44.4%, and 43.1%, respectively) compared to the unimodal system (51.2%, 39.5%, and 37.3%). Accuracy was comparable between the two systems (63.0% for multimodal vs. 63.3% for unimodal). Furthermore, the correctness of system predictions significantly influenced the calculated performance metrics, with notable differences observed between correct and incorrect predictions. Further analysis based on optometrists' experience levels revealed additional variations in these metrics, indicating that modality, system accuracy, and user expertise collectively shaped diagnostic performance. Overall, the multimodal system appeared to better align with the expectations of optometrists.

In this study, we address the challenge of understanding how data modalities influence clinicians' trust in AI systems and how their diagnostic performance is affected. By developing and evaluating both unimodal and multimodal AI mock-up systems, conducting user studies and quantitative analyses, we found that the optometrists had higher trust in the multimodal system, which also enhanced their diagnostic performance. These findings provide valuable guidance for designing AI systems that are trusted by clinicians in general and for glaucoma diagnosis specifically.

Significance of Study

Problem Examining the effect of AI system modality on optometrists' trust and decision accuracy in human-AI collaborative environment.
What is Already Known Existing research has established that trust is crucial for human-AI collaboration. However, there is a limited exploration of how the modality of AI systems affects optometrists' trust and decision-making accuracy in human-AI collaborative environments specifically in glaucoma diagnosis.
What This Paper Adds This study aims to explore the effect of AI system data modality on user trust and decision accuracy. Unimodal and multimodal AI mock-up systems were developed for glaucoma diagnosis and optometrists' trust and accuracy were measured through interview and trust questionnaire.

6.3 Methodology

This section details the research methodology we used to investigate trust in unimodal and multimodal AI systems within a collaborative environment. The methodology comprises several steps: the data collection process from a clinical setting, the development of prototype interfaces for glaucoma diagnosis, and a series of structured interviews with optometrists. Each subsection addresses specific aspects of the methodology, including outreach strategies, participant demographics, the AI-optometrist collaborative decision-making process, and the subsequent data analysis.

6.3.1 Ethical Consideration

This study, under application number 45248, received ethical approval from the Delegated Research Ethics Committee (DERC) of the University of Waterloo on December 4,

2023 and given in Appendix B.1. Informed consent was obtained from all participating optometrists prior to the commencement of the interviews given in Appendix B.2. Participants were fully briefed on the study's objectives, procedures, potential benefits, and risks, as well as their right to withdraw from the study at any point. Data collected during the interviews were anonymized and deidentified to ensure confidentiality, with all information securely stored on a protected server. Each participant was compensated CAD 150 for their two-hour involvement, recognizing the value of their professional input and encouraging detailed participation.

6.3.2 Patients' Data Collection

To develop representative AI mock-up interfaces for assessing collaborative decision-making between AI systems and optometrists, we used real patient data to ensure clinical realism and contextual validity. We accessed the entire available database of glaucoma patients from the School of Optometry and Vision Sciences at the University of Waterloo. From this database, we selected records from individuals who had explicitly consented to the use of their data for research purposes and chose eight patient records based on the availability of extensive longitudinal data and evidence of diagnostic ambiguity or changes across at least three clinical visits. The selection of data attributes was guided by prior research in the field, including studies [Leske et al., 2008, Le et al., 2003, de Voogd et al., 2006, Gordon et al., 2002, Group et al., 2007]. The dataset comprised comprehensive ocular health metrics and external factors, which enabled us to develop a representative unimodal and multimodal interface for assessing the AI-optometrist collaborative decision-making process.

6.3.3 Data Refinement and Processing

We refined and processed the collected data for developing an AI mock-up system and for the interview component of the study. From the initial pool of patients, two were selected for use in the development of mock-ups and interviews based on comprehensive discussions with both an AI expert and an optometrist. The selection criteria were targeted at selecting instances that would be ideal cases for both the unimodal and multimodal AI systems, as well as those with sufficient representative data for glaucoma diagnosis. The selected cases were drawn from the suspect diagnostic range, which is typically associated with clinical ambiguity; however, no formal difficulty classification was applied. We selected these cases to ensure that the AI system would have the ability to assist in glaucoma diagnosis and to promote a collaborative decision-making environment between the AI and the optometrist. Moreover, it would take a significant amount of time for an optometrist to review more than two records with multiple visits. Each of the two selected patients had three visits, and we anticipated that participants would take about 10-15 minutes to review each patient visit and consult each AI system. Given that there were two versions of the AI system, each participant reviewed a total of 12 visits. This limit to two cases ensured a manageable workload while also providing sufficient data to assess the user performance and trustworthiness of the AI systems.

6.3.4 AI Mock-Up System Development

The AI mock-up system was designed and developed using real patient data in Figma, providing an interactive interface for clinicians. The system combines patient information and AI outputs to facilitate clinical decision-making, as shown in Figure 6.1

Patient information is classified into four sections based on the data source: general

information, fundus imaging, OCT, and visual field marked as c1, c2,c3 and c4 in Figure 6.1 respectively. Each section is accessible through a separate tab. This general section includes patient demographics, ocular history, medical history, medication usage, social history, and family history. Additionally, it includes ocular information such as pupil testing, extraocular muscle movements, and anterior segment details for both the right eye (OD) and the left eye (OS). The fundus imaging section displays the fundus photographs of both OD and OS. The OCT section contains the retinal nerve fiber layer (RNFL) single exam and trend reports for both OD and OS. Finally, the visual field section includes single field analysis and glaucoma progression analysis (GPA) with baseline GPA follow-up data.

The AI system output is divided into two parts: AI diagnosis and explanation. The first part is the AI diagnosis part [Figure 6.1(A)], which contains all the main decisions provided by the AI system. It includes the primary diagnosis [Figure 6.1(a1)](indicating whether the patient has glaucoma, is a glaucoma suspect, or does not have glaucoma), the severity level [Figure 6.1(a2)] that provides a granular assessment of the diagnosis (high, medium, and low for patients classified as suspect and early, mild, moderate, or severe for patient classified as having glaucoma), and the confidence level [Figure 6.1(a3)] of the AI system in its diagnostic decision. Diagnoses and suspicion levels in the mock system were calculated based on the majority choice from a prior study [Ghaffar, 2024], while confidence scores were computed as the percentage of participants selecting that majority diagnosis. A summary of these values, across all cases and visits, is provided in the Figure 6.1, which presents the calculated diagnosis decisions, severity levels, and confidence scores used for each case in the mock AI system and shows consistency across the systems.

The explanation section [Figure 6.1(B)] details the reasoning behind the AI system's decision, a conclusion, and recommended actions for the next steps. Explanations were manually crafted to reflect the reasoning process of each system. These explanations

varied by modality: unimodal systems referenced only one type of input data (Fundus image), whereas multimodal explanations integrated all type of available input data. Each explanation consisted of identified decision factors [Figure 6.1(b1)], a diagnostic findings [Figure 6.1(b2)] , and conclusion and recommended action [Figure 6.1(b3)]. The initial explanations were derived from qualitative analysis of that prior study [Ghaffar, 2024], which explored how participants interpreted and reasoned through similar patient data. Those qualitative data provided the basis for developing explanation content that aimed to mirror authentic clinical reasoning processes. The explanations were then iteratively refined in consultation with a practicing optometrist to ensure clinical accuracy, relevance, and consistency across all cases.

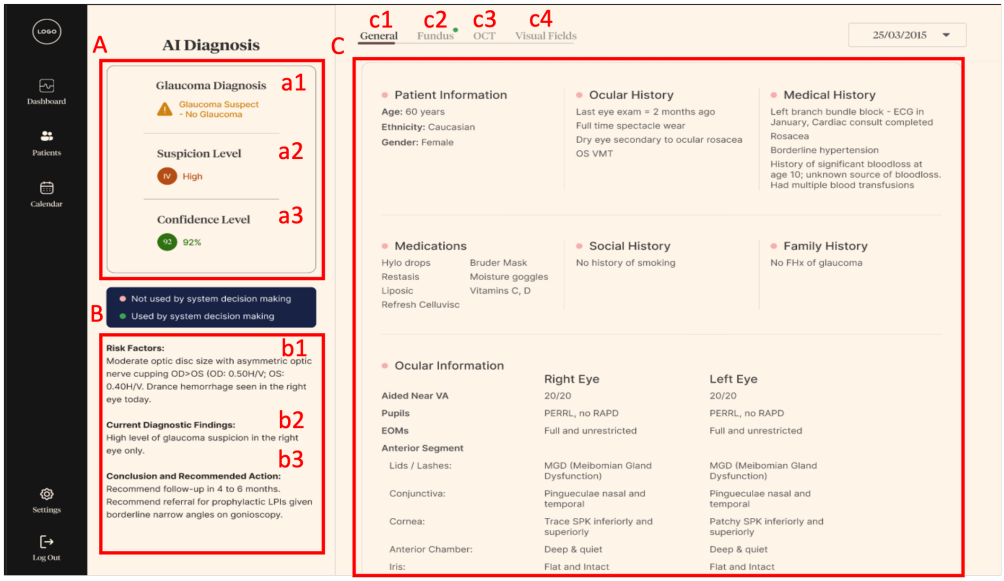


Figure 6.1: AI mock-up system interfaces (Left-hand side shows the AI global output and the right-hand side includes all the patient information)

The designs for both the unimodal and multimodal AI systems had some key differences. The participants were informed of these specific features to ensure participants were aware

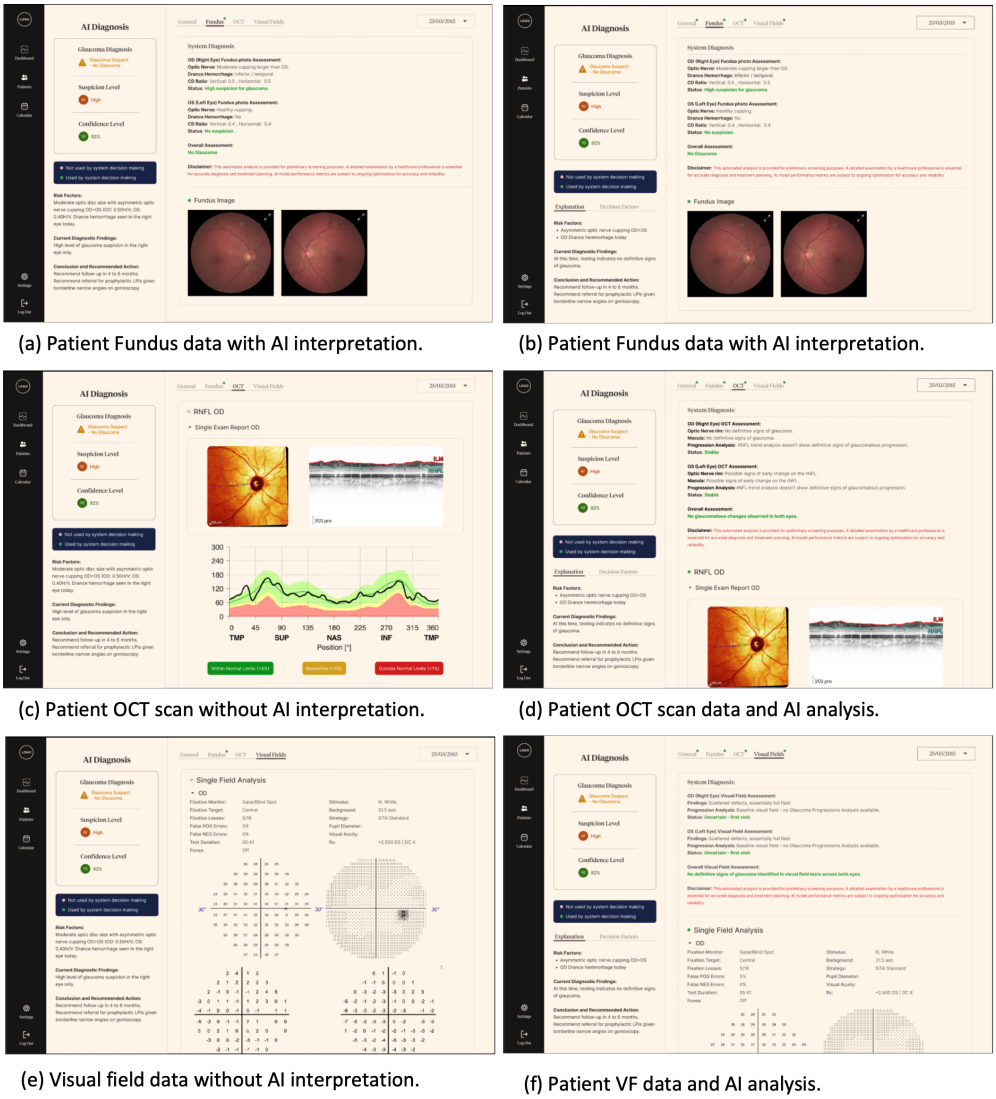
of the system modality they were interacting with. For the unimodal system, decision-making was based solely on fundus images, and an AI explanation was provided for both the OD and OS within the fundus imaging tab. An overall output of diagnostic decisions was also present on the AI diagnosis panel. However, explanations for the unimodal system were based only on fundus imagery.

The multimodal AI system used an ensemble model or multi-agent approach, combining individual AI systems for fundus images, OCT, and visual field data. Each system independently generated predictions and explanations in its own section, providing local explanations when a user would click on a given tab. These explanations, combined with the general patient information, were used to generate the global explanation visible in the AI diagnosis panel. The decision factor visualization was included in the multimodal system only, as the unimodal system used only fundus images as a decision factor. Detailed side-by-side views are provided in Figure 6.2.

6.3.5 Optometrist Interviews

We recruited and interviewed 20 optometrists from the Waterloo region of Canada. Most of them were affiliated with the University of Waterloo's School of Optometry and Vision Sciences. Each optometrist participated in a 100–120-minute online interview conducted in English via Microsoft Teams or Zoom.

We employed a within-subjects design with repeated measures to facilitate a direct comparison of the systems. Each participant was presented with both unimodal and multimodal AI system cases in a specific sequence to prevent bias (As in our previous study, optometrists showed preference for multimodal system). Participants first reviewed Case A using the unimodal AI system, followed by Case B with the unimodal AI system. Next,



(a) Patient Fundus data with AI interpretation.

(b) Patient Fundus data with AI interpretation.

(c) Patient OCT scan without AI interpretation.

(d) Patient OCT scan data and AI analysis.

(e) Visual field data without AI interpretation.

(f) Patient VF data and AI analysis.

Figure 6.2: Unimodal and multimodal AI mock-up system, side-by-side comparison of all sections. The left-hand side shows a unimodal system and the right-hand side shows a multimodal system

participants reviewed Case A again, but this time using the multimodal AI system. Finally, participants reviewed Case B again using the multimodal AI system. This sequence and structure ensured that participants did not develop a preference for the multimodal

system due to the initial encounter. Each case included three visits except the unimodal Case A. Figure 6.3 shows the interview flow and detail of data.

After each visit, participants were asked questions to evaluate the AI system's predictions for each visit, marking them as correct or incorrect and whether they agreed or disagreed with the AI decision, as well as providing their confidence in the system on a Likert scale. Additionally, after completing each case, participants filled out a questionnaire related to their trust in the AI systems mentioned in AppendixB.3. The trust questionnaires were Likert scale questionnaires comprised of 18 questions, which we classified into five targets. The low value "1" on the Likert scale represented very strong agreement, and the high value "5" represented disagreement. To avoid potential confusion and to ensure that higher values reflected higher trust, the responses were inverted during the analysis stage. The questionnaire is provided in B.3. This approach allowed us to evaluate trust and the user's accuracy using different modalities.

Throughout the interviews, participants used the 'think aloud' protocol to review patient records and consult the AI systems, aiding in understanding their decision-making processes.

6.3.6 Data Analysis

We used the open-source statistical software JASP for statistical analysis. To begin, we provided a summarized statistic of the demographics of the participating optometrists. Key metrics included mean, median, and standard deviation for variables such as years of professional experience, the number of patients seen weekly, and the number of glaucoma patients seen weekly. We also analyzed and plotted the gender, ethnicity, and age distributions. The participants were classified into novice and experts were carried out in line

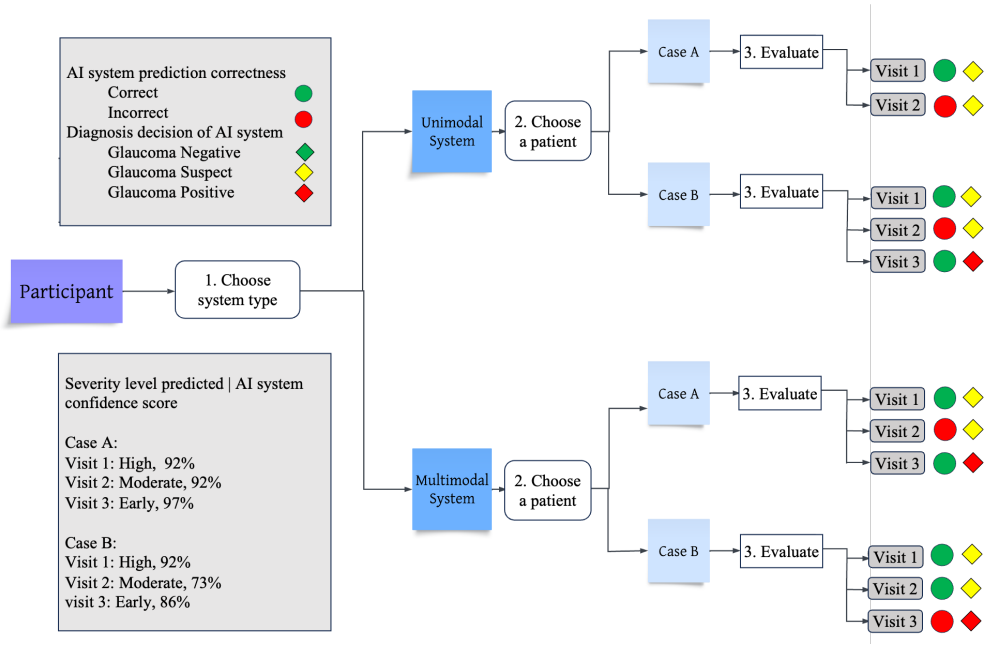


Figure 6.3: Interview flow, detail of cases, and patient

with previous study based on their years of experience and frequency of consultations with patients with glaucoma inline with previous study [Ghaffar, 2024]. Further, we carried out a detailed analysis as follows.

First, we evaluated the reliability of the trust questionnaire using Cronbach's α (α) to ensure internal consistency across both unimodal and multimodal conditions. Cronbach's alpha coefficients and inter-item correlations were calculated for each dimension of trust to confirm that each group of questions represents a single dimension. After that, we aggregated the responses for each dimension and used the aggregated score in further statistical analysis. Further, we computed Inter-item correlation (ICC) across all data and separately for the unimodal and multimodal systems to ensure the reliability of the data and consistency across participant evaluations.

Second, participants' responses across the five trust dimensions were analyzed. De-

scriptive statistics were calculated, and distributions were plotted to compare unimodal and multimodal systems. The Shapiro-Wilk test (SW) was used to assess normality at both the item level and the composite (target) level. Based on these results, either the non-parametric Wilcoxon Signed-Rank test (WSR) or the paired t-test (Student's t) was applied to evaluate differences between modalities at both levels. To examine potential interaction effects between modality and trust dimensions, a 2 (Modality: Unimodal vs. Multimodal) $\times$ 5 (Targets: Trust Dimensions) repeated measures ANOVA was conducted using composite scores (averaged across three items per target). Omega squared (ω^2) was reported as an estimate of effect size. This multifactorial analysis evaluated the effect of modality on trust across all targets simultaneously.

Next, the influence of user and system factors were analyzed. A Repeated Measure ANOVA test was employed to assess the user factors' influence included as exploratory variables, and their interaction with data modality. The modality was considered as within-subject variables (Modality), while the user factors were considered as between-subject variables. For the system factors, the Mann-Whitney U test was used to identify significant differences in trust scores between the unimodal and multimodal AI systems.

Lastly, the optometrists' performance was evaluated by calculating Accuracy (ACC), Precision (PREC), Recall (REC), and F1-Score (F1). Following each consultation, optometrists recorded their diagnostic decision by selecting from predefined categories (Glaucoma Positive, No Glaucoma, Suspect). A confusion matrix was constructed by comparing the optometrists' diagnostic decisions to the ground truth clinical diagnoses across the three categories. For each class, a True Positive (TP) was defined as a case where the optometrist correctly classified a patient into the true diagnostic category (e.g., correctly diagnosing a Glaucoma Positive case as Glaucoma Positive). A False Positive (FP) was defined as a case where the optometrist incorrectly classified a patient into a class that

did not match the true diagnosis (e.g., diagnosing a No Glaucoma case as Glaucoma Positive). A False Negative (FN) occurred when the optometrist failed to classify a patient into their true class, instead assigning a different diagnosis (e.g., diagnosing a Glaucoma Positive case as Suspect or No Glaucoma). A True Negative (TN) was defined as a case where the optometrist correctly identified that a patient did not belong to a particular class (e.g., correctly not diagnosing a No Glaucoma case as Glaucoma Positive). Based on this confusion matrix, ACC was calculated using the overall proportion of correct predictions. PREC, REC, and F1 were macro-averaged across the three diagnostic categories to ensure that each class contributed equally to the final metrics, thereby addressing potential class imbalance and providing a more balanced evaluation of performance.

ACC was calculated as the proportion of correct diagnostic decisions by optometrist among all diagnosis decisions.

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.1}$$

PREC was calculated by averaging the precision scores across all diagnostic classes, where each class-specific precision was defined as the proportion of correct positive diagnoses among all diagnoses identified as positive by the optometrist.

$$\text{PREC} = \frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{TP_i + FP_i} \tag{6.2}$$

REC was calculated by averaging the recall scores across all diagnostic classes, where recall for each class was defined as the number of correct diagnoses identified by the optometrist out of all actual cases of that class.

$$\text{REC} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i} \tag{6.3}$$

F1 was calculated by averaging the F1 scores across all diagnostic classes, where the F1 score for each class was defined as the harmonic mean of precision and recall.

$$\text{F1} = \frac{1}{C}\sum_{i=1}^{C}\frac{2 \times TP_i}{2 \times TP_i + FP_i + FN_i} \tag{6.4}$$

By applying these statistical methods, we provided a comprehensive evaluation of the trust and performance of optometrists using unimodal and multimodal AI systems.

6.4 Results

6.4.1 Demographic Characteristics of Optometrists

We recruited a total of 20 optometrists in the study. The mean years of experience as an optometrist was 6.613 years (SD = 8.08). The optometrist saw an average of 59.12 (SD = 24.72) patients weekly. Specifically, the number of glaucoma patients seen weekly averaged 7.875 (SD = 6.24) with a median of 6.75. The gender distribution of the optometrists was relatively balanced, mainly dominated by 10 women, 8 men, and only 2 Trans/Non-binary. The ethnicity distribution was diverse. To maintain participant anonymity, ethnicities were grouped, and one participant's self-identified background was reported as 'Other'. The majority of participants identified as being from Asian (South, East, West, including Iranian and Middle Eastern; n = 10) backgrounds followed by Caucasian/European (n

= 9) and one participant identifying as 'Other'. Participants also represented a diverse age range: 9 optometrists were aged 20–30, 5 were aged 30–40, 3 were aged 40–50, 1 was aged 50–60, and 2 preferred not to disclose their age. The gender, ethnicity, and age distributions are plotted to provide a clear understanding of the demographics in Figure 6.4.

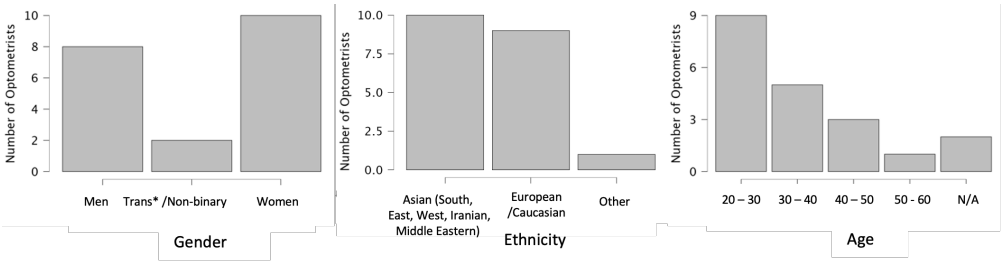


Figure 6.4: Demographic distribution of optometrists

6.4.2 Assessment of Trust Questionnaire Reliability, Validity, and Inter-class consistency

Before conducting comparative analyses, the reliability and validity of the trust questionnaire were assessed to determine whether each group of questions effectively represented a single construct in both unimodal and multimodal conditions.

We first measured the internal consistency of each construct with Cronbach's alpha, which indicated high internal consistency across all targets, with alpha values exceeding 0.7. This suggests that the items within each construct are reliably measuring the same underlying concept. Additionally, Bartlett's Test for Sphericity yielded significant results ($p < .001$) for all constructs, confirming that the data is suitable for factor analysis. Further, Exploratory Factor Analysis (EFA) confirmed the uni-dimensionality of the constructs,

with all eigenvalues greater than 1, supporting a strong and consistent factor structure in both unimodal and multimodal conditions. Confirmatory Factor Analysis (CFA) also corroborates these findings, with factor loading greater than 0.6 for unimodal and above 0.4 for multimodal systems, indicating an acceptable to strong model fit. Collectively, these analyses affirm that the groups of questions reliably and validly measure their respective constructs across different modalities, supporting their use in subsequent comparative analyses.

Furthermore, a high Intraclass Correlation Coefficient (ICC) with values greater than 0.5 for both unimodal and multimodal indicated excellent agreement between participant evaluations, suggesting that the ratings are highly consistent across the 20 raters. The narrow confidence interval of ICC further reinforces the robustness of this finding. These findings suggest that both systems provide reliable and valid measurements of trust, and therefore, averaging the values across constructs is a reasonable approach for further comparative analysis.

Table 6.4.2 shows detailed psychometric analysis of optometrist responses for both unimodal and multimodal systems across all five targets.

6.4.3 Trust Disparities Between Unimodal and Multimodal AI Systems

There are significant differences in the levels of trust exhibited by optometrists when collaborating with unimodal versus multimodal AI systems. Our findings indicate that optometrists exhibited higher trust in the multimodal system (mean = 3.98) compared to the unimodal system (mean = 3.00). A significant difference was found, t (37) = 9.43, p <.001, based on Student's paired t-test, with a large effect size (Cohen's d = 1.53). To

Table 6.1: Summary of psychometric properties for Uni-modal and Multi-modal constructs.

Targets	**Tests**	**Uni-modal**	**Multi-modal**
IDM	α (95% CI)	0.84 (0.78-0.89)	0.82 (0.74-0.88)
	ICC (95% CI)	0.64 (0.52-0.73)	0.6 (0.48-0.7)
	Bartlett's Test	$\chi^2(3) = 109.9, p < .001$	$\chi^2(3) = 93.9, p < .001$
	Eigenvalue	2.29	2.21
	Factor loading (range)	0.62-0.92	0.63-0.99
IQ	α (95% CI)	0.76 (0.71-0.86)	0.71 (0.58-0.8)
	ICC (95% CI)	0.56 (0.45-0.67)	0.45 (0.3-0.58)
	Bartlett's Test	$\chi^2(3) = 55.76, p < .001$	$\chi^2(3) = 47.6, p < .001$
	Eigenvalue	2.01	1.9
	Factor loading (range)	0.63-0.86	0.54-0.87
TA	α (95% CI)	0.94 (0.92-0.96)	0.9 (0.86-0.93)
	ICC (95% CI)	0.8 (0.74-0.85)	0.68 (0.6-0.8)
	Bartlett's Test	$\chi^2(6) = 338.5, p < .001$	$\chi^2(6) = 238, p < .001$
	Eigenvalue	3.41	3.08
	Factor loading (range)	0.86-0.91	0.76-0.89
OE	α (95% CI)	0.92 (0.89 - 0.94)	0.51 (0.40 - 0.63)
	ICC (95% CI)	0.74 (0.67-0.81)	0.51(0.40-0.63)
	Bartlett's Test	$\chi^2(6) = 241.9, p < .001$	$\chi^2(6) = 119, p < .001$
	Eigenvalue	3.25	2.58
	Factor loading (range)	0.80-0.94	0.41-0.87
SQ	α (95% CI)	0.85 (0.81-0.90)	0.87 (0.82-0.90)
	ICC (95% CI)	0.6 (0.49-0.69)	0.62 (0.52 - 0.72)
	Bartlett's Test	$\chi^2(6) = 159.6, p < .001$	$\chi^2(6) = 201.7, p < .001$
	Eigenvalue	2.86	3.03
	Factor loading (range)	0.6-0.93	0.72-0.90

Abbreviations: α (Cronbach's alpha), ICC (Intraclass Correlation Coefficient) CI, (Confidence Interval)

further investigate, we analyzed each target of trust separately.

6.4.3.1 Descriptive Statistics of Trust Targets

We calculated descriptive statistics of the trust questionnaire to get a better idea of the participant rating of the systems. The mean trust scores for these targets in unimodal systems were 2.11 (SD = 0.74) for input data modality, 2.9 (SD = 0.84) for information quality, 2.61 (SD = 0.89) for technical accuracy, 3.18 (SD = 0.99) for output explainability, and 4.28 (SD = 0.78) for system quality. In contrast, the mean trust scores for multimodal systems were higher, with 3.90 (SD = 0.74) for input data modality, 3.85 (SD = 0.64) for information quality, 3.57 (SD = 0.79) for technical accuracy, 3.94 (SD = 0.67) for output explainability, and 4.65 (SD = 0.47) for system quality. The aggregated mean score distribution for each target, seen in Figure 6.5, also shows a clear variation between the optometrist's trust in unimodal vs multimodal systems. The unimodal distributions mostly lie in the range of 1-3 on the Likert scale, except for system quality, while for multimodal, the distribution spread mostly between 3-5.

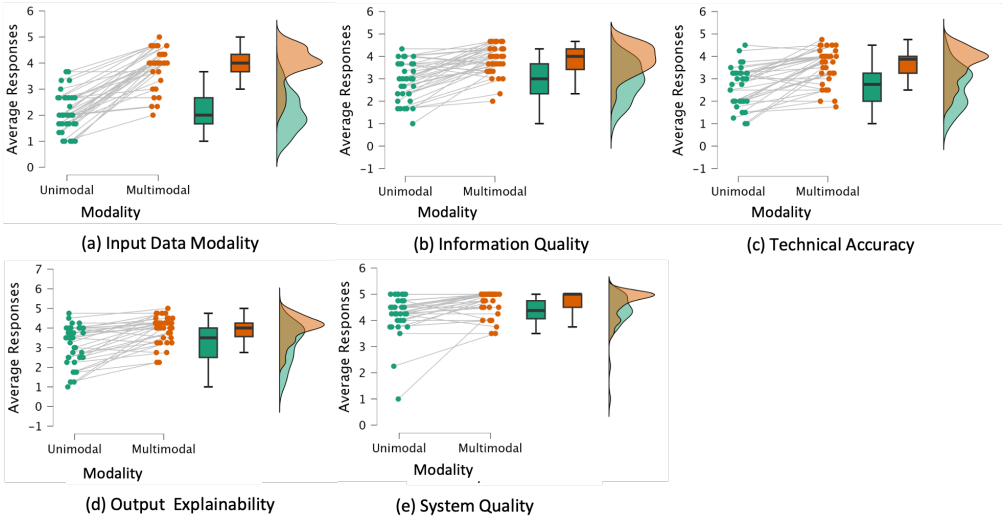


Figure 6.5: Comparison of the distribution of mean score of each target of unimodal and multimodal AI mock-up systems

Across all five targets of trust, the multimodal system received higher mean ratings than the unimodal system, indicating greater trust in the multimodal system. This suggests that participants expressed stronger agreement with trust in multimodal systems. The relatively lower standard deviations observed in the multimodal system in most of the targets suggest that there is less variability in optometrists' responses, which could imply a more uniform perception of trust in a multimodal system. Further details of each question item also follow the same trend. The distributions seen in Figure 6.6 show that for the multimodal system, all items' distributions peak near high value while for the unimodal, it peaks near low value. Additional detailed statistics for each question are provided in B.4.

6.4.4 Inferential Analysis of Trust Variations Between Unimodal and Multimodal AI

To evaluate whether there were significant differences in trust between each target of unimodal and multimodal AI systems, a paired-t test was conducted to assess whether these differences were statistically significant.

First, the normality of the trust ratings was evaluated using the Shapiro-Wilk test. The results indicated significant deviations from normality ($p<.05$) only in the system quality target, justifying the use of non-parametric tests for this target and Student's paired t-tests for all other targets to compare means between unimodal and multimodal systems. The tests revealed significant differences in trust scores between the two modalities.

For input modality, the t-test identified a significant difference, t (37) =11.16, $p<.001$, with the multimodal system showing significantly higher ratings, indicating higher trust with a mean difference of 1.8 (SE = 0.16). The effect size, as measured by Cohen's d, was 1.81, suggesting a large effect favoring the multimodal system. A significant difference was

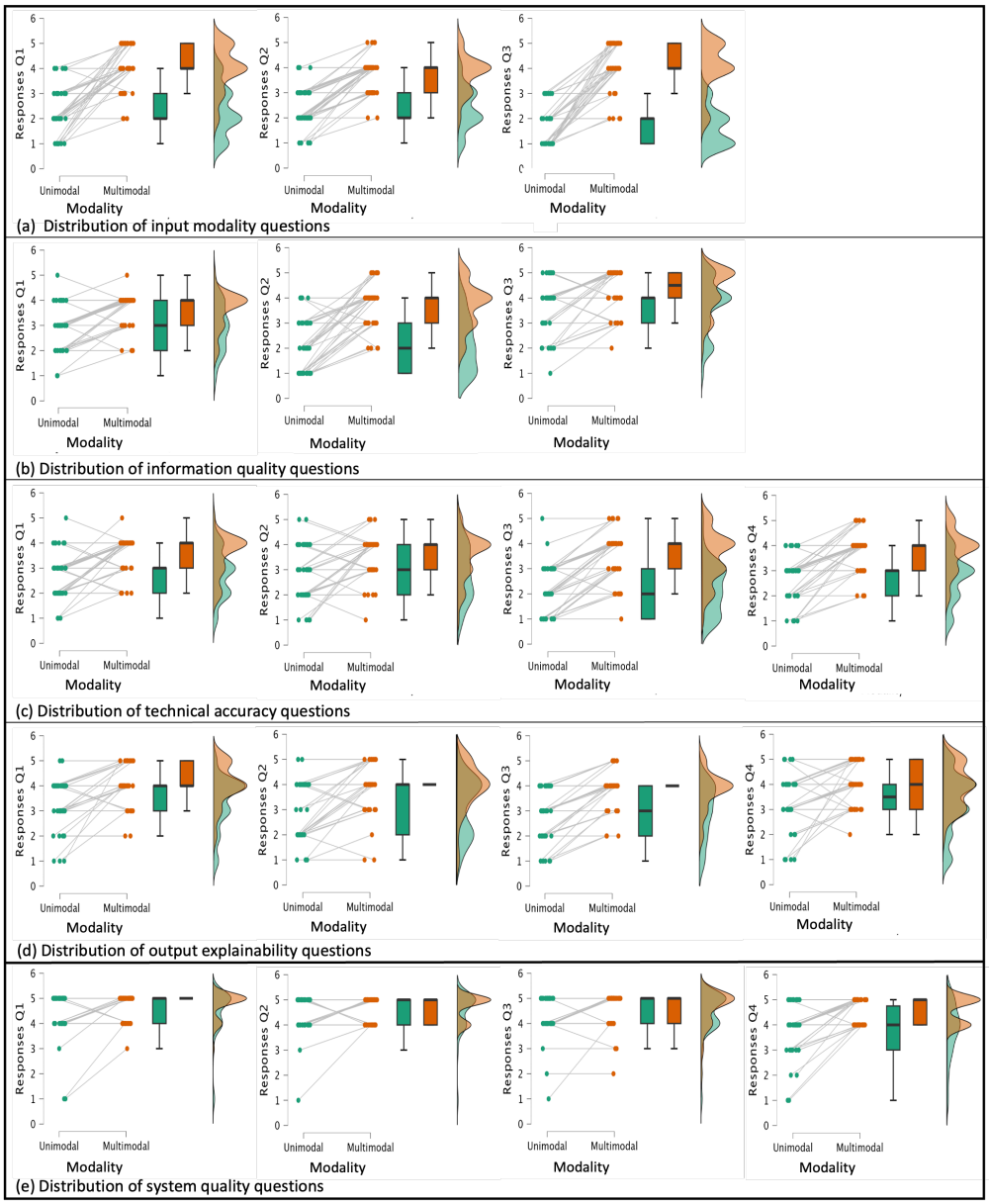


Figure 6.6: Distribution of responses for each question in the data modality target, with means for each group marked.

also found in the information quality target, t (37) =7.1, p<.001, with a mean difference of 0.96 (SE = 0.13) and effect size d=1.15, where multimodal systems received lower ratings indicating higher trust. Similarly, the technical accuracy target showed a significant difference, t (37) =7.0, p<.001, with a mean difference of 0.961 (SE = 0.13) and effect size d=1.14, with multimodal systems being rated more favourably. The output explainability target also showed a significant difference, t (37) =7.0, p<.001, with a mean difference of 0.76 (SE = 0.13) and effect size d=1.39, again favouring multimodal systems. Finally, in the system quality target, the Wilcoxon signed-rank test confirmed a significant difference, W=273.0, p<.001, with an effect size rank biserial correlation of 0.82.

While the paired comparisons revealed significant differences in trust between unimodal and multimodal systems within each trust target, these tests did not account for potential interaction effects across targets. To address this, a repeated measures ANOVA was conducted. The analysis revealed a significant main effect of modality, F(1, 37) = 88.96, p<.001, ω^2 = 0.376, confirming the patterns observed in the pairwise tests, where multimodal systems consistently received higher trust ratings across all trust targets. A significant main effect of trust targets was also found, F(2.80, 103.51) = 74.77, p<.001, ω^2 = 0.389, indicating that trust ratings varied significantly across the five targets. Most importantly, a significant modality × trust targets interaction was observed, F(2.96, 109.39) = 27.76, p<.001, ω^2 = 0.127, suggesting that the difference in trust between unimodal and multimodal systems depended on the specific target being evaluated.

These findings are consistent with the trends observed in the descriptive statistics, where multimodal systems generally received higher trust ratings. The reliability of the trust measures, as indicated by Cronbach's alpha values, along with the significant results from both the Wilcoxon signed-rank and paired t-tests, provides strong support for the conclusion that trust ratings for multimodal systems were significantly higher than for uni-

modal systems across all trust targets. The repeated measures ANOVA further confirmed these differences, revealing a significant interaction between modality and trust targets. Together, these results demonstrate a consistent and robust pattern: trust in AI systems is influenced by the data modality of the AI system, with multimodal systems eliciting significantly greater trust. A detailed breakdown of individual item-level comparisons is presented in Table 6.2.

6.4.5 Influence of Optometrists and System Factors on Trust and Confidence in Unimodal and Multimodal AI

6.4.5.1 Effect of Experience on Trust

A Repeated Measures ANOVA was conducted to assess the effect of expertise. The analysis showed no significant impact on trust ratings across most targets except for the system quality. Here, a significant effect of expertise was observed F (1,37) =10.809, p=.002, indicating that trust varied with the level of expertise. However, there was no significant interaction between modality and expertise, suggesting that expertise did not impact trust in unimodal versus multimodal systems.

6.4.5.2 Effect of Gender on Trust

To assess the influence of participant gender on trust ratings in unimodal and multimodal AI systems, a Repeated Measures ANOVA was conducted. Levene's test confirmed that the assumption of homogeneity of variances was not violated for both the unimodal and the multimodal system (p>0.05) across all targets, suggesting that the variances across

Table 6.2: The table shows Shapiro-Wilk (SW) test for data normality and Wilcoxon Signed-Rank (WSR) test for variability analysis in unimodal and multimodal systems across the items of the user's trust questionnaire.

Target	**Items**	**S-W test**		**WSR test**			
		W	***P*-value**	**W**	**z-score**	***P*-value**	**RBC**
IDM	Q1	0.9	0.014	460	4.7	$< .001$	1.0
	Q2	0.9	$< .001$	528	4.9	$< .001$	1.0
	Q3	0.9	0.002	595	5.1	$< .001$	1.0
IQ	Q1	0.9	$< .001$	344	3.7	$< .001$	0.8
	Q2	0.9	0.018	537	4.6	$< .001$	0.9
	Q3	0.9	0.001	223	3.1	0.001	0.8
TA	Q1	0.9	0.005	356	3.5	$< .001$	0.8
	Q2	0.9	0.004	277	3.1	0.002	0.7
	Q3	0.9	$< .001$	520	4.8	$< .001$	1.0
	Q4	0.9	0.004	427	4.5	$< .001$	1.0
OE	Q1	0.9	$< .001$	197	3.4	$< .001$	0.9
	Q2	0.9	$< .001$	234	3.5	$< .001$	0.8
	Q3	0.8	$< .001$	276	4.2	$< .001$	1.0
	Q4	0.9	$< .001$	230	2.8	0.003	0.7
SQ	Q1	0.7	$< .001$	68	2.3	0.019	0.7
	Q2	0.7	$< .001$	50	1.5	0.12	0.5
	Q3	0.6	$< .001$	45	1.8	0.059	0.6
	Q4	0.8	$< .001$	224	3.8	$< .001$	0.9

Abbreviations: RBC (Rank-Biserial Correlation)

groups were not significantly different, and thus the Repeated Measures ANOVA could be appropriately used.

While gender did not independently affect trust across most targets (information quality, output explainability, and system quality), a significant interaction between modality and gender was found in the input data modality target F (2,36)=3.410, p=.044 and the technical accuracy target F(2,36)=4.137, p=0.024. This suggests that gender may influence how trust is perceived in these targets.

These findings highlight the importance of considering both gender and expertise in the design and evaluation of AI systems in clinical settings, ensuring that these systems foster trust among diverse user groups.

6.4.5.3 Effect of AI System Accuracy on Optometrists Confidence

To assess the effect of system accuracy on confidence, we analyzed confidence levels in relation to the accuracy of predictions made by unimodal and multimodal AI systems. After each visit, optometrists were asked to rate their confidence in the system on a Likert scale of 1-5, and these responses were used for the analysis. The results revealed distinct differences between the two systems. Optometrists exhibited higher confidence in the multimodal system when its predictions were correct (M = 4.1, SD = 0.7). Even when the multimodal system's predictions were incorrect, confidence levels remained relatively high (M = 3.5, SD = 1.1), indicating a moderate decrease but still maintaining a strong level of confidence. This suggests that clinicians continued to perceive the system as trustworthy, likely because its use of diverse data sources created an impression of high reliability. In contrast, the unimodal system showed little variation in confidence levels between correct (M = 3.1, SD = 1.2) and incorrect (M = 2.9, SD = 0.9) predictions, reflecting a more

generalized skepticism or lower overall confidence in its decisions.

The significant difference in confidence between correct and incorrect predictions in the multimodal system, as confirmed by the Mann-Whitney U test (W = 1289.5, p=.01), suggest the system's ability to inspire confidence when it performs well, which is less pronounced in the unimodal system. Further analysis revealed the influence of expertise in the multimodal AI system. Novice optometrists (M = 4.1, SD = 0.7) showed confidence levels almost equal to experts (M = 4.1, SD = 0.9) when the system was correct. However, when the system was incorrect, novices exhibited higher confidence ratings (M = 3.7, SD = 0.9) compared to experts (M = 3.1, SD = 1.3). Notably, a significant difference was observed between the confidence ratings of experts for correct versus incorrect decisions (Mann-Whitney U = 198.5, p=.02), whereas novices did not show this difference. This finding suggests that novice optometrists are more prone to be over-trusting the system. Additional statistical details are provided in Appendix Table B.6.

6.4.6 Performance of Optometrists in Clinical Decision-Making using Unimodal and Multimodal AI Systems

The comparison between the unimodal and multimodal systems revealed substantial differences in diagnostic performance. When the AI system was correct, optometrists' accuracy (Eq. 6.1) was 87.5% with the multimodal system, compared to 69.8% with the unimodal system. The precision (Eq. 6.2) of the multimodal system was 76.7%, notably higher than the 58.8% achieved with the unimodal system. In terms of recall (Eq. 6.3), optometrists achieved 72.2% when assisted by the multimodal system, substantially outperforming the 34.5% recall observed when assisted by the unimodal system. The F1 score (Eq. 6.4), which reflects the harmonic mean of precision and recall, was 74.1% for optometrists using

the multimodal system and 32.4% when using the unimodal system.

When the AI system was incorrect, optometrists' accuracy dropped to 54.1% in the unimodal condition and declined sharply to 15.1% in the multimodal condition. Precision (Eq. 6.2) was comparable between the two settings, with 50.5% for the unimodal system and 48.1% for the multimodal system. However, recall (Eq. 6.3) decreased to 36.7% in the unimodal system condition and fell to just 9% in the multimodal system. This resulted in lower F1 scores (Eq. 6.4) of 31.3% and 14%, respectively.

Upon further exploration of cases involving incorrect AI system outputs, we found that in the unimodal-incorrect condition, expert optometrists classified the patient as Suspect in 13/13 cases (100%), while novices classified the patient as Suspect in 20/24 cases (83.3%). In the multimodal-incorrect condition, experts classified the patient as Suspect in 11/13 cases (84.6%), whereas novices classified the patient as Suspect in 12/22 cases (54.5%).

The results suggest that while the multimodal system enhances diagnostic support when accurate, its errors lead to more conservative judgments, particularly among expert optometrists, who often classified patients as Suspect in line with trust calibration theory, whereas novices, potentially influenced by overconfidence, were more likely to make definitive but incorrect decisions.

6.5 Discussion

This study examined how the modality of AI interface design, unimodal versus multimodal, affects clinicians' trust, system acceptance, and diagnostic decision-making in the context of glaucoma diagnosis. The results demonstrate that multimodal AI systems consistently elicit higher levels of trust and system acceptance compared to unimodal systems.

These systems also had a positive influence on optometrists' diagnostic accuracy when the AI's recommendations were correct. These findings suggest that integrating multiple data modalities not only results in calibrated users' trust in AI tools but also better aligns with the complexity of clinical reasoning, thereby supporting more effective human–AI collaboration in diagnostic settings.

6.5.1 Development and Validation of the Trust Instrument

To interpret these findings meaningfully, it is important to reflect on how trust was measured. While we considered several validated instruments such as the Checklist for Trust [Jian et al., 2000b], TOAST [Wojton et al., 2020], Krober's trust scale [Körber, 2019] and the Multi-Dimensional Measure of Trust (MDMT) [Malle and Ullman, 2021], none were optimal for the specific demands of our study. The Checklist for Trust, TOAST, and Körber's trust scale all provide general assessments of trust but do not differentiate between specific system targets. While they may capture overall trust differences between systems, they lack the granularity needed to attribute those differences to functional targets such as input modality, information quality, or output explainability. Specifically, TOAST focuses on broad perceptions of trustworthiness, and Körber's scale, while multidimensional, emphasizes general trust constructs (e.g., reliability, familiarity) that are less suited for capturing dynamic, task-specific trust in AI-assisted decision-making contexts. MDMT, which was developed for human–robot interaction, includes constructs such as moral trust that are not applicable to non-embodied clinical AI systems. Given these practical limitations and the need to assess user trust in distinct system targets, we developed a custom trust questionnaire tailored to our use case. The creation of context-specific trust measures is both common and often necessary when standardized instruments do not

adequately capture the structure or intended use of a system [Kohn et al., 2021].

Our instrument construction is explicitly grounded in the conceptual frameworks, ensuring alignment with both the design features of our AI system and the theoretical dimensions of trust. It is focused on five trust-related targets: IDM, IQ, TA, OE, and SQ. While these targets do not align exclusively with a single category of the Performance–Process–Purpose (PPP) framework [Lee and See, 2004], they can be conceptually situated within this model based on their underlying theoretical rationale and functional role in shaping trust. These constructs are inherently multidimensional, and their influence on users' trust may span across multiple components of the framework. Nevertheless, a reasoned mapping is possible.

IDM and TA are most closely associated with Performance, as they pertain to users' perceptions of the system's competence and reliability. OE and dimensions of IQ (accuracy, and relevance of the information) relate to Process, emphasizing the interpretability and transparency of the system's outputs. SQ aligns with Purpose, reflecting how well the system supports clinical goals and values. However, several constructs may also be interpreted as contributing to multiple dimensions. For instance, IQ may inform both Performance (through the accuracy, relevance) and Process (via clarity and understandability). Similarly, aspects of SQ, such as usability and interface design, may influence perceptions of Performance, while IDM may also bear on Purpose if the data sources reflect alignment with clinicians' information needs. While these associations are not one-to-one mappings, they provide a structured framework for interpreting how different system attributes influence trust, and can further be conceptually related to the broader trust antecedents in organizational trust model [Mayer et al., 1995b].

Along with this conceptual grounding in established theoretical models of trust, the statistical analysis in results section further supported the robustness of our questionnaire,

indicating strong internal consistency and clear factor structure across the five trust targets, thereby validating its utility for capturing meaningful variation in clinician trust across interface conditions.

6.5.2 Interpretation of Trust Targets

The results across all five targets collectively point to the role of system design in fostering clinician trust in AI. The higher rating in target IDM suggests that the integration of multiple data types not only aligns with clinical practice but also facilitates the interpretation of relevant patient information. The observed preference for systems that mirror how clinicians naturally collect and evaluate information indicates that users found these systems more congruent with their cognitive workflow, thereby reducing cognitive load. Cognitive Load Theory (CLT) [Sweller, 1988] states that minimizing the effort required to process information enhances decision-making. Multimodal AI supports this by consolidating complex data into more digestible formats, enabling clinicians to interpret information more efficiently and fostering greater trust. It also establishes the foundation for other targets: richer inputs facilitate clearer outputs, influencing perceived quality, accuracy, and explainability.

Similarly, significant differences in the IQ target indicate that clinicians perceived multimodal systems as more trustworthy. This builds on the input modality advantage: when input is diverse and integrated, the resulting information appears more clinically useful and relevant. Information quality theory [Wang and Strong, 1996] identifies accuracy, completeness, and relevance as central to perceived information value. By providing more complete and relevant information, multimodal systems fostered calibrated trust among optometrists.

The multimodal system was also rated as more effective in communicating the reasoning behind its recommendations (OE target). This reflects a preference not only for accurate outcomes but for systems that make their decision-making process transparent and understandable. In explainability theories, explanation is viewed as a communicative process involving the identification and simplification of relevant aspects of a decision in alignment with the user's goals and knowledge [Miller, 2019b]. Multimodal systems likely support this more effectively because integrated data provides a richer base from which explanations can be constructed. When diverse inputs inform the system's logic, clinicians receive explanations that are more grounded in multiple corroborating sources. This also contributes to higher ratings in IQ, as explanations feel more complete and meaningful. Additionally, when clinicians understand how the AI arrived at its diagnosis, it reinforces their perception of TA. Thus, explainability is not an isolated function, it is supported by the quality and structure of inputs and, in turn, reinforces confidence in the information presented.

The user perception that multimodal systems are more technically accurate, as reflected in the TA target, suggests the cumulative influence of prior targets, DM, IQ, and OE, since both systems demonstrated the same actual accuracy. This indicates that clinicians were not judging accuracy based on output correctness alone but on how comprehensive and cognitively coherent the information felt. Trust in automation is shaped by system transparency, predictability, and alignment with user expectations, factors that make a system seem more reliable even when objective performance is unchanged [Lee and See, 2004].

The observed differences in system quality were primarily related to perceived support, with the multimodal system being viewed as more supportive. This likely stemmed from strengths in prior targets, IDM, IQ, and OE, which contributed to responses that felt more coherent and contextually relevant. As a result, clinicians found the multimodal

system easier to engage with, reinforcing trust through its informational richness rather than interface design differences.

6.5.3 Impact of Modality on Optometrists' Confidence and Performance

Optometrists demonstrated greater diagnostic accuracy and confidence when using the multimodal AI system, particularly when its predictions were accurate. The system's richer, more clinically intuitive inputs likely enhanced clinical reasoning and increased perceived trustworthiness. Notably, even when the AI predictions were incorrect, confidence in the multimodal system remained moderately high, indicating a residual trust in its reliability, while diagnostic performance declined sharply. A deeper analysis revealed the critical role of clinical expertise in shaping both diagnostic performance and trust calibration. Expert optometrists exhibited more calibrated trust, adjusting their confidence and diagnostic decisions more sensitively based on the AI system's accuracy. When confronted with uncertain or incorrect AI outputs, experts often adopted a more conservative diagnostic strategy, frequently categorizing cases as Suspect, in line with trust calibration [Lee and See, 2004] and recent empirical evidence (Pareek et al., 2024).

When faced with uncertain or erroneous AI outputs, experts tended to adopt a more conservative diagnostic strategy, often categorizing cases as Suspect. Rather than placing blind trust in the AI, they calibrated their trust, consistent with the principles of trust calibration [Lee and See, 2004], and recent findings showing that individuals adjust their reliance on AI based on prior observations of its accuracy [Pareek et al., 2024].

In contrast, novices tended to maintain high confidence and frequently issued definitive classifications even when the AI was inaccurate. This pattern suggests an overre-

liance on the system, potentially driven by cognitive biases such as the overconfidence effect, and is consistent with prior findings indicating that less experienced clinicians are more susceptible to automation bias [Nourani et al., 2020] and reduced error awareness [Lee and Chew, 2023]. This discussion illustrates the dual-edged nature of multimodal AI systems: while they can improve diagnostic performance and support appropriate trust among experienced users, they may also unintentionally encourage overtrust among novices. Therefore, although multimodal systems offer considerable promise for enhancing clinical decision-making, their design should be firmly rooted in user-centered principles, with built-in mechanisms to support calibrated trust, particularly to reduce the risk of overreliance by less experienced clinicians.

6.5.4 User Characteristics and Trust

Analyses of expertise and gender suggest that individual differences may influence how users perceive and trust AI systems. Experts are more likely to calibrate their trust in line with system performance, while less experienced users may be more prone to overreliance. Gender related differences in trust emerged in specific contexts, suggesting that trust perceptions may vary and interact with system modality. These findings should be interpreted with particular caution given their exploratory nature and the ethical sensitivity involved in analyzing protected characteristics. Nonetheless, they highlight the importance of designing AI systems that are user-centered and capable of fostering equitable, calibrated trust across diverse populations.

6.5.5 Limitations and Future Research

This study has several limitations, including its controlled research setting, the use of a simulated rather than fully operational AI system, the short duration of user interaction, and the broad comparison limited to only unimodal and multimodal system categories. These factors may influence how the findings translate to routine clinical practice, particularly in terms of trust dynamics and system interaction. Future research should examine AI integration in real clinical environments using fully functional unimodal, multimodal, and hybrid systems, and investigate the development of trust and performance over extended periods through longitudinal studies.

6.6 Conclusion

Trust is a critical factor in successful human-AI collaboration in clinical decision-making. This study explored how optometrists' trust differed between unimodal and multimodal AI systems, revealing that multimodal systems fostered higher calibrated trust and improved decision accuracy. These outcomes highlight the importance of aligning AI design with clinicians' diagnostic reasoning processes. We advocate for the development of AI systems that integrate multimodal data to have calibrated trust on the system and higher performance of optometrist. Future work should build on these findings through real-world implementation and longitudinal evaluation to better understand trust dynamics in clinical settings.

6.7 Acknowledgment

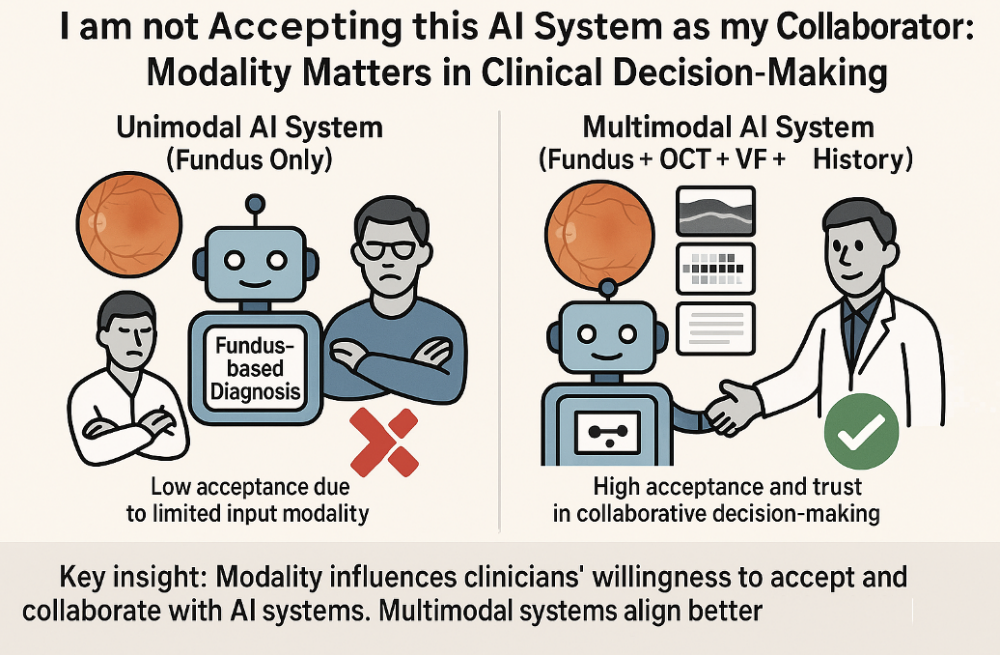
This research benefitted from collaborations with the School of Optometry and Vision Sciences (thanks to Dr. Stanley Woo, Director) and was supported through funding from the Natural Sciences and Engineering Research Council of Canada (NSERC-CRD Grant) (independent of study design/analysis). We are also thankful to our participants who gave us their precious time. FG's research was funded by the University of Waterloo (Graduate Research Studentship and International Doctoral Student Award). We also acknowledge the use of generative AI tools, specifically Grammarly and ChatGPT-4o, for grammar checks and flow of writing. Generative AI tools are not used for any type of data or content generation in this paper.

Chapter 7

Behavioral Determinants of Acceptance of an Artificial intelligent (AI) System: The Impact of Data Modality in human-AI collaboration in Optometry

This chapter is based on the submitted article:

Ghaffar F, Woima R, Furtado N, Burns C. "I am not accepting this AI system as my collaborator": Modality matters. *Translational Engineering in Health and Medicine.*



Graphical Abstract

7.1 Abstract

Background

Artificial intelligence (AI) is transforming healthcare, including disease diagnosis and management. In glaucoma diagnosis, AI systems analyze medical data to identify patterns and provide actionable insights. However, clinician acceptance of AI, particularly regarding unimodal (single data type) versus multimodal (integrated data types) systems, remains underexplored.

Objective

This study aims to investigate how data modality (unimodal vs. multimodal) affects clinician acceptance of AI systems for glaucoma diagnosis. It examines differences in acceptance levels, changes after user interaction with both unimodal and multimodal systems, and resilience to errors in multimodal systems.

Method

Unimodal and multimodal AI mock-up systems tailored for glaucoma diagnosis were developed using actual patient data and designed to reflect real-world decision-making processes. A within-subjects study design was employed, where optometrists assessed both systems. User acceptance was measured using a modified Unified Theory of Acceptance and Use of Technology (UTAUT) model, incorporating five dimensions: performance expectancy (PE), effort expectancy (EE), social influence (SI), facilitating conditions (FC), and a newly introduced dimension, modality preference (MP). In-depth interviews and paired t-tests were conducted to analyze the findings.

Results

Multimodal systems were significantly more accepted than unimodal systems, with a mean user acceptance rating of 1.7 compared to 2.45 ($p < .001$, Cohen's $d = 1.2$). Across all UTAUT dimensions, multimodal systems outperformed unimodal ones, showing statistically significant differences in PE, EE, SI, FC, and MP ($p < .001$). Post-interaction, multimodal systems showed improved acceptance, while unimodal systems experienced

declines in PE and EE. Errors in multimodal systems did not significantly impact user acceptance.

Conclusion

Multimodal AI systems are more acceptable to clinicians due to their enhanced performance and holistic insights, despite increased complexity. These findings show the importance of data modality considerations in AI system design to improve clinician acceptance and smooth integration into clinical workflows.

7.2 Introduction

Artificial intelligence (AI) is transforming disease diagnosis, treatment, and management across various healthcare fields. It is an adaptive technology that utilizes advanced statistical algorithms to analyze both structured and unstructured medical data, with an aim to predict hidden patterns, and future outcomes, and extract clinically relevant actionable information tailored to specific situations [Choudhury et al., 2022], thereby improving clinical accuracy and patient outcomes. One area where AI shows significant promise is in glaucoma diagnosis [Campbell et al., 2020, Li et al., 2018, Qidwai et al., 2023]. Glaucoma is a group of eye conditions characterized by damage to the optic nerve, often caused by increased intraocular pressure [Schuster et al., 2020]. It is a leading cause of irreversible blindness worldwide and can progress silently, with few early symptoms [Shalaby et al., 2022]. In human-AI collaborative healthcare environments, AI provides timely insights that enhance and augment care providers' knowledge and expertise, thereby supporting more informed decision-making [Bhatt and Bae, 2023]. However, the effective

utilization and collaboration with these systems depend on user acceptance, and this acceptance may be significantly influenced by the data modality. Understanding the impact of different data modalities on user acceptance is essential to explore how AI can be successfully integrated into clinical practice.

AI demonstrates significant promise in detecting glaucoma by analyzing either a single type of data (unimodal) or by integrating multiple types of data (multimodal) with various systems developed using these different data types. For example, [Singh et al., 2022] demonstrated that AI can effectively identify glaucomatous damage with high accuracy using a single data type, such as fundus images, while [Xiong et al., 2022] developed AI systems that utilize multimodal data, incorporating both optical coherence tomography and visual field tests. Unimodal AI systems may positively influence clinician acceptance due to their simplicity and minimal training requirements, thereby enhancing usability and decreasing the required effort but their limited data scope may reduce the performance which the users expect as they offer less support for complex clinical decision-making. Conversely, multimodal AI systems, by integrating multiple data types, provide more holistic and richer insights that may increase performance, but their complexity may impact usability due to a steeper learning curve or may lead to high cognitive load because of more complex interfaces.

User acceptance is defined in the Technology Acceptance Model (TAM) by two interconnected components: perceived usefulness and ease of use. Perceived usefulness is the extent to which users think a specific system may enhance their ability to accomplish a task, whereas ease of use is a measure of how simple the user thinks the system is to use [Davis, 1989]. Based on TAM, UTAUT [Venkatesh et al., 2003] and, more recently, the AI Device Use Acceptance model (Gursoy et al., 2019) have been developed to gauge user acceptance of AI. These models have widely been used by researchers for assessing

the acceptance of technology in various fields including healthcare [AlQudah et al., 2021]. For instance, the acceptability and feasibility of the glaucoma easy screener for glaucoma screening in rural areas was evaluated using UTAUT [Nida et al., 2021]. Similarly, another study investigates factors influencing patients' intentions to use telemedicine in ophthalmology during COVID-19, utilizing the TAM [Tsatsos et al., 2024].

Despite extensive research on physician adoption of AI systems [Chen et al., 2022b, Sivaraman et al., 2023], little is known about how various data modalities (unimodal versus multimodal) affect clinician acceptability, particularly in the diagnosis of glaucoma. The main focus of current glaucoma AI research is technical diagnostic efficacy; nevertheless, care provider acceptability and perceptions of unimodal and multimodal systems have not yet been examined and there remains a significant gap. This study aims to fill this gap by investigating the following research questions:

1. Is there a difference in clinicians' acceptance of unimodal and multimodal AI systems? Which system is accepted more?

2. How does clinician acceptance evolve before and after interactions with unimodal and multimodal AI systems?

3. Does user acceptance of multimodal AI systems remain stable after clinicians encounter errors?

   Our study significantly advances the answers to these problems by creating and thoroughly testing unimodal and multimodal AI mock-up systems tailored for glaucoma diagnosis using actual patient data. Working closely with specialists in AI, human factors, and optometry, these systems were designed to make sure they reflect real-world decision-making procedures and meet clinical requirements. Using a

within-subjects design, where each participant assessed both systems, we conducted in-depth interviews with optometrists to measure user acceptance of these systems. User acceptance was measured using a modified UTAUT model, which identifies five critical dimensions: performance expectancy (PE), effort expectancy (EE), social influence (SI), facilitating conditions (FC), and an additional factor, modality preference (MP), that we introduced into the framework in this study. PE specifically relates to people's expectations around the possible productivity benefits and the perceived ability of a technology to improve job performance [Kim et al., 2024]. EE, on the other hand, measures perceived ease of use and represents a person's expected degree of technological comfort [Du et al., 2023]. SI also has a big impact because people are often influenced by what other people think about adopting technology [Park et al., 2020]. FC is the availability of technological and organizational resources [Agyei and Özge Razi, 2022]. Finally, MP refers to the type of data used by AI systems, indicating the user's perception of the system's reliability and effectiveness.

Our findings indicate that multimodal systems are significantly more accepted by optometrists compared to unimodal ones. Specifically, multimodal systems received a mean user acceptance rating of 1.7, while unimodal systems had a higher, less favorable mean rating of 2.45 on a 5-point Likert scale, with 1 representing highest acceptance and 5 representing lowest acceptance. This difference was statistically significant, with a high effect size ($p < .001$, Cohen's $d = 1.2$), as determined by a paired t-test. When examining the five UTAUT dimensions—PE, EE, SI, FC, and MP, multimodal systems consistently outperformed unimodal systems. The mean ratings for multimodal across these dimensions were 1.7, 2.3, 1.3, 2.1, and 2.1, respectively, compared to unimodal ratings of 2.8, 2.8, 1.6, 2.5, and 2.5, with all differences reaching statistical significance ($p < .001$). Additionally, a comparison of

pre and post interaction shows that multimodal systems have higher user acceptance across all dimensions (with three dimensions being significantly high) compared to the pre-interaction phase, while the unimodal system showed a notable decline in PE, a significant increase in EE ratings, and only met initial acceptance levels in other dimensions. Furthermore, AI mistakes in the multimodal system had no observed effect on user acceptance. User acceptance remained consistent before and after interaction, with analysis showing no statistically significant difference.

In this study, we examine how clinicians' acceptance of AI systems is affected by data modalities. We conducted in-depth interviews with optometrists and presented them with different modality AI systems for glaucoma diagnosis and found that they exhibited substantially higher acceptance of the multimodal AI system, surpassing their initial acceptance of AI in general and unimodal system in particular. These findings have important implications for AI system design, emphasizing the necessity of carefully evaluating data modalities to enhance clinician acceptance and facilitate smooth integration into clinical workflows.

7.3 Methodology

7.3.1 Study Design

This study employed a within-subject design to assess user acceptance of glaucoma diagnosis AI systems among optometrists. The study explicitly focuses on how unimodal and multimodal system affect clinicians' acceptance. Two AI system prototypes (mock-up systems) with real patient data were developed to assist and provide a second opinion in the clinical decision-making. Optometrists interacted with both systems, and data were col-

lected through both observational and quantitative methods. In this paper we specifically focus on the quantitative data collected through user acceptance questionnaires.

7.3.2 Ethical Consideration

This study received ethical approval (Application No. 45248) from the Delegated Research Ethics Committee (DERC) at the University of Waterloo on December 4, 2023 (Appendix B.1). We obtained consent from all the participating optometrists about the use of data collected through the interviews (Appendix B.2). They were also briefed thoroughly on the study's purpose, procedures, potential benefits, risks, and their right to withdraw at any time without consequence. After the interview the collected data were anonymized and stored in a protected server. In recognition of their professional contributions, each participant received CAD 150 for their two-hour participation, encouraging detailed engagement with the study. Patients' data collection and selection

We used the same patient data used by (Ghaffar et al., 2024) for the development of AI-mock-up systems. This patient data was obtained from the School of Optometry and Vision Sciences at the University of Waterloo using records from patients who had explicitly consented to the use of their data for research purposes. Out of the collected eight patients' data, two specific patients' data were chosen for mock-up system development. Those cases were selected in which longitudinal data of maximum number of exam or assessment were present. Additionally, the cases needed to be ideal candidates for both unimodal and multimodal AI systems. This allowed for a detailed simulation of clinical interactions with AI system and decision-making processes. The data was limited to two patient cases each with three visits, to balance the depth of review with a manageable workload, as each optometrist required about 10-15 minutes reviewing per visit. This setup provided

sufficient data to assess user acceptance across a total of 12 visits. However, due to time constraints, one visit was later omitted from the unimodal case leaving a total of 10 visits.

7.3.3 AI Mock-up System Development

The AI mock-up system was designed and developed in Figma, using real patient data, providing an interactive interface for clinicians. The selection of data attributes included in the system and provided on the interface was guided by the research studies that explicitly mentioned glaucoma attributes necessary for decision making [Leske et al., 2008, Le et al., 2003, de Voogd et al., 2006, Gordon et al., 2002, Group et al., 2007] as well as through extensive discussion with experts in AI, human factors, interface design and optometry experts.

Two versions of AI system were developed one each for unimodal and multimodal. The unimodal system used only fundus imaging data and multimodal used patient history, clinical information, fundus imaging, OCT and visual field data. The system has multiple sections as shown in Figure 7.1. The AI system provides outputs related to disease diagnosis (the patient has glaucoma, is a glaucoma suspect, or does not have glaucoma), the staging level for diagnosed as glaucoma (early, mild, moderate, or severe) and suspect (high, medium, and low) as shown in section A. The system also provides an explanation for its decisions, a conclusion, and recommended actions as shown in section B. The system provides a tab for viewing patient information such as histories, fundus, visual field as shown in section. The section D displays the information related to each tab.

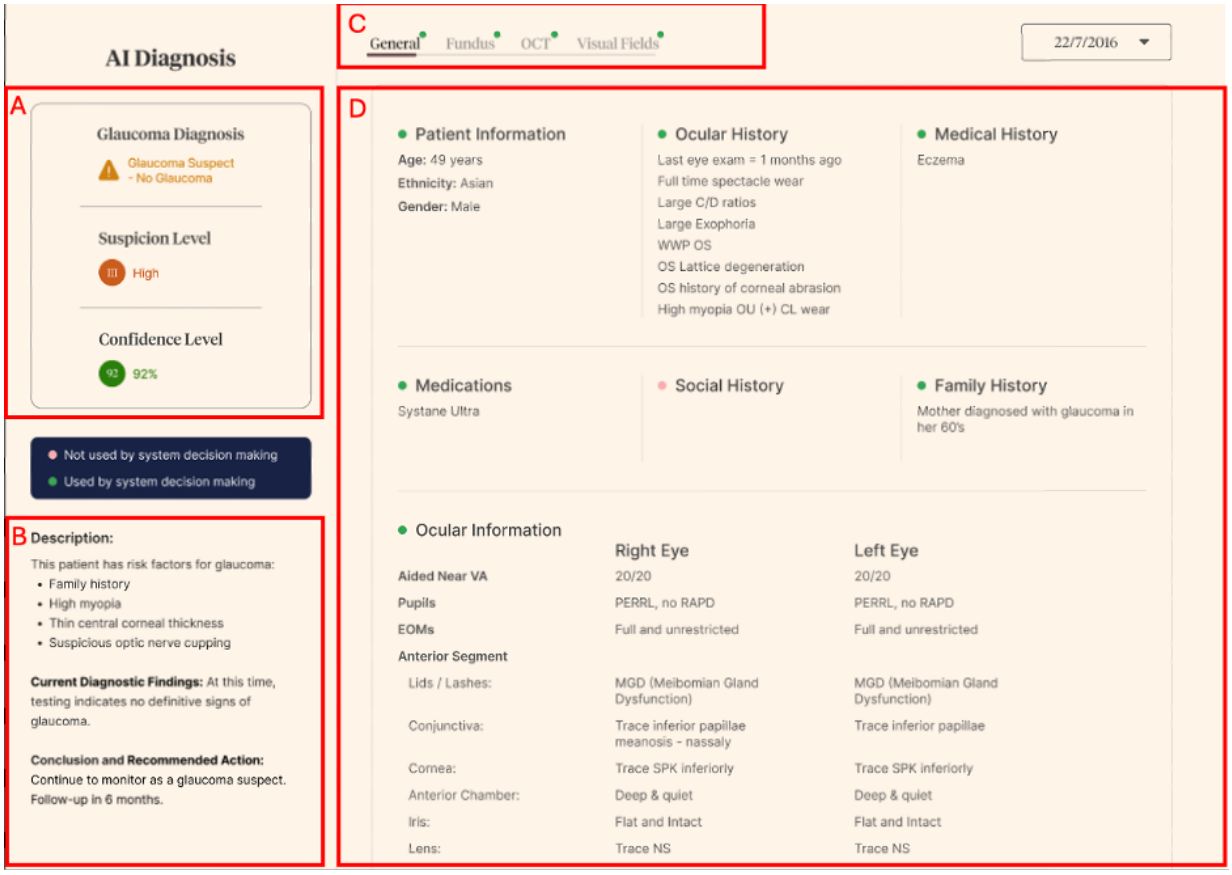


Figure 7.1: Screenshot of proposed mock interface used. Section A highlights the diagnosis and systems confidence, Section B provides details on the decision shown in Section A. Section C is where different data types can be selected for review, and Section D is the main display for whichever section is under review

Both systems have differences in the fundus imaging, OCT and visual field tabs where the unimodal provide local explanation of OD and OS for fundus imaging only while the multimodal provides local explanation for OD and OS of all fundus imaging, OCT and also visual field. A side-by-side view as shown in Figure 7.2. In the unimodal system, the fundus imaging OD and OS decisions combines and generates the global AI output shown in section A and B of Figure B while in multimodal system, the global output is generated

by combining OD and OS and of fundus, OCT and visual field.

7.3.4 Optometrist Interviews

Twenty optometrists were contacted and interviewed; the majority were connected to the School of Optometry and Vision Sciences at the University of Waterloo. Using Zoom and Microsoft Teams, each optometrist took part in a 100–120-minute online interview in English. During the interviews, participants reviewed patient records using the "think aloud" protocol and consulted the AI systems to aid in their decision-making.

Both unimodal and multimodal systems were shown to each participant; to avoid bias, the unimodal system was shown first, followed by the multimodal system. Because of the initial encounter, this sequence and structure made sure that participants did not grow to prefer the multimodal system. To measure the general acceptance of AI systems, the participants were requested to complete a questionnaire prior to the cases being presented. Following that, each case was shown to them one visit at a time. At the end of each visit review, participants were asked if they agreed or disagreed with the AI diagnosis, staging, and explanation decisions. On completion of each case review, participants were then asked to fill modified UTAUT questionnaire (provided in Appendix B.3) for measuring the user acceptance of AI systems. The questionnaire had 18 questions, classified into five dimensions. The lowest value "1" on the Likert scale represented very strong agreement and the highest value "5" represented disagreement. Participants were asked to complete an additional questionnaire in a multimodal case prior to encountering an intentionally introduced incorrect AI evaluation. This approach allowed us to measure user acceptance both before and after experiencing an AI error, providing insights into how user acceptance might shift in response to system mistakes. Participants were remunerated 150 CAD for

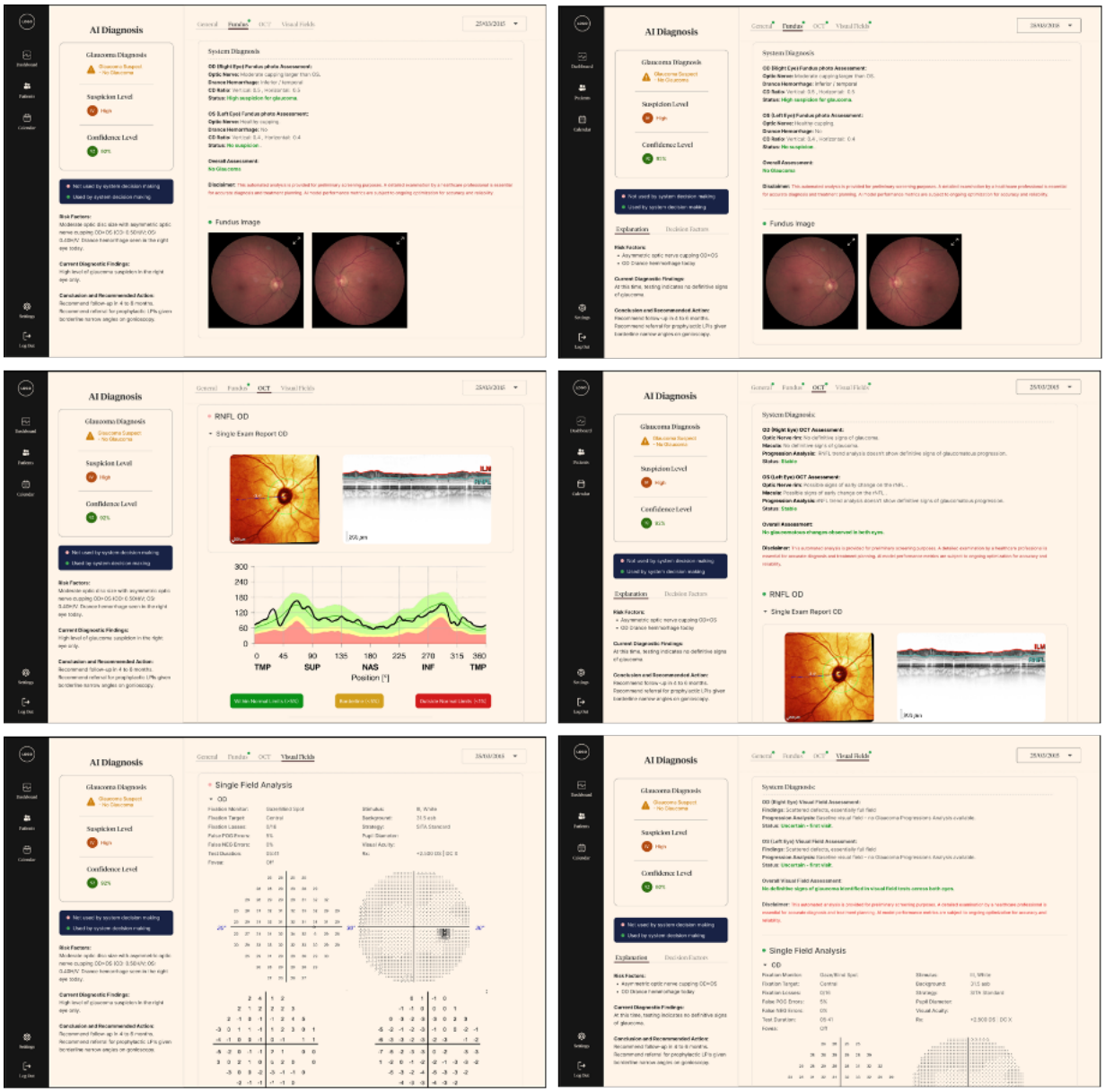


Figure 7.2: Different interface views for the data types available: Fundus imaging, OCT and visual fields. The left side is from a unimodal mock-up and the right side is images from a multimodal mock-up

their two-hour commitment, recognizing their contribution.

7.3.5 Data Analysis

We conducted our statistical analysis using the open-source software JASP, beginning with a descriptive summary of the demographics of the participating optometrists. Key demographic variables included gender, ethnicity, age, years of professional experience, weekly patient count, and weekly glaucoma patient count, with the mean, median, and standard deviation calculated for each.

To analyze user acceptance of an AI system, we first computed descriptive statistics (mean, median, standard deviation, confidence interval, and interquartile range) for each of the five dimensions of acceptance without dividing the data into unimodal or multimodal. This analysis provided a detailed overview of the data, highlighting the distribution of responses, central tendencies, and variability. We also created visual plots to illustrate these metrics, facilitating a preliminary examination of trends in the collected responses.

To assess user acceptance of unimodal and multimodal AI systems across dimensions, we began by evaluating the reliability and validity of the dimensions. Cronbach's alpha with a 95% confidence interval (CI) was calculated to evaluate internal consistency, and Intraclass Correlation Coefficients (ICC) with 95% CI were computed to assess reliability across raters. Bartlett's Test of Sphericity was used to verify the suitability of our data for factor analysis, where we examined eigenvalues and factor loadings (range) to confirm the dimensionality and strength of associations for each factor.

After establishing reliability, we calculated descriptive statistics for each system and employed the Shapiro-Wilk test to assess the normality of the data distributions, which is essential for determining the appropriate statistical tests to use. Given that our data

did not meet the assumptions for normality, we opted for the Wilcoxon paired test (non-parametric test) to examine differences between the unimodal and multimodal systems across dimensions. To track changes in user acceptance over time, we compared pre-interaction responses with post-interaction responses for both conditions using a repeated-measures ANOVA, allowing us to analyze the within-subject effects of the interactions over time. Finally, we assessed user acceptance before and after encountering an intentional AI error by plotting response distributions and again using the Wilcoxon paired test to determine statistical significance, ensuring our analysis remained consistent with the non-parametric approach suitable for our data characteristics.

7.4 Results

7.4.1 Demographic Characteristics of Optometrists

A total of 20 optometrists were recruited for this study. The gender distribution of the group reflected a reasonable balance, with 40% identifying as male, 50% as female, and 10% as transgender or non-binary. In terms of ethnicity, 90% of the optometrists were either of Caucasian or Asian (South, East, or West) descent, equally divided between these two categories. The remaining 10% comprised individuals of African (5%) and Middle Eastern (5%) heritage. Participants' ages spanned a wide range: 45% fell between the ages of 20 and 30, 25% between 30 and 40, 15% between 40 and 50, and 5% between 50 and 60. An additional 10% chose not to disclose their age. Figure 7.3 provides a graphical representation of these demographic variables, offering further insight into the composition of the participant group.

The overall average length of experience in the field was 6.61 years (SD = 8.08), illus-

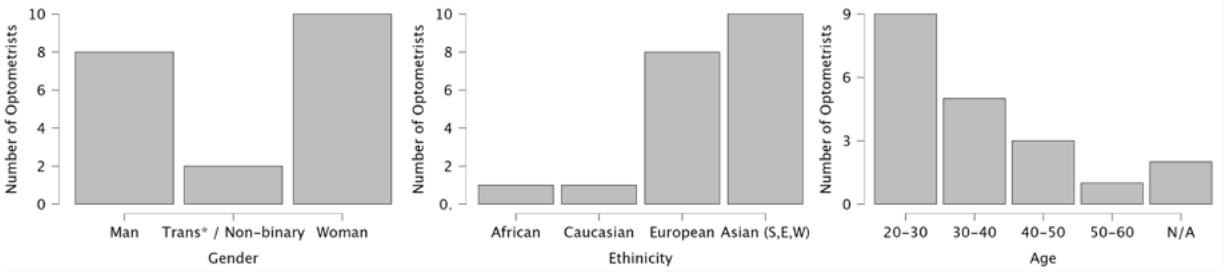


Figure 7.3: Demographic information (gender, ethnicity and age) of recruited optometrists

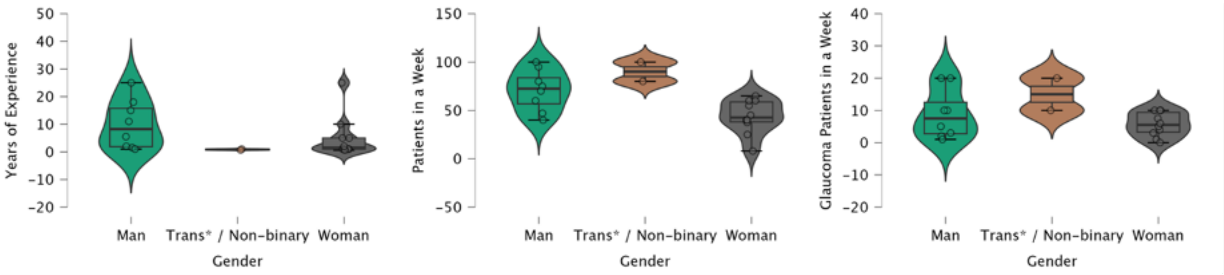


Figure 7.4: Distribution of year of experience (left), patients seen in a week (middle), and glaucoma patients seen in a week (right) in relation to gender

trating a broad range of professional backgrounds among participants. On a weekly basis, the optometrists reported seeing an average of 59.12 patients (SD = 24.72). Of these, approximately 7.88 were glaucoma patients (SD = 6.24), with a median value of 6.75. A gender-wise distribution of these experience metrics is shown in Figure 7.4.

7.4.2 Descriptive Statistics of User Acceptance Dimensions

In our analysis of user acceptance, PE dimension was slightly higher, with a mean of 2.6 (SD = 0.9, CI = 2.4–2.8) and a median of 2.3 (IQR 1.8–3.1), reflecting that optometrists had moderate confidence in the system's ability to enhance diagnostic accuracy and efficiency. The EE dimension yielded a lower mean of 1.5 (SD = 0.6, CI = 1.4–1.6), and

a median of 1.3 (IQR 1–2), suggesting that participants found the system relatively easy to use, requiring less effort. This aligns with the strong belief that the system would not impose a significant cognitive or operational load. For SI , the mean was 2.3 (SD = 0.7, CI = 2.1–2.4) with a median of 2.0 (IQR 1.6–2.6), indicating moderate peer and supervisory encouragement towards system adoption. This shows that while optometrists are somewhat influenced by their professional community, the effect is neither highly positive nor negative. Finally, FC had a mean of 2.3 (SD = 0.9, CI = 2.1–2.5) and a median of 2.6 (IQR 1.6–3.0), reflecting moderate confidence that the necessary resources and support structures would be in place for successful system implementation. We did not include the modality preference dimension in this analysis as their items were unique for unimodal and multimodal systems and couldn't be analyzed as an average together. The visualized distributions of each dimension, as shown in Figure 7.5, suggest that most of the items exhibit either uniform distribution or a peak in the middle range showing a moderate level of confidence in system performance, and social support for adoption. However, in the effort expectancy dimension, responses were more concentrated in the lower range (equating to highest acceptability), suggesting a strong belief among participants that minimal effort would be required to effectively use the system.

7.4.3 Variation in User Acceptance of Unimodal vs Multimodal system

Before conducting comparative analyses, we assessed the reliability and validity of the user acceptance responses to ensure that each group of questions effectively represented a single construct in both unimodal and multimodal conditions. As summarized in Table 7.1. Cronbach's alpha values showed strong internal consistency across most dimensions, with

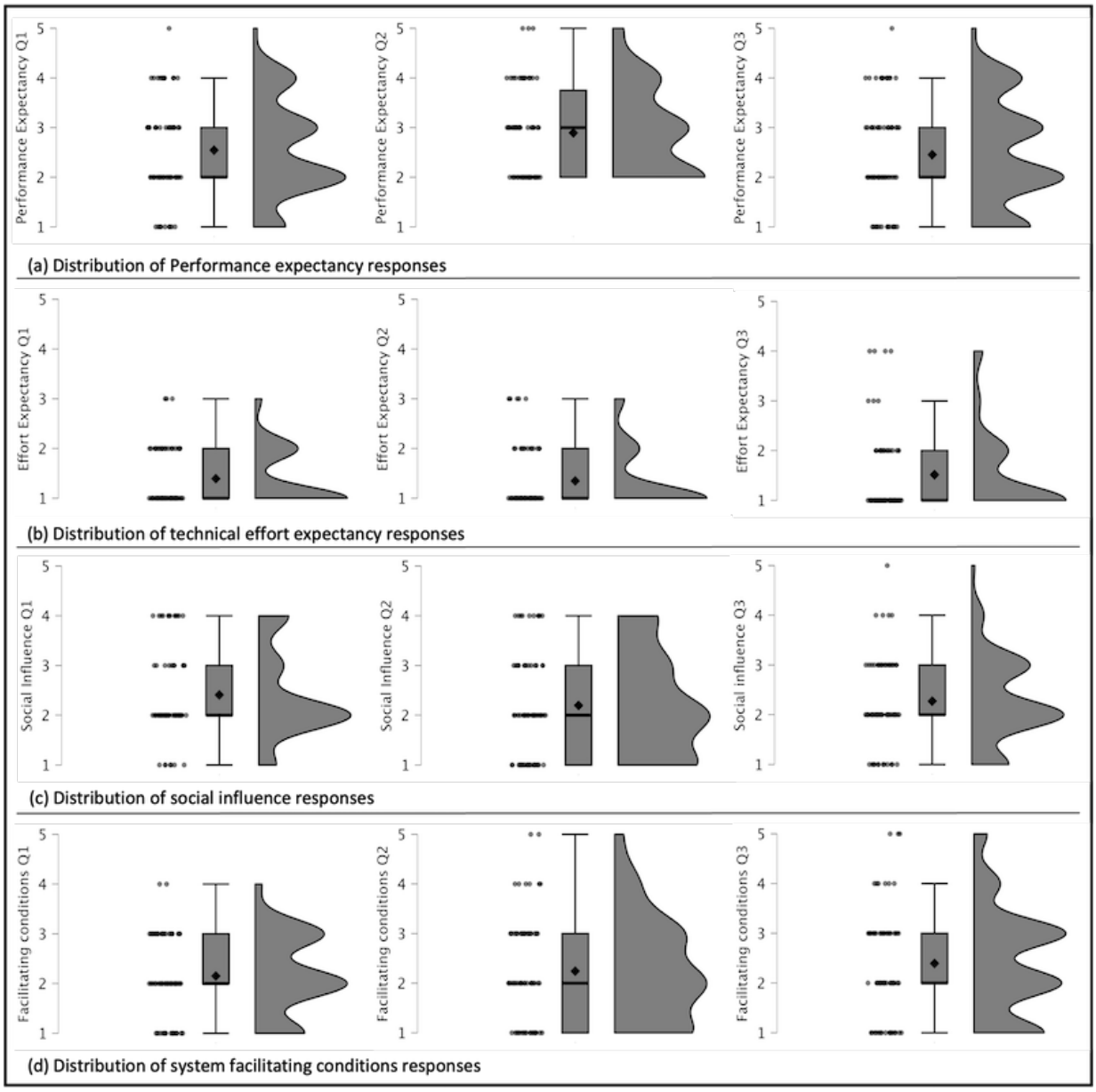


Figure 7.5: Distribution of responses against each item of all dimensions of user acceptance except modality preference.

values exceeding 0.7 for PE, EE, and MP. Although SI and FC had lower alpha values, all dimensions had eigenvalues greater than 1 in the Exploratory Factor Analysis (EFA), confirming the one-dimensionality of the constructs. Bartlett's Test of Sphericity was significant for all dimensions, further supporting the appropriateness of factor analysis. These results demonstrate that the items within each dimension reliably measure their intended constructs, validating the use of these dimensions for further analysis.

Table 7.1: Summary of reliability tests for Uni-modal and Multi-modal Systems for all dimensions of user acceptance

Dimension/ Construct	**Tests**	**Uni-modal**	**Multi-modal**
MP	α (95% CI)	0.76 (0.57-0.86)	0.76 (0.59-0.89)
	ICC (95% CI)	0.64 (0.2-0.77)	0.52 (0.31-0.69)
	Bartlett's Test	$\chi^2(3) = 27.91$, $p < 0.001$	$\chi^2(3) = 30.54$, $p < 0.001$
	Eigenvalue	2.01	2.05
	Factor loading (range)	0.6-0.91	0.58-0.88
PE	α (95% CI)	0.89 (0.81-0.94)	0.9 (0.82-0.94)
	ICC (95% CI)	0.73 (0.61-0.83)	0.75 (0.62-0.84)
	Bartlett's Test	$\chi^2(3) = 63.44$, $p < 0.001$	$\chi^2(3) = 69.51$, $p < 0.001$
	Eigenvalue	2.47	2.51
	Factor loading (range)	0.81-0.89	0.83-0.93
EE	α (95% CI)	0.86 (0.77-0.92)	0.84 (0.74-0.91)
	ICC (95% CI)	0.69 (0.56-0.80)	0.65 (0.45-0.80)
	Bartlett's Test	$\chi^2(3) = 55.84$, $p < 0.001$	$\chi^2(3) = 49.78$, $p < 0.001$
	Eigenvalue	2.39	2.31
	Factor loading (range)	0.74-0.88	0.69-0.87
SI	α (95% CI)	0.71 (0.51-0.84)	0.46 (0.05-0.70)
	ICC (95% CI)	0.46 (0.26-0.64)	0.26 (0.05-0.50)
	Bartlett's Test	$\chi^2(3) = 23.16$, $p < 0.001$	$\chi^2(3) = 12.43$, $p = 0.006$
	Eigenvalue	1.924	1.566
	Factor loading (range)	0.52-0.8	0.48-1.0
FC	α (95% CI)	0.82 (0.7-0.90)	0.68 (0.49-0.81)
	ICC (95% CI)	0.60 (0.45-0.75)	0.40 (0.20-0.61)
	Bartlett's Test	$\chi^2(3) = 54.64$, $p < 0.001$	$\chi^2(3) = 52.95$, $p < 0.001$
	Eigenvalue	2.22	2.12
	Factor loading (range)	0.56-1.01	0.53-1.03

7.4.3.1 Descriptive Statistics

We compared the optometrists' perceptions of unimodal and multimodal systems across different dimensions, such as modality preference, PE, EE, SI, and FC. The analysis shows that the multimodal system is viewed more favourably. Across all these categories, the multimodal system shows lower mean, median, and standard deviation values, indicating that optometrists found it more reliable in improving performance, rated it as high performing, easier to integrate, and less effort intensive. The responses also suggest that the multimodal system is perceived to have better social acceptance within their professional environment and is seen as having more available resources and institutional support. In contrast, the unimodal system tended to score higher, suggesting it may present more challenges in these areas. The detailed statistics presented in the Table 7.2 shows the overall preference for multimodal systems, highlighting their perceived ease of use and greater compatibility with the working conditions and expectations of optometrists.

Table 7.2: Descriptive statistic of user acceptance of unimodal and multimodal AI system across all dimensions

Dimension/Construct	System	Mean (95% CI) ± SD	Median (IQR 25%-75%)
MP	Multimodal	1.74 (1.5–2.0) ± 0.56	1.66 (1.3–2.3)
	Unimodal	2.81 (2.4–3.1) ± 0.88	2.67 (2.3–3.6)
PE	Multimodal	1.68 (2.1–2.6) ± 0.67	2.33 (1.6–2.8)
	Unimodal	3.01 (2.7–3.3) ± 0.90	3.0 (2.3–3.6)
EE	Multimodal	1.39 (1.2–1.5) ± 0.50	1.0 (1.0–1.6)
	Unimodal	1.67 (1.4–1.9) ± 0.87	1.67 (1.0–2.1)
SI	Multimodal	2.11 (1.9–2.3) ± 0.59	2.0 (1.6–2.6)
	Unimodal	2.51 (2.2–2.7) ± 0.87	2.33 (1.8–3.0)
FC	Multimodal	2.11 (1.8–2.4) ± 0.75	2.0 (1.5–3.0)
	Unimodal	2.53 (2.2–2.8) ± 0.88	2.33 (1.8–3.2)

7.4.3.2 Significance of Variation

The mean and median values indicate that the multimodal system consistently outperforms the unimodal system across all dimensions. However, to determine whether these observed differences are statistically significant, we conducted both the Shapiro-Wilk test and the Wilcoxon signed-rank test. The Shapiro-Wilk test was used to assess whether the differences between the unimodal and multimodal systems followed a normal distribution. It revealed that while MP and PE followed a normal distribution, the other dimensions EE, SI, and FC, deviated from normality.

Given the non-normal distribution in some dimensions, the Wilcoxon signed-rank test was applied to evaluate whether the differences in the median responses between the two systems were significant. The results showed significant differences across all five dimensions, confirming that the superior performance of the multimodal system was statistically significant ($p < 0.05$ for each dimension).

Table 7.3 provides precise p-values and other relevant metrics to justify the significance of variation. Additionally, the distribution of responses for each dimension is visualized in Figure 7.6. It shows a more intuitive understanding of how the responses were distributed between the two systems, complementing the numerical data from the table.

These analysis with lower mean and median scores with statistically significant differences in all these dimensions shows user acceptance of multimodal AI system is higher than unimodal system in clinical practice in general and optometry in specific. In the system adoption dimension, stronger acceptance is shown, meaning optometrists found the multimodal system to be more advantageous over traditional or any other diagnostic methods and the data it uses align with their mental model. The PE dimension shows that optometrists have higher confidence in multimodal AI system effectiveness and show

Table 7.3: Significance of variation of user acceptance of unimodal and multimodal system across all dimensions

Dimension/Construct	Shapiro-Wilk Test		Wilcoxon Signed-Rank Test			Rank-Biserial Correlation
	W	p	W	z	p	
Adoption Modality	0.95	0.27	311	4.00	<.001	0.91
Performance Expectancy	0.98	0.80	536	4.07	<.001	0.80
Effort Expectancy	0.94	0.07	282.5	3.23	0.001	0.74
Social Influence	0.85	<.001	307.5	2.85	0.005	0.63
Facilitating Conditions	0.90	0.002	314	3.52	<.001	0.79

skepticism in using the unimodal system in clinical practice. They believe that a multimodal system could improve their diagnostic accuracy, produce reliable results, and save time and effort more than unimodal. The effort expectancy dimension indicates that optometrists found the multimodal AI system more user-friendly, less challenging, and less difficult even though it produces more results and explanations than the unimodal. The social influence dimension shows that optometrists believe that the multimodal system will be more accepted as they would more likely recommend it to their peers in the professional community for diagnostic purposes than the unimodal system. The facilitating condition for the multimodal system indicates greater confidence that the necessary infrastructure is in place or could be arranged to support the use of multimodal AI system in practice. The statistical differences for each item across all dimensions are presented in tabular format and visualized as distributions in Appendix B.6.

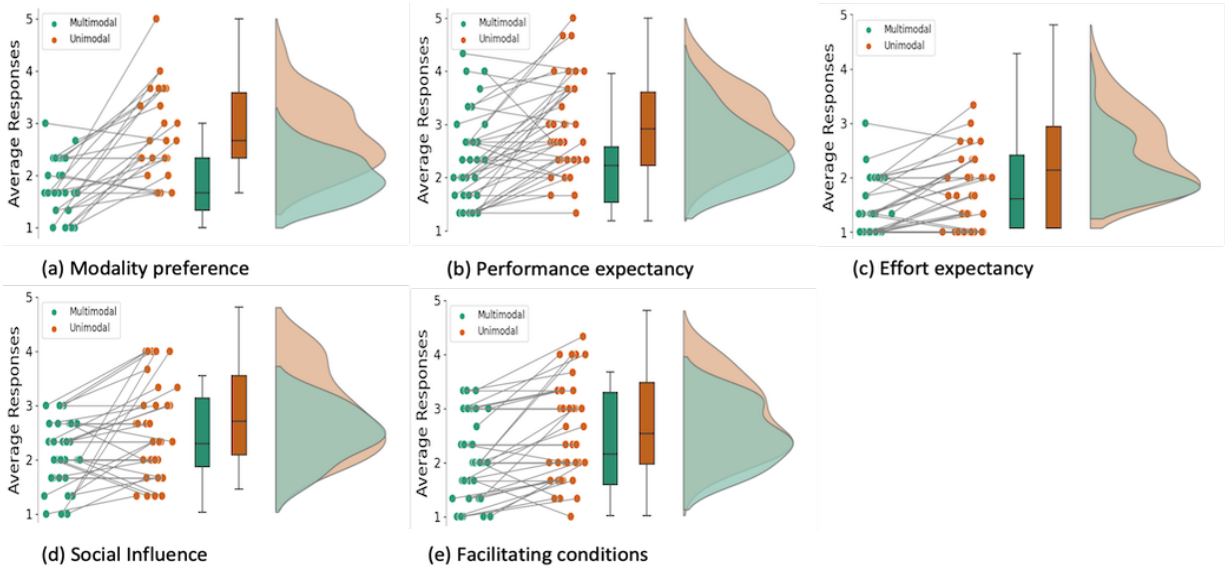


Figure 7.6: Distribution of responses by optometrist averaged across dimensions for both uni-modal and multi-modal systems

7.4.4 Progressive Changes in User Acceptance: Pre-AI System Interaction (PAI), Post-Unimodal AI system Interaction (PUI), and Post-Multimodal AI System Interaction (PMI)

Progressive changes in user acceptance were observed across three stages: Pre-AI system Interaction (PAI), Post-Unimodal AI system Interaction Post-Unimodal AI system Interaction (PUI), and Post-Multimodal AI system Interaction Post-multimodal AI system Interaction (PMI). This analysis was carried out for all dimensions except for the MP dimension, which was not measured at the PAI stage due to the study's design. The MP items were specific to the unimodal and multimodal systems, they were only relevant after user interaction with these systems. The descriptive statistics were first calculated for all dimensions of user acceptance at the three mentioned stages. To confirm the varia-

tions and changes in user acceptance at different stages, a repeated measures ANOVA was conducted, revealing significant differences across the stages.

The users initially had higher PE ratings at PAI stage which decreased at the PUI stage. However, at the PMI stage, the PE returned closer to PAI stage. The mean user's EE at the PAI stage were lowest which then increased at the PUI stage, and increased even further at PMI stage. The user's mean social influence rating was similar at PAI and PUI stage, with a little increase at PMI stage. The FC rating is also at same level at PAI and PUI stage with a decrease at PMI. The detailed statistics are given in Table 7.4

Table 7.4: The table shows comparative descriptive statistics of user acceptance at pre and post interaction stages across all dimensions

Dimension/Construct	**Mean (95% CI) ± SD**	**Median (IQR 25%-75%)**
	PAI	
PE	2.3 (1.9–2.7) ± 0.7	2.3 (1.7–2.7)
EE	2.2 (1.9–2.5) ± 0.7	2.0 (1.7–2.7)
SI	2.5 (2.1–2.9) ± 0.9	2.5 (1.9–3.0)
FC	2.6 (2.1–3.0) ± 1.0	2.3 (2.0–3.1)
	PUI	
PE	3.2 (2.8–3.7) ± 0.9	3.0 (2.7–4.0)
EE	1.6 (1.3–2.0) ± 0.7	1.7 (1.0–2.0)
SI	2.6 (2.2–3.0) ± 0.9	2.3 (2.0–3.2)
FC	2.5 (2.1–3.0) ± 1.0	2.3 (1.8–3.2)
	PMI	
PE	2.4 (2.0–2.9) ± 0.9	2.3 (1.7–3.0)
EE	1.4 (1.1–1.6) ± 0.5	1.0 (1.0–1.7)
SI	2.2 (1.7–2.5) ± 0.6	2.1 (1.7–2.7)
FC	2.1 (1.6–2.4) ± 0.8	2.0 (1.3–3.0)

The repeated measures ANOVA test indicates significant changes in users' acceptance across the three stages (PAI, PUI, and PMI) for three of the four dimensions as shown in Table 7.5 with its post hoc analysis shown in Figure 7.6. Specifically, the PE dimension showed a significant difference ($p = 0.003$), with post-hoc Bonferroni correction tests revealing significant differences ($p = 0.007$) in PAI and PUI, with PE at PUI stage dropping below initial expectancy. No significant difference was observed between PAI and PMI, suggesting that the multimodal AI system met users' initial PE. EE also varied significantly ($p<.001$) with post-hoc test Bonferroni correction showed that the variation is significant ($p<.05$) between PAI and PMI and between PUI and PMI stages. EE at PUI and PMI exceeded PAI, indicating that users found both of the systems easier to use than anticipated. Although SI changes were not significant, a trend towards increased recognition of multimodal systems within the community was observed. Finally, FC varied significantly ($p = .005$), with post-hoc test Bonferroni correction showing that the variation is significant ($P<0.05$) between PAI and PMI and between PUI and PMI stages. The facilitating conditions rating is also better for PMI than PAI and PUI. Figure 7.7 provide a visual representation of these trends and changes in users' perceptions across the three stages.

These findings indicate that user acceptance evolved progressively before interaction, after unimodal interaction, and after multimodal interaction, across three of the measured dimensions. It highlights modalities influence on user perceptions and acceptance, offering insights for improving future AI systems.

Table 7.5: ANOVA Results

Dimension/Construct	F (p-value)	ω^2
PE	7.1 (0.003)	0.1
EE	13.9 (<0.001)	0.2
SI	2.1 (0.1)	0.02
FC	6.7 (0.005)	0.05

Table 7.6: Bonferroni Post Hoc Comparisons (t, p-value)

Dimension/Construct	Before vs After Unimodal	Before vs After Multimodal	After Unimodal vs After Multimodal
PE	-3.6 (0.007)	-0.2 (1.0)	3.5 (0.007)
EE	3.0 (0.02)	4.6 (<0.001)	2.6 (0.05)
SI	-	-	-
FC	0.4 (1.0)	4.0 (0.002)	3.1 (0.01)

7.4.5 Sustained User Acceptance Despite AI System Errors in Multimodal AI Systems

The analysis demonstrates that user acceptance of the multimodal AI system remained consistent, even when users encountered a situation where the AI made an error. As illustrated in Figure 7.8 the distribution of responses shows only negligible variation before and after the error, highlighting the system's resilience in terms of user acceptance. The results of the Wilcoxon paired t-test further substantiate this finding, indicating no statistically significant differences across all measured dimensions of acceptance. Even when examining individual items within each dimension, no meaningful change was observed in user acceptance following the AI error. These results suggest that the benefits of the multimodal system are sufficiently compelling to uphold user confidence and acceptance, even in the face of inaccuracies. Users appear to prioritize the system's overall utility and smooth integration into their workflow, suggesting that isolated errors do not undermine their willingness to collaborate with or adopt the technology. This resilience in acceptance

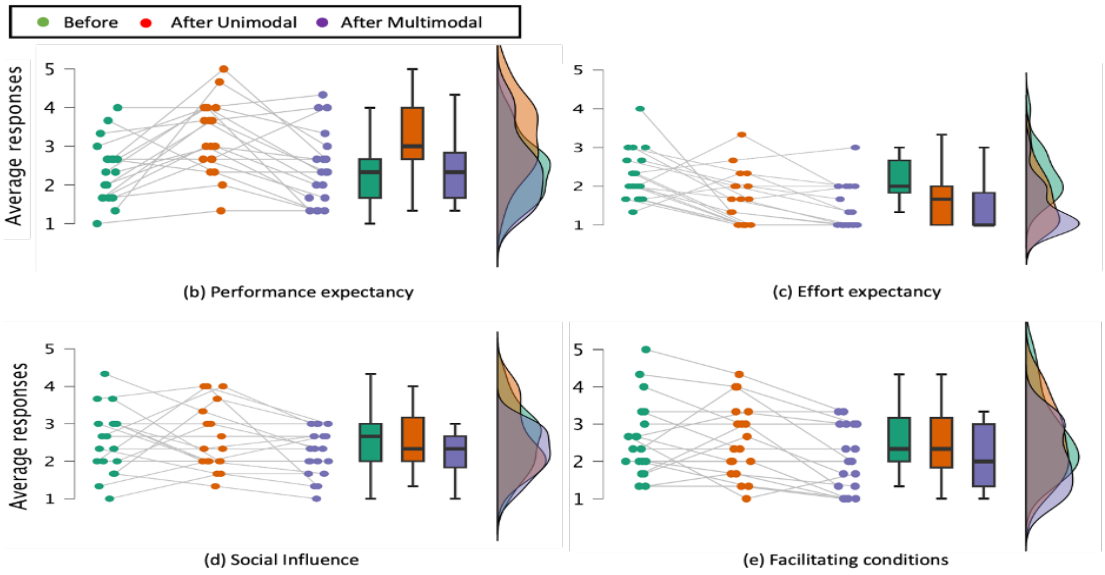


Figure 7.7: The figure shows distribution of acceptance ratings from optometrists from before interacting with the AI system (green), after interacting with the unimodal system (red) and after interacting with the multimodal system (purple) for four dimensions.

shows the multimodal AI system's potential for use in clinical settings.

MP measures how well the system integrates into the diagnostic process and its advantages over traditional methods. PE examines the belief that the system can improve diagnostic accuracy, produce reliable results, and save time. EE measures how easy the system is to use. SI reflects the likelihood of recommending the system to colleagues and the expected support from supervisors. Finally, FC assesses whether optometrists feel there are sufficient resources and organizational support for system adoption.

7.4.6 Conclusion

In this work the effect of data modalities on the acceptance of AI systems was explored through the use of a modified UTAUT framework. We designed both unimodal and mul-

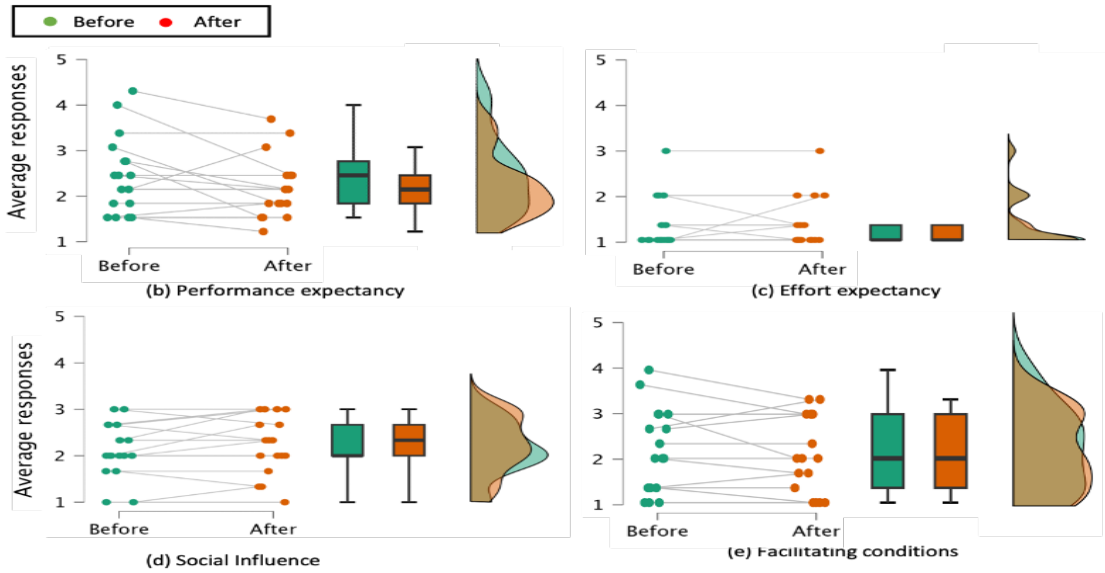


Figure 7.8: Distribution of acceptance ratings from optometrists before error introduction into the system (green) and after (red) for all four dimensions.

timodal mock-up AI systems with real patient data for glaucoma diagnosis and recruited optometrists to evaluate these systems. It was found that a multimodal AI system for glaucoma diagnosis and management had highest acceptance among optometrists. Future research should focus on the development and training of an accurate multimodal AI system for glaucoma detection for use in optometry practice. Furthermore, the adoption of a multimodal system into clinical workflows should be assessed through continued user-centred design practices and collaboration with optometrists, ensuring translation of user acceptance findings in this study to a real-world setting.

Chapter 8

Discussion and Conclusion

This chapter provides a comprehensive discussion of the key findings of this research, their implications, contributions, limitations, and directions for future work. It concludes with a reflection on the overall significance of the study and its contributions to the field of AI and human factors in healthcare specifically in optometry.

8.1 Discussion and Interpretation of Key Findings

This section provides a discussion and interpretation of the key findings from each chapter, with a focus on how they relate to the research questions. Each subsection corresponds to a specific chapter and addresses the related research question outlined in Section1.6.

8.1.1 Variability in Decision-Making

In chapter 3 We first investigated how optometrists' experience levels affect their decision-making in glaucoma diagnosis, a crucial factor for patient outcomes to answer the research

question 1 which is further divided into sub-questions mentioned in section 3.2. The results showed significant differences in diagnostic approaches between novices and experts. The discussion ahead will dive deeper into these differences in decision-making, particularly focusing on how experience levels affect priorities in data interpretation and heuristic usage in diagnosis and will correlate our findings in section 3.4 with the research questions in section 3.2 of chapter 3.

8.1.1.1 Decision Making Processes

Our findings in Chapter 3 suggest a clear distinction in decision-making behavior between expert and novice optometrists, particularly in situations where prior patient data was unavailable. Experts demonstrated significantly higher alignment with original clinical decisions, even when working with incomplete information. This suggests that clinical expertise enables more accurate, intuitive judgments based on subtle pattern recognition and accumulated experience.

This observation is supported by Hammond's Cognitive Continuum Theory (CCT) [Hammond and R, 1981], which describes decision making as existing along a spectrum from highly analytical to highly intuitive reasoning. Analytical reasoning is step-by-step and often used by novices or in unfamiliar situations. In contrast, intuitive reasoning is relying on pattern recognition and past experience. Experienced clinicians tend to move more fluidly along this continuum, adjusting their reasoning style based on the situation [McCartney, 2017, Yang et al., 2012]. This supports our findings: novices relied on analytical reasoning and improved with detailed patient information, while experts showed cognitive flexibility, shifting between intuitive and analytical approaches based on the situation.

To further understand these variations, we can draw on Rasmussen's Skills-Rules-Knowledge (SRK) framework [Rasmussen, 1983] and the Naturalistic Decision Making (NDM) approach [Klein, 1993].

The SRK model categorizes decision-making into three levels: skill-based reasoning (automatic and intuitive), rule-based reasoning (based on familiar procedures or guidelines), and knowledge-based reasoning (analytical and used when facing novel or complex problems)[Rasmussen, 1983]. These levels of reasoning connect directly to our glaucoma decision ladder (Figure2.2) and help explain how experts and novices approached diagnosis when patient information was limited.

Novice optometrists, who have limited clinical experience, typically operated at the knowledge-based level. They follow the decision ladder in a step-by-step and linear manner to reach a diagnosis, progressing from data observation to system state interpretation, goal selection, and action formulation. This analytical approach reflects their reliance on explicit reasoning at each stage of the decision process. Novices tended to interpret clinical findings as signals—raw, disconnected pieces of information that lacked immediate clinical meaning. Without prior patient history, they had difficulty transforming these signals into meaningful patterns, which reduced their confidence in diagnosis and often led to divergence from the original clinical decision.

In contrast, expert optometrists were able to operate at the rule-based level, drawing on internalized procedures and prior experience with similar cases. Their greater familiarity with glaucoma presentations allowed them to interpret clinical findings not as isolated signals, but as signs (recognizable patterns) or even as symbols, which represent deeper, abstract clinical meaning. This ability enables experts to take shortcuts on the decision ladder, skipping intermediate steps such as conscious state evaluation or goal formation. Instead, they moved quickly from initial observations to action selection, guided by pattern

recognition and heuristics. Even in the absence of prior patient history, these clinicians could make decisions that closely aligned with the original clinical assessment.

These findings also align with the broader principles of the Naturalistic Decision Making (NDM) framework, which emphasizes how experienced professionals make effective decisions in real-world environments that are often uncertain, time constrained, and information limited [Klein et al., 1989, Klein, 1993]. Unlike classical models that assume decision-makers systematically compare multiple options, NDM suggests that experts rely heavily on experience, context, and pattern recognition to guide their reasoning. Their decisions are shaped by familiarity with similar situations rather than formal analysis. In our findings, expert optometrists appeared to reflect this naturalistic approach. Even when patient history was incomplete, they could draw on their prior clinical experience to recognize meaningful patterns and make confident judgments. Their ability to operate effectively under such constraints contrasts with that of novice optometrists, who were more dependent on structured and complete data to guide their reasoning. This supports the idea that clinical expertise involves not only what clinicians know but also how they apply it under real-world conditions.

Lastly, when prior patient history was provided, the performance gap between experts and novices narrowed. This shows the role of contextual information in clinical reasoning. For novices, access to detailed patient history supported a more accurate and structured diagnostic process, helping compensate for their lack of pattern recognition and experiential shortcuts.

Together, these theoretical models, CCT, SRK, and NDM, provide a comprehensive framework for understanding the cognitive processes that differentiate novice and expert decision-making.

8.1.1.2 Data Prioritization in Clinical Assessments

The study in chapter 3 also explored how optometrists at different experience levels prioritize data during clinical assessments. Initially, both experts and novices engaged comprehensively across various diagnostic areas, but with a lower focus on progression, change, and correlation analysis. This trend reflects common practice for initial assessments, where establishing a broad understanding of the patient's condition is prioritized over detailed analysis. Interestingly, subsequent visits revealed a significant increase in attention to progression and change, particularly among experts. Experts demonstrated a more comprehensive approach, integrating clinical findings and patient history from the outset and emphasizing structural-function correlation more than novices. This comprehensive risk assessment aligns with the notion that experienced clinicians can synthesize complex information more effectively [McCartney, 2017] and that is why their concordance was quite high than novices.

Novices, despite a higher initial engagement progression and change analysis, might misunderstand the optimal timing for such analyses, typically requiring follow-up data. Although this increased for novices in subsequent visits, the experts showed a robust increase almost double that of novices. This indicates a strategic approach to deeper analysis in the later stages of patient management by experts.

8.1.1.3 Heuristic Approaches and Cognitive Efficiency

Our results in chapter 3 suggest that experienced optometrists may employ heuristic approaches more frequently, particularly when data is limited, which aligns with Dual-Process Theory [Kahneman, 2011]. This theory explains that experts often rely on System 1 thinking—fast, intuitive, and heuristic-driven rather than System 2, which is slower and more

analytical. Experienced optometrists likely rely on System 1 thinking because their years of practice have exposed them to many cases, allowing them to recognize patterns quickly and make rapid judgments without consciously analyzing every detail. This heuristic-driven approach is particularly useful when data is incomplete, as they can fill in gaps based on past experience rather than engaging in slower, effortful reasoning. In contrast, less experienced optometrists may depend more on System 2, carefully analyzing each case step by step due to their lack of extensive pattern recognition.

The nuanced internalization of a patient's historical data by experts, evidenced by their subtle differences in conducting longitudinal comparisons and intercase comparisons, also suggests cognitive efficiency. This efficiency allows experts to retain and integrate key details of disease progression into their clinical judgments without frequent referrals to past records or references to other cases. This cognitive agility is indicative of their evolved diagnostic expertise and pattern recognition in glaucomatous progression [Tsang et al., 2023].

The reliance on external case references by novices, as noted in some cases, may be a compensatory mechanism for their relative inexperience. This behavior highlights the gap in their approach to long-term patient management and the development of pattern recognition skills.

8.1.2 Decision Factors Variations

Diving deeper into the variations of experts and novices, we further investigated research question 2 that is related to the diagnostic factors and staging decisions made by novice and expert optometrists when assessing glaucoma in chapter 4. We found significant differences in diagnostic approaches, confidence and the difference in factors and reasons that influence their decision. In this section I will further discuss and correlate findings in section 4.4

with the research questions in section 4.2 of chapter 4.

8.1.2.1 Differences in Factors that Influence the Decisions of Experts and Novices.

In our study in chapter 4, in all three diagnostic decisions where the patient is classified is having glaucoma, suspect or not having glaucoma, experts displayed a more holistic and comprehensive approach, integrating their extensive clinical experience and intuition. According to Cognitive Load Theory (CLT) [Sweller, 1988], experts have automated many aspects of their diagnostic process, reducing their cognitive load. This allows them to quickly recognize patterns and consider a wider range of factors, including subtle signs, without feeling overwhelmed. In contrast, novices experience a higher cognitive load because they must consciously analyze each piece of information. Since their working memory is occupied with processing basic details, they struggle to integrate multiple factors at once. This difference explains why experts take a more holistic and intuitive approach, while novices rely on a slower, more analytical process that may overlook subtle signs.

Experts considered a broader range of factors, including subtle signs. The novices were focusing on only a subset of data in majority of the cases. This finding convey that novices are greatly influenced by only subset of examinations to take decision while experts take into account the majority of the factors before taking a decision. Further we have seen that in cases where glaucoma was diagnosed or was classified as suspect the experts provided the disease progression and change analysis as justification quite higher almost double to that of novices. In the realm of medical decision-making, these findings are supported by literature suggesting that experienced clinicians often rely on a 'pattern recognition' approach, whereas less experienced clinicians lean towards a more algorithmic, data-intensive approach [Caddick et al., 2023].

The information filtration processes of experts for decision making from the all the information they have acquired while reviewing the record is much efficient than novices. Previous literature on the cognitive strategies employed by expert problem solvers also discusses the differences in the efficiency of both groups. According to [Schunk, 2012] rather than marshalling vast arrays of information that might bear on a problem, skilled individuals pinpoint essential features, align them with their extensive background knowledge, and conceive a limited number of viable solutions. When analyzing patient records, optometrists streamline the decision-making process by simplifying complex information. They achieve this by discarding extraneous data and focusing solely on critical relevant details. The difference in deterministic and exclusion reasons further employs that decisions of each group is influenced by different factors.

These analyses provide information that both experts and novices decisions are influenced by different factors. Their decisions are not only influenced by the available factor but also by employing different cognitive strategies.

8.1.2.2 Balancing Positive (in Favor of Glaucoma) and Negative (Against Glaucoma) Factors/Reasons

Our analysis of the justification's ratios in chapter 4 also showed distinct patterns in justifying their decisions with experts providing very less reason overall and also when break it down to positive and negative when looking at overall reasons. Specifically, in glaucoma diagnosed cases, experts provide significantly fewer justifications against glaucoma compared to novices and more justifications supporting a glaucoma diagnosis. This disparity suggests that experts are more confident in their decisions than novices. Similarly in suspected cases, experts averaged less justifications against and in favor of glaucoma. Moreover, in the staging stage, we observed that experts balance positive and negative reasons when

staging glaucoma, with an increase in positive reasons and a decrease in negative reasons as the disease progresses. While novices initially show a similar trend, their reasoning becomes more plateaued in the later stages for both positive and negative reasons. This indicates that experts have a better understanding of glaucoma staging and can clearly define stages based on decision factors, whereas novices struggle with this aspect.

These analysis reveals distinct differences in how novices and experts balance positive (in favor of glaucoma) and negative (against glaucoma) factors when staging their decisions. Therefore, experts are better at dynamically integrating decision factors to accurately stage glaucoma, while novices lack the nuanced understanding to do so effectively.

Furthermore, the clustered decisions of experts and the bell-shaped curve decisions of novices, resonate with the Dreyfus model of skill acquisition (Dreyfus and Dreyfus, 1980). This model suggests a progression from rigid adherence to rules (novices) to an intuitive approach based on deep experience (experts). Our findings, showing that experts prioritize efficiency and depth by focusing on critical data points (after filtration), align with the model's characterization of expert behavior as intuitive and holistic. Similarly, novices' breadth-focused approach, considering a wider array of factors, reflects their developmental stage of relying more on guidelines and external factors, consistent with the early stages of the Dreyfus model. This is evident in our study, where experts use progression analysis in their diagnostic process, indicative of an intuitive, pattern-based approach, while novices' consistent focus on a broader range of variables aligns with a more analytical, rule-based method at the end prove that priorities are different. This unequivocally demonstrate differing priorities in their diagnostic processes. Experts tend to prioritize efficiency and depth, focusing on critical data points and using a more intuitive approach. In contrast, novices display a preference for a broader range of variables, adhering to a more analytical and rule-based method. These contrasting approaches suggest a fundamental difference in

how optometrists at different stages of their careers prioritize and interpret clinical data in glaucoma diagnosis, aligning with established theories in skill acquisition and cognitive psychology.

8.1.3 Human-Centric AI System Guideline to Support Variations

In this section, we propose a set of guidelines for developing an AI system for glaucoma diagnosis based on our findings and to answer research question 3. These guidelines are intended for a human-AI collaborative environment, where the AI system will be used as a decision support tool rather than a standalone decision maker. We suggest that by following these guidelines, the AI system will effectively support optometrists with different experience levels, bridging the gap between novices and experts, and ultimately improving patient outcomes through enhanced diagnostic accuracy. The collaborative model will also uphold ethical and professional standards, mitigating any legal, ethical, and professional barriers posed by standalone system. As a result, it will facilitate effective integration of AI into clinical practice.

8.1.3.1 Minimizing the Effect of Limited Data Availability

As our findings show variability in diagnostic concordance between novice and expert optometrists when comprehensive patient histories are not available. Our findings reveal variability in diagnostic concordance between novice and expert optometrists when comprehensive patient histories are unavailable. Experts, leveraging heuristic processing [Kahneman, 2011] and pattern recognition through mental schemas [Klein, 2017], efficiently extract key diagnostic cues from incomplete data. This reduces cognitive load

and allows them to maintain higher concordance. In contrast, novices rely on systematic processing [Chaiken, 1980] and rule-based reasoning, which increases cognitive load and makes it difficult to prioritize critical information, leading to lower diagnostic accuracy.

When comprehensive data is available, novices experience reduced cognitive load, enabling them to apply analytical reasoning more effectively, thereby narrowing the gap with experts. Experts, in turn, refine their judgments by cross-referencing the additional data, mitigating the effects of confirmation bias and further improving their diagnostic accuracy.

These findings suggest that AI systems should have a robust module that provides detailed patient histories. If a patient has limited historical data, the system should present novices with two features. It should identify and display data from the most similar patient cases with the same conditions [Parimbelli et al., 2018] along with the decisions made in that case. It should also highlight the dissimilarities between the selected patient and the current patient. The AI system should be trained with expert decision-making patterns by modelling the diagnostic approaches and visualizing the variables by estimating the missing data [Berchuck et al., 2019] backward in time to give them an idea of how history would look. These two features would minimize the limited data availability challenge and allow novices to make more accurate and consistent decisions in glaucoma diagnosis.

8.1.3.2 Evolving AI System for Initial and Subsequent Visits

Our result shows that in the initial visit, the expert optometrists adopt a comprehensive approach to assess all factors, except they don't carry analytical analysis (any progression and change analysis, and structure and function correlation analysis). As they move to subsequent visits, they still follow a comprehensive approach but with limited focus on patient historical information, instead, their focus shifts to analytical assessment. We

accordingly suggest that an AI system should behave in a similar way and evolve along with the visits. In the initial visits, the AI system should prompt a novice to adopt a comprehensive approach and highlight important factors such as family history, age etc. In the subsequent visits, the AI system should highlight any analytical changes. It would be better that the AI system present a hierarchy of features, by demonstrating which features require more or less importance to assess during both the initial visit and subsequent follow-up visits. These features of AI system will help novices in understanding the optimal time for any analysis.

8.1.3.3 Follow-up Time Suggestion

The follow-up time analysis show variation among optometrists. We suggest that for better patient management the AI system should provide an optimal time frame for follow-up visits [Bian et al., 2020] which should be based on patient data and clinical indicators. Along with follow-up interval, the system should also display the variables that influence this decision, such as disease progression, patient history, and symptoms. With this feature we can bring more consistency in follow up time suggestion by optometrists.

8.1.3.4 Enable Efficient Data Filtration and Prioritization

Design the AI to efficiently filter and prioritize critical diagnostic factors, mimicking experts' ability to discard irrelevant data and focus on essential details. This will streamline the decision-making process, especially for complex patient records. Moreover, the AI system should present decision factors in a hierarchical manner, prioritizing the most influential factors at the top [Bozorgmehr and Weltermann, 2023]. This approach ensures that users can quickly grasp the key elements influencing the decision and would be bet-

ter able to detect and correct for its sizable mistakes [Poursabzi-Sangdeh et al., 2021b]. Novices can benefit from a clear structure that highlights critical diagnostic indicators, while experts can dive deeper into the nuanced details of less critical factors. For instance, prioritizing factors such as intraocular pressure (IOP), retinal nerve fiber layer (RNFL) thickness, and visual field test results can help streamline the decision-making process.

8.1.3.5 Balance Positive and Negative Justifications Dynamically

The AI system should provide bipolar decision support system support both novices and experts by providing detailed explanations and highlighting positive (in favor of glaucoma) and negative (against glaucoma) factors separately and their interaction [Natarajan and Augustin, 2024]. This setup helps novices understand the weightage and balance of positive and negative factors necessary for making a diagnosis (glaucoma, no glaucoma, or suspect) and determining the stage (early, mild, moderate, or severe). By learning how these factors interact and contribute to different diagnostic outcomes, novices can develop a deeper understanding of the decision-making process. It validates experts' clinical judgment by showing a comprehensive and dynamic interaction of factors, allowing them to confirm their assessments and refine their diagnostic accuracy. This approach ensures a balanced, transparent, and educational diagnosis process, enhancing accuracy and confidence for all users.

8.1.3.6 Educational Module

Lastly, incorporating a training and educational component within the AI system increase learning efficiency and can effectively bridge the knowledge gap between novice and expert optometrists [Dai et al., 2024]. For novices, interactive learning modules such as case stud-

ies, simulated diagnoses can help them understand the weightage and balance of positive and negative factors necessary for making decisions and determining. For instance, the system could allow users to dynamically adjust positive and negative factors to observe how these changes affect decisions and staging. This comprehensive educational support not only enhances individual competence but also promotes overall improvements in diagnostic consistency and accuracy.

8.1.3.7 General AI system Features and Behavior

Most of the AI systems developed by researchers for glaucoma diagnosis focus on detection of glaucoma with binary outputs [Hemelings et al., 2020, Saxena et al., 2020, Lin et al., 2022] . However, to gain optometrists' trust the AI system developed for glaucoma should also include a confidence score, a detailed explanation of how it arrived at that specific decision and should provide a visualization of the risk factors in a hierarchical manner [Kamal et al., 2022a, Chayan et al., 2022]. This will ensure that the optometrist does not overlook any important details. Additionally, the AI system should have an interactive interface that offers an adjustable variables feature [Beauxis-Aussalet et al., 2021] that allows the optometrist to modify the values of any factor and see how different scenarios would affect the outcomes, such as diagnosis, severity, and follow-up time. For example, if the optometrist changes a variable to reflect a worsening condition, the AI system will adjust the follow-up recommendation accordingly. Such an AI system design for glaucoma will ensure more consistent decision making regardless of expertise level.

8.1.4 Effect of AI System Data Modality on Trust and Implications for AI System Design

In chapter 6, I examined the impact of AI system data modalities on user trust, addressing the research question 4. This primary question was further divided into three sub-questions, as outlined in Section 6.2. The findings revealed significant differences in user trust when interacting with unimodal and multimodal systems. Additionally, demographic factors and levels of expertise influenced certain dimensions of trust. Overall, this study demonstrates that users' trust in AI systems is significantly enhanced with the adoption of multimodal architectures. The findings show the critical role of integrating diverse data modalities in AI development and emphasize key design considerations for fostering trust and promoting adoption. By systematically addressing trust dimensions, multimodal AI systems have the potential to bridge the gap between technological innovation and practical implementation, ultimately transforming decision-making processes.

The discussion ahead dive deeper into these differences into effect of data modalities on user trust and will correlate the findings in Section 6.4 with research question in Section 6.2 of chapter 6.

8.1.4.1 Difference in User Trust on Unimodal and Multimodal AI System

The findings in 6 and 7 provide evidence that clinicians' trust in AI systems differs significantly between unimodal and multimodal systems. Across all trust dimensions examined, including input modality, information quality, technical accuracy, output explainability, and system quality, multimodal systems consistently outperformed unimodal systems in eliciting higher trust ratings. This consistent pattern shows the enhanced perception of reliability, accuracy, and transparency associated with multimodal systems, contributing

to their broader acceptance in clinical settings.

A key result emerged in the input modality dimension, where multimodal systems demonstrated a significant advantage, likely due to their ability to integrate diverse data sources, similar to how clinicians approach diagnoses. CLT [Sweller, 1988] explains that reducing the effort required to process information improves decision-making. Multimodal AI achieves this by consolidating complex data into a more digestible format, allowing clinicians to verify and interpret information more efficiently, thereby increasing trust.

Similarly, significant differences in the information quality and technical accuracy dimensions highlight the perceived richness, precision, and reliability of multimodal systems. Lee and See's model of trust in automation [Lee and See, 2004] suggests that trust is built on perceived system reliability and predictability. Since multimodal AI synthesizes multiple data types, it offers more corroborative evidence, reducing uncertainty and reinforcing clinicians' confidence in the system's accuracy.

The output explainability dimension further shows the importance of transparency, as multimodal systems provide clearer and more actionable insights. Explainable AI principles suggest that AI explanations must be interpretable, actionable, and aligned with user expertise. Multimodal AI enhances explainability by presenting diverse data formats allowing clinicians to cross-check information intuitively. In contrast, unimodal AI often presents outputs in a single format, making validation more difficult. CLT [Sweller, 1988] explains that complex or ambiguous outputs increase mental effort, reducing trust, whereas multimodal AI reduces this burden by synthesizing information in a structured, clinician-friendly manner. Additionally, Cognitive Fit Theory [Vessey, 1991] posits that decision-making improves when information presentation aligns with users' analytical approaches. Multimodal AI supports this by integrating complementary data sources, providing a clearer, context-rich foundation for clinical judgments.

In the system quality dimension, the Wilcoxon signed-rank test also confirmed a significant advantage for multimodal systems, though variability in perceptions of system quality suggests the need for further investigation into factors such as user interface design and response times.

From a clinical perspective, the results highlight the potential of multimodal AI to improve diagnostic workflows for glaucoma and assists the optometrist in decision making. Given the complexity of glaucoma diagnosis, which often requires the synthesis of structural and functional data, the use of multimodal systems aligns with the practical demands of clinical decision-making. Moreover, the higher trust ratings for multimodal systems suggest that clinicians may be more likely to rely on these tools in practice, ultimately leading to better patient outcomes.

However, the study also raises questions about the factors that influence trust in unimodal systems. While they performed significantly worse in all dimensions, understanding their limitations could inform efforts to improve their design and utility in specific contexts where multimodal systems may not be feasible.

8.1.4.2 Demographics Effect on User's Trust and AI Design

The findings in 6 and 7 provide important insights into how expertise and gender influence trust in unimodal and multimodal AI systems. Expertise was found to significantly impact trust in the system quality dimension, indicating that individuals with varying levels of expertise evaluate system quality differently. However, this effect was not influenced by system modality, suggesting that expertise affects trust uniformly across both unimodal and multimodal systems. In contrast, gender demonstrated a more specific influence, with significant interactions between modality and trust observed in the input data modality

and technical accuracy dimensions. These results imply that gender differences may shape how users evaluate certain aspects of system functionality, particularly related to data inputs and accuracy.

These findings highlight the importance of considering individual characteristics such as expertise and demographic factors like gender when designing AI systems. Addressing these differences can help create systems that are perceived as trustworthy and reliable by a diverse range of users, thereby enhancing their overall acceptance and usability.

8.1.4.3 User's has Higher Performance While Interacting with Multimodal System

The users higher performance while interacting with multimodal highlights its ability to mimic clinicians' natural decision-making processes. By synthesizing information across modalities, it supports optometrists in addressing cognitive gaps, ensuring more comprehensive evaluations. For AI design, this shows the importance of prioritizing context-rich, multimodal architectures to enhance decision alignment. The improved alignment with optometrists' mental models builds trust, encouraging reliance on the system. Unlike unimodal systems, multimodal AI reflects clinicians' reasoning processes, fostering confidence. Designers should focus on creating systems that mirror domain-specific workflows and present information intuitively to reduce cognitive strain.

The similarity in specificity between unimodal and multimodal systems highlights a need for better mechanisms to identify disagreements. Designers must enhance systems to detect and communicate areas of uncertainty, using features like confidence scores and explainable insights to ensure critical evaluation rather than blind trust.

Multimodal systems have the potential to reduce cognitive load and improve workflow

efficiency. However, to achieve smooth integration, they must present actionable, concise insights. Designing interfaces that prioritize key data points while preserving clinician autonomy will encourage collaboration and better workflow compatibility.

8.1.5 Effect of AI System Data Modality on User's Acceptance

In chapter 7, I explored the effect of AI system data modalities on user acceptance to answer the research question 5. This main research question was further broken down into 3 research question mentioned in chapter 7. To answer these research questions first the overall acceptability of an AI system was evaluated, followed by differences in user acceptance depending on the type of data used by the system. Further, the evolution of acceptability was assessed before and after interacting with the AI systems. Finally, the impact of errors in the AI system on acceptability was explored.

8.1.5.1 Overall Acceptability of AI for Glaucoma Diagnosis

When assessing the acceptability of an AI system for use in glaucoma diagnosis in general, user acceptance was highest for effort expectation, suggesting that optometrists did not encounter issues in learning the proposed system. It further suggests less effort is required for using this system, indicating there is no disturbance in their workflow. There was fair support for the other dimensions evaluated: performance expectation, social influence and facilitating condition. Indicating a fairly supportive opinion among optometrists regarding the expected productivity of the system, what others in the field would think about this system as well as the availability of resources to implement and use this technology. The modality preference dimension was not applicable for overall acceptance analysis as the questions were specific to each type of system. A survey study assessing optometrist

perspectives of AI in retinal disease found similar findings for measures that would translate closely to perspectives of performance expectancy and social influence in this study [Ho et al., 2022]. However, no finding from their study translates well to the facilitating conditions dimension in our study and it should be noted that their study was of a different design than this present work and further was focused more on user perspectives and not acceptability of AI.

8.1.5.2 User's Acceptance of Multimodal is Higher Than Unimodal AI Systems

When looking closer at the type of data used in a proposed AI model, this study found a preference for a multimodal system. This system encompasses fundus imaging, OCT, visual field, clinical data and more, compared to the unimodal model proposed, which only utilized fundus imaging. Previous literature in glaucoma diagnosis is mainly focussed on enhancing accuracy using either unimodal or multimodal AI systems [Fu et al., 2018, Lim et al., 2022]. A generalized study on retinal disease detection claim that unimodal AI systems may be more relevant to the scope of optometry practices as a screening tool, and multimodal systems would serve best for advanced ophthalmology care [AlShawabkeh et al., 2024]. However, our findings indicate that from an optometrist acceptability standpoint, a multimodal system would be more accepted.

This preference can be explained through CLT [Sweller, 1988], human-centered design principles, task-technology fit theory (TTF) [Goodhue and Thompson, 1995], and risk perception in decision-making. From a CLT perspective, while multimodal systems may initially appear more cognitively demanding due to their complexity, they ultimately reduce uncertainty by consolidating diverse diagnostic inputs into a unified, interpretable format, enhancing diagnostic confidence. Human centered design principles further support this

preference, as multimodal AI systems are designed to align with the natural diagnostic workflow of optometrists, making them feel more intuitive and professionally relevant.

TTF theory emphasizes that technology is most effective when it aligns with the demands of a specific task. Given that diagnosing complex conditions like glaucoma requires integrating multiple data points—such as fundus imaging, OCT, and clinical data—multimodal systems inherently provide a better fit for clinical decision-making than unimodal alternatives. Risk perception also plays a crucial role; optometrists may view unimodal systems as riskier due to their limited diagnostic scope, increasing the potential for misdiagnosis or missed findings. Multimodal AI mitigates this risk by offering a more holistic dataset, reinforcing optometrists' confidence in AI-assisted decision-making and aligning with the risk-averse nature of healthcare practice.

We therefore argue that implementing a multimodal system at the general optometrist practice stage would provide access to more comprehensive diagnostic systems. This would assist their diagnostic decision making, increasing the optometrist acceptance of an AI system. To the best of our knowledge, there is no direct comparison of user acceptance ratings for unimodal and multimodal AI systems for glaucoma in the literature.

8.1.5.3 Progressive Changes to User's Acceptance

Comparing acceptability ratings before and after interaction with AI systems showed a significant difference in user acceptance for three of the four dimensions assessed. We saw a significant decrease in acceptance between PAI and PUI for performance expectancy, however acceptance between PAI and PMI were similar. This suggests that the multimodal AI system is more aligned with the user's initial performance expectations of an AI system, in comparison to the unimodal system. The observed decrease in acceptance after

interacting with unimodal AI, compared to the stable acceptance of multimodal AI, can be best explained by Expectation-Confirmation Theory Expectation-Confirmation Theory (ECT) [Oliver, 1980, Bhattacherjee, 2001]. According to ECT, user acceptance depends on whether a system meets, exceeds, or falls short of initial expectations. If users anticipate AI to provide comprehensive, high-quality decision support, then a unimodal system, which relies on limited input sources, may feel insufficient, leading to a drop in performance expectancy. The mismatch between expectations and actual system capability results in disconfirmation, reducing user confidence. In contrast, multimodal AI confirms user expectations by integrating multiple data points, delivering a richer, more aligned diagnostic tool. This reduces disappointment, reinforces system credibility, and sustains acceptance levels

In the effort expectations dimension, the user acceptance increased after interacting with both unimodal and multimodal systems. This provides insight into perspectives of the users and their acceptance of the proposed interface design used in the mock AI systems for this study. There was no significant difference between PAI and PUI for facilitating conditions but there was an increase in acceptance between PAI and PMI for this dimension. This suggests their optimism in organizational support and willingness to provide resources to adopt the multimodal system. Lastly, there was no significant difference in user acceptance for the social influence dimension. This suggests that data modality has no effect on changes in peer and colleague influences, however we cannot exclude the possibility of the effects of other factors not explored in this study on this dimension.

8.1.5.4 Error May Not Impact User Acceptance

Interestingly, this study found that optometrists' acceptance remained at a comparable level for the multimodal system, even after error was introduced. An infallible AI system cannot be achieved [Kaas and Habli, 2024], similar to hypothetical perpetual motion machines. This suggests that the system should be designed so that occasional mistake does not impact user acceptance. In this study, we found that acceptability of the multimodal AI system remained unaffected by error.

Optometrists may exhibit automation bias, meaning they tend to over-rely on AI systems even when errors occur. If they perceive the system as generally accurate, they may continue to trust its outputs despite occasional mistakes. This aligns with trust calibration, where users adjust their trust in a system based on its performance over time. If the multimodal AI system consistently provides useful recommendations, occasional errors may not significantly reduce acceptance.

Additionally, risk compensation theory [Wilde, 1982] suggests that when people use technology they perceive as reducing risk, they may adjust their behavior accordingly. If optometrists believe the AI system enhances decision-making, they might be more forgiving of errors because they still feel their overall accuracy is improved. This could explain why acceptance remained stable despite the introduction of errors.

This maintenance of acceptance is ideal for deployment of an AI system in a more real-world scenario.

8.1.5.5 Final Reflections on Data Modality for AI System Design and Generalizability of Findings

We suggest that in the context of glaucoma, the system should not rely on only one source of data, but that it should combine multiple sources of data as used by optometrists. Our findings indicate that optometrists are more likely to trust AI systems when they reflect familiar diagnostic workflows and use clinically relevant combinations of data.

While these results are grounded in glaucoma diagnosis, the underlying principle of aligning AI decision-making with the clinician's multimodal reasoning process may generalize to other domains. In radiology, for instance, clinicians interpret a combination of imaging modalities (e.g., MRI, CT, X-rays) alongside patient history and lab results. In oncology, decision-making incorporates imaging, biopsy results, genomic profiles, treatment response data, and clinical notes. Similarly, in neurology, clinicians integrate structural and functional imaging (e.g., MRI, EEG), cognitive assessments, and longitudinal history to monitor disease progression and make diagnostic decisions. In all of these fields, interpreting and synthesizing multiple data sources is essential for accurate and confident decision-making.

Beyond healthcare, similar patterns of human decision-making occur in fields such as finance and engineering, where professionals routinely integrate multiple types of information to form judgments under uncertainty. In finance, for example, investment decisions are based on a combination of quantitative data (such as stock performance and economic indicators) and qualitative inputs (such as market news or geopolitical developments). In engineering domains like predictive maintenance, decisions rely on real-time sensor readings, historical maintenance records, and contextual factors like operating conditions. In both cases, experts do not rely on a single data stream but synthesize information from

diverse sources to inform action. AI systems that are designed to reflect this multimodal reasoning process by presenting integrated, context aware insights, are more likely to align with human decision-making, enhance trust, and support effective collaboration.

8.2 Limitations and Future Directions

While this research makes significant contributions to human factors, artificial intelligence, and optometry, certain limitations must be acknowledged. These limitations provide context for interpreting the findings and offer avenues for future exploration.

8.2.1 Participant Diversity

Although the study included a diverse participant pool in terms of gender and clinical experience, most participants were recruited from the Waterloo region in Ontario, Canada, and were likely trained at the University of Waterloo School of Optometry. Given the region's strong technological focus and urban-suburban setting, participants may be more comfortable integrating digital tools into their clinical practice. They may also serve a patient population that is generally more health literate, tech savvy, and accustomed to routine eye care thus shaping particular diagnostic workflows and decision-making approaches. These contextual influences could affect how participants engage with diagnostic technologies and decision-support systems.

As a result, the findings may not fully generalize to optometrists in other regions, particularly those in rural or resource limited settings, where access to technology, clinical workflows, and patient demographics may differ. Future research should aim to capture

a broader range of practice environments to better understand how these factors shape decision-making.

8.2.2 Limitation of Case Numbers

A limitation of this research is the restricted number of patient cases included in both studies. In the variability study, only four glaucoma cases were used, and in the modality study, two cases were examined across different system conditions. While these cases were carefully selected, clinically validated, and structured longitudinally to capture realistic diagnostic reasoning and intra-individual dynamics, the limited diversity of cases constrains the generalizability of the findings.

The choice to limit the case count was deliberate, aimed at balancing cognitive load, session duration, and data quality. Nonetheless, the relatively small number of unique cases remains a limitation, and future work should consider incorporating a broader range of cases and pathologies to strengthen external validity and explore additional variability in clinician decision-making and interaction with AI systems.

8.2.3 Limitation of Novice and Expert Clusters

The classification of participants into novice and expert clusters in this study was based on two glaucoma-specific criteria: years of clinical experience and the number of glaucoma patients seen per week. While this threshold-based approach was grounded in theoretical models, empirical precedents, and validated by practicing optometrists, it does present certain limitations.

First, clinical expertise exists on a continuum rather than as two discrete categories.

By employing a binary classification, this study may have oversimplified the nuanced progression of diagnostic expertise, potentially overlooking intermediate levels of proficiency. Second, the thresholds were determined based on prior literature and expert consultation but may not generalize universally across different clinical settings or specialties. Finally, the small sample size ($n = 14$) limited the viability of more sophisticated clustering techniques, such as K-means or DBSCAN, which could have identified natural groupings without imposing predefined cut-offs.

Future research with larger and more diverse participant samples should consider adopting continuous or multi-level measures of expertise to capture more granular differences in diagnostic behavior and interaction with AI systems.

8.2.4 Artificial Research Setting

The first study was conducted in a controlled, research-focused environment, where participants completed diagnostic tasks, questionnaires, and interviews. While this structured setup ensured consistency and allowed for a focused analysis of variations in decision-making between novice and expert optometrists, it does not fully replicate the complexities of real-world clinical practice. Factors such as time constraints, patient interactions, and workflow interruptions were absent, potentially limiting the ecological validity of the findings.

Future studies should extend these findings by evaluating decision-making variability and AI system interactions in real-world clinical environments. Field studies or in-clinic observations could provide insights into how time pressure, multitasking, and patient interactions influence diagnostic behaviors and system usability.

8.2.5 Generalizability Across Clinical Specialties

This research was designed specifically for optometry and glaucoma diagnosis. While the proposed AI system guidelines and human-centric design principles may have broader applicability, their effectiveness cannot be assumed in other healthcare domains without further investigation.

Future research could explore the adaptation of these principles in other medical specialties, such as oncology, cardiology, or radiology. Investigating whether similar multimodal AI designs are effective in contexts requiring different diagnostic workflows or data integration challenges would expand the applicability of these findings across healthcare.

8.2.6 Mock AI System and Real-World Deployment

The mock AI system used in this research was developed using real patient data and informed by expert input, which enhanced the authenticity of user trust and acceptability assessments. However, it is important to note that the system did not have an actual AI model running in the background. Instead, it functioned as a simulated interface designed to explore how clinicians might interact with AI-generated outputs. While valuable for eliciting initial impressions, this limits the realism of user responses, particularly regarding trust in dynamically generated or uncertain AI decisions. Additionally, the system operated in a simulated research environment rather than being deployed in real-world clinical workflows. As a result, the study does not fully evaluate the feasibility, trust, and acceptability of AI systems in dynamic, real-world conditions.

Future work should focus on deploying operational AI systems in clinical settings to assess their usability, scalability, and integration into daily workflows. Studies that evaluate

how AI systems impact diagnostic accuracy, clinician efficiency, and patient outcomes in real-world environments will be critical for transitioning from research prototypes to practical tools.

8.2.7 Idealized Design and Limitations of the AI Prototype

The AI systems presented in this study were implemented as mock-ups to simulate realistic decision-support behavior in both unimodal and multimodal conditions. While this approach enabled controlled experimentation and meaningful assessment of user interaction, it also has inherent limitations. Though the technical feasibility of the features proposed in mockup-AI system is discussed in Section 5.5, seamless integration of diverse data modalities and reliable context-aware explanations, remain technically challenging to implement with current AI capabilities.

Future work should involve close collaboration with AI development experts to assess the data integration and explanations feature proposed in our mockup-AI system. This would enable alignment between user-centered design aspirations and the current state of AI technology, ensuring that future systems are both realistic and impactful.

8.2.8 Focus on Multimodal and Unimodal Systems

This research compared unimodal and multimodal AI systems, highlighting the advantages of multimodal systems for trust, performance, and acceptability. However, the study did not explore hybrid systems—those capable of dynamically combining modalities based on clinical needs or user preferences. Such systems may provide additional flexibility in scenarios where resource constraints or incomplete data are present. In addition, in the

unimodal condition, only fundus photographs were used as the input modality. Other potential unimodal modalities, such as OCT or visual field data, which are also commonly used in clinical practice, were not examined. This represents a limitation of the study, as these alternative modalities could influence trust, performance, and acceptance differently. Future work should consider comparing different unimodal inputs to better understand how specific modalities shape user perceptions and decision-making.

Future research could examine hybrid AI systems that adaptively prioritize data modalities depending on the clinical context. For example, studies could assess the performance of systems that adjust their decision-making approach based on urgency or the availability of certain diagnostic data. Investigating these hybrid systems could offer valuable insights for optimizing AI system design.

8.2.9 Limited Longitudinal Interaction

The study primarily examined diagnostic decisions, trust, and acceptability during brief interactions with the AI systems. While this provided valuable insights into initial user perceptions, it does not account for how trust, reliance, or user performance may evolve over prolonged use. Understanding these dynamics is essential for assessing long-term adoption and efficacy.

Longitudinal studies could track user interactions with AI systems over several months to evaluate how trust develops, how clinicians refine their use of AI tools, and how sustained exposure impacts diagnostic accuracy and efficiency. These studies could also explore how AI systems influence training and skill development among clinicians, particularly novices.

8.3 Contributions of the Study

This PhD thesis makes significant contributions across three domains: human factors, artificial intelligence (AI), and optometry. These contributions address fundamental research questions and challenges in clinical decision-making posed because of decision making variations and data modality and provide guidance on the design of AI systems.

8.3.1 Contributions to Human Factors

This thesis contributes to the field of human factors by advancing the understanding of cognitive processes and decision-making strategies. Key contributions include:

8.3.1.1 Insights into Variability in Decision-Making

Extensive human factors research has shown that experts rely on pattern recognition and holistic decision-making, whereas novices adopt a rule-based, data-intensive approach. Building on this foundation, our study advances the understanding of cognitive differences in glaucoma diagnosis by demonstrating how experience shapes clinical assessment strategies and the use of available data.

Our work contributes to human factors research by refining existing literature of medical decision-making, emphasizing that expertise influences not only efficiency but also adaptability in complex diagnostic tasks. We highlight how experts dynamically balance competing diagnostic factors, selectively filter information, and integrate progression analysis into their judgments—capabilities that novices struggle to develop. Additionally, our study provide insights on patient management decisions, including risk factors assessment and follow-up planning by both novices and experts. These findings offer implications for

the design of decision-support systems aimed at supporting both expert and novices in clinical reasoning.

8.3.1.2 Data Prioritization and Cognitive Efficiency

Chapter 3 and 4 confirms that experts prioritize critical factors and discard extraneous information more effectively than novices, reducing cognitive load and improving decision-making efficiency, an observation well established in prior research. However, we extend this understanding by identifying which factors are prioritized and when these prioritization strategies are applied in glaucoma diagnosis. This deeper analysis reveals how experts dynamically adjust their focus based on disease severity, diagnostic uncertainty, and data availability, providing new insights into the cognitive mechanisms that differentiate expert and novice decision-making in glaucoma diagnosis.

8.3.1.3 Heuristic Approaches and Clinical Judgment

The findings in chapter 3 show that experienced clinicians rely on heuristics, while novices use more structured, rule-based approaches in glaucoma decision-making. While this pattern aligns with established models of clinical reasoning, our contribution lies in illustrating how these strategies manifest in the specific context of glaucoma care. By situating reasoning processes within real outpatient assessments, this work adds contextual depth to existing understandings of how expertise shapes decision-making in complex clinical environments

8.3.1.4 Design of Human-Centric Decision Support Systems

By linking cognitive variability to clinical outcomes, this research provides insights into the design of AI-driven decision support systems that accommodate users' cognitive processes and limitations. The proposed guidelines in section 8.1.3 consists of some principles extracted from glaucoma study but can be applicable across medical decision-making. A key focus is ensuring that AI recommendations align with the expertise level of the user, adapting decision support for both novice and expert clinicians.

The glaucoma-specific guidance extends these principles by addressing the impact of limited data availability as mentioned in Section 8.1.3.1 and proposing AI strategies to support clinical decision-making at different stages of disease progression as suggested in Section 8.1.3.2. Additionally, we also proposed in Section 8.1.3.5 to provide balanced positive and negative justifications dynamics so that it can be easily transformed to disease severity, helping optometrists refine diagnostic and treatment decisions.

By providing these human centric AI system guidelines, this research provides a structured foundation for developing AI systems that not only enhance glaucoma management but also inform decision-support design in other medical domains.

8.3.2 Contributions to Artificial Intelligence

The thesis makes substantial contributions to the development and implementation of AI systems for glaucoma diagnosis, particularly in addressing the challenges of variability in expertise and the integration of multimodal data. Key contributions include:

8.3.2.1 Bridging Expertise Gaps with AI Systems

The study provides a detailed analysis of how AI can support users at varying levels of expertise:

- **Human-AI collaboration models:** This thesis explores Human-AI collaboration models by providing guidelines for developing AI systems that support diverse decision-making processes for both experts and novices. It further evaluates unimodal and multimodal AI systems in terms of trust and user acceptance, ensuring their effectiveness in augmenting clinical decision-making rather than replacing human expertise.
- **User-centered AI systems:** The findings inform the design of AI tools that adapt to the user's cognitive and experiential needs. Novices benefit from systems that provide structured, step-by-step decision support, while experts are better supported by systems that emphasize efficiency and critical data prioritization.

8.3.2.2 Clinician-Driven Approach to AI Design

This study contributes to AI development by taking a clinician-driven approach to understanding glaucoma decision-making. By starting with the experiences of optometrists and examining how they reason through real cases, the work ensures that AI systems are informed by actual clinical practices. This approach helps identify where AI can support decision-making for both novice and expert clinicians—offering guidance where needed without overriding expertise. It supports the development of AI tools that are not only useful and well-integrated into clinical workflows but also trustworthy and acceptable to end users.

8.3.2.3 Enhancing diagnostic accuracy through multimodal AI systems

The study reveals the critical role of multimodal AI systems in improving diagnostic accuracy and user performance:

- **Data Integration and User Accuracy:** Multimodal systems, which synthesize diverse data such as fundus imaging, OCT, and clinical indicators, could significantly improve human and AI alignment and results in expert-level decisions. This integration enhances diagnostic accuracy across experience levels. For novice clinicians, the AI system provides interpretive support by highlighting relevant features across modalities, helping them recognize patterns they may not yet be familiar with. For experts, it acts as a second reader, streamlining decision-making in complex or ambiguous cases without disrupting their workflow. This form of multimodal AI support is particularly valuable in domains where diagnostic decisions rely on synthesizing diverse and often subtle signals—such as retinal disease, oncology (e.g., integrating imaging and pathology), or cardiology (e.g., combining ECG, imaging, and clinical risk factors). The framework demonstrated here in glaucoma decision-making can thus be generalized to other specialties that demand integration of heterogeneous data and decision support adaptable to different user expertise levels.

- **Trust and acceptance:** Chapters 6 and 7 provided empirical evidence that multimodal AI systems outperform unimodal systems in fostering trust and user acceptance. While previous research has extensively examined the impact of data modality on the accuracy and efficiency of AI systems, its role in shaping trust and acceptance, especially in healthcare and focused on input data modality, has been largely overlooked. This work addresses this gap by systematically investigating how different input modalities influence clinicians' confidence and adoption of AI systems. Addi-

tionally, it establishes the importance of aligning AI design with clinicians' diagnostic workflows to enhance trust. These contributions offer a new perspective on the development of AI systems that are not only accurate but also intuitive and smooth integration into clinical practice

- **Error resilience:** This research contributes into the dynamics of user acceptance in the presence of AI system errors. Unlike prior work that emphasizes algorithm aversion, where users tend to reject AI after observing mistakes [Zerilli et al., 2022, Dietvorst et al., 2018], this study shows that multimodal systems, can maintain clinician trust and acceptance despite occasional errors. This finding advances the field by identifying design conditions under which AI can remain resilient in real-world clinical settings.

- **Design recommendations for human centered AI:**

  This thesis contributes to human-centred AI design by, first, proposing a role-sensitive system that supports both novice and expert clinicians in decision-making, tailored specifically to glaucoma diagnosis. Second, it evaluates unimodal and multimodal AI systems and finds that multimodal systems better align with clinical workflows and foster greater user trust. Based on these findings, the thesis offers design guidelines that go beyond model-centric priorities such as accuracy, focusing instead on clinician needs.

By addressing both expertise variability and data modality, this research advances the development of robust, user-centric AI systems that improve diagnostic outcomes in healthcare.

8.3.3 Contributions to optometry

This thesis makes meaningful contributions to the field of optometry by addressing critical challenges in glaucoma diagnosis and offering practical solutions for integrating AI tools into clinical practice. Key contributions include:

8.3.3.1 Understanding Decision-Making in Glaucoma Diagnosis

This research offers detailed insights into how optometrists prioritize and interpret clinical data when diagnosing glaucoma. It reveals gaps in novice decision-making, such as an over-reliance on specific subsets of data, and contrasts these with expert strategies like progression analysis and structure-function correlation. While previous studies have examined variability in diagnostic approaches, they have primarily focused on comparing the accuracy of experts and novices, finding their agreements [Koklanis and Thorburn, 2021] or on training interventions [Zhu et al., 2024]. In contrast, this study contributes to a deeper understanding of diagnostic reasoning, which can inform the development of AI systems that more effectively align with clinical decision-making processes

8.3.3.2 Guidelines for AI Integration in Optometric Practice

The proposed AI system guidelines are specifically tailored to the clinical and diagnostic needs of optometrists. While earlier work has primarily focused on developing diagnostic models [Krishnan et al., 2025], offering training-related AI recommendations for optometrists [Zhu et al., 2024], or exploring practitioner attitudes toward AI adoption [Ho et al., 2022], comprehensive design and integration guidelines remain scarce. This thesis addresses that gap by providing practical recommendations in section 8.1.3, aimed

at supporting both novice and expert clinicians in glaucoma diagnosis and broader optometric decision-making. The research also highlights the importance of multimodal data integration, for fostering trust and acceptance of AI systems in everyday optometric practice.

8.3.3.3 Enhancing Training and Education

The research highlights significant opportunities to incorporate AI system and tools into optometric education. By revealing specific gaps in novice decision-making—such as difficulty in interpreting progression. this research provides a foundation for targeted educational interventions. AI systems designed with educational features, such as step-by-step decision support, visual explanations, and expert comparison modes, can help novices develop critical diagnostic skills, including pattern recognition, structural-function correlation, and data prioritization. AI systems have the potential to accelerate learning curves, and support self-directed learning, ultimately helping trainees transition more effectively toward expert-level performance. Furthermore, such tools could complement traditional clinical supervision, making training more scalable and tailored to individual learning needs."

8.3.3.4 Practical Implications for Glaucoma Management

The study provides actionable insights into how AI tools can support glaucoma diagnosis, accurate staging, and effective patient management in glaucoma care. This has implications for improving long-term outcomes and reducing the burden of this sight-threatening disease.

These contributions advance optometry by addressing clinical challenges, integrating advanced technologies, and providing a roadmap for the future of AI-driven diagnostic

care.

8.3.4 Final Contribution Remarks

This thesis makes theoretical and practical contributions to the fields of human factors, artificial intelligence, and optometry. By addressing cognitive variability in glaucoma diagnosis and advancing human-centric AI system design, the research provides a comprehensive foundation for enhancing healthcare delivery through human-centric and technology-driven solutions. These contributions pave the way for future research and development in AI-assisted healthcare systems.

8.4 Conclusion

This PhD thesis has explored the intersection of human factors, artificial intelligence, and clinical decision-making, specifically within the context of glaucoma diagnosis. Through a series of comprehensive studies, this research has provided new insights into the variability of diagnostic decision-making, the influence of expertise levels, and the critical role of AI system design in supporting clinicians. The findings contribute to both theoretical understanding and practical advancements in these fields, paving the way for future research and implementation.

The first study examined the decision-making processes of novice and expert optometrists, highlighting significant differences in their cognitive strategies, data prioritization, and diagnostic reasoning. These findings demonstrate how experience shapes clinical judgment and underline the importance of tailoring decision-support tools to different expertise levels.

The subsequent studies investigated the role of AI system data modalities, comparing unimodal and multimodal systems, in fostering trust and acceptance. The results unequivocally showed the superiority of multimodal systems, which integrate diverse data sources to align with clinicians' natural diagnostic workflows. These systems not only improved user trust and increased user acceptance but also enhanced diagnostic performance across expertise levels, bridging the gap between novice and expert decision-making.

In addition to these findings, this thesis makes a significant contribution by proposing design guidelines for human-centric AI systems. These guidelines emphasize the development of systems that cater to users with diverse levels of experience and select data modalities that align with users' mental models, ensuring optimal trust and high acceptance.

Despite its contributions, this research acknowledges certain limitations, such as the reliance on simulated environments, the geographical focus of the participant pool, and the brief nature of interactions with the mock AI system. These limitations provide avenues for future research, including the deployment of fully operational AI systems in real-world clinical settings, longitudinal studies of user adaptation, and investigations into the generalizability of findings across other medical domains.

In conclusion, this thesis shows the transformative potential of AI, demonstrating how thoughtfully designed, human-centric systems can improve decision making, foster optimum trust and increase AI system acceptance among users. By addressing the variability in expertise and leveraging the power of multimodal data integration, AI systems can bridge critical gaps in clinical decision-making. This work provides a foundation for advancing AI-assisted healthcare and sets the stage for future innovations that prioritize both technological sophistication and human factors.

References

[Abràmoff et al., 2018] Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., and Folk, J. C. (2018). Pivotal trial of an autonomous ai-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine*, 1:39.

[Afroogh et al., 2024] Afroogh, S., Akbari, A., Malone, E., Kargar, M., and Alambeigi, H. (2024). Trust in ai: progress, challenges, and future directions. *Humanities and Social Sciences Communications*, 11:1568.

[Agyei and Özge Razi, 2022] Agyei, C. and Özge Razi (2022). The effect of extended utaut model on efls' adaptation to flipped classroom. *Education and Information Technologies*, 27:1865–1882.

[Ahn et al., 2024] Ahn, D., Almaatouq, A., Gulabani, M., and Hosanagar, K. (2024). Impact of model interpretability and outcome feedback on trust in ai. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–25. ACM.

[Ajenaghughrure et al., 2020] Ajenaghughrure, I. B., da Costa Sousa, S. C., and Lamas, D. (2020). Risk and trust in artificial intelligence technologies: A case study of autonomous vehicles. In *2020 13th International Conference on Human System Interaction (HSI)*, pages 118–123. IEEE.

[Akter et al., 2022] Akter, N., Fletcher, J., Perry, S., Simunovic, M. P., Briggs, N., and Roy, M. (2022). Glaucoma diagnosis using multi-feature analysis and a deep learning technique. *Scientific Reports*, 12:8064.

[Akter et al., 2025] Akter, N., Gordon, J., Li, S., Poon, M., Perry, S., Fletcher, J., Chan, T., White, A., and Roy, M. (2025). Glaucoma detection and staging from visual field images using machine learning techniques. *PloS one*, 20(1):e0316919.

[AlQudah et al., 2021] AlQudah, A. A., Al-Emran, M., and Shaalan, K. (2021). Technology acceptance in healthcare: A systematic review. *Applied Sciences*, 11:10537.

[AlShawabkeh et al., 2024] AlShawabkeh, M., AlRyalat, S. A., Bdour, M. A., Alni'mat, A., and Al-Akhras, M. (2024). The utilization of artificial intelligence in glaucoma: diagnosis versus screening. *Frontiers in Ophthalmology*, 4.

[Andersen et al., 2023] Andersen, T. O., Nunes, F., Wilcox, L., Coiera, E., and Rogers, Y. (2023). Introduction to the special issue on human-centred ai in healthcare: Challenges appearing in the wild. *ACM Transactions on Computer-Human Interaction*, 30:1–12.

[Andrews et al., 2023] Andrews, R. W., Lilly, J. M., Srivastava, D., and Feigh, K. M. (2023). The role of shared mental models in human-ai teams: a theoretical review. *Theoretical Issues in Ergonomics Science*, 24:129–175.

[Armutat et al., 2024] Armutat, S., Wattenberg, M., and Mauritz, N. (2024). Artificial intelligence: Gender-specific differences in perception, understanding, and training interest. In *International Conference on Gender Research*, pages 36–43. Academic Conferences International Limited.

[Asan et al., 2020] Asan, O., Bayrak, A. E., and Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: Focus on clinicians. *Journal of Medical Internet Research*, 22:e15154.

[Ashtari-Majlan et al., 2024] Ashtari-Majlan, M., Dehshibi, M. M., and Masip, D. (2024). Glaucoma diagnosis in the era of deep learning: A survey. *Expert Systems with Applications*, 256:124888.

[Ashtari-Majlan and Masip, 2025] Ashtari-Majlan, M. and Masip, D. (2025). Spatial-aware transformer-gru framework for enhanced glaucoma diagnosis from 3d oct imaging. *IEEE Journal of Biomedical and Health Informatics*.

[Azuara-Blanco et al., 2007] Azuara-Blanco, A., Burr, J., Thomas, R., Maclennan, G., and McPherson, S. (2007). The accuracy of accredited glaucoma optometrists in the diagnosis and treatment recommendation for glaucoma. *British Journal of Ophthalmology*, 91:1639–1643.

[Bai et al., 2024] Bai, A., Si, M., Xue, P., Qu, Y., and Jiang, Y. (2024). Artificial intelligence performance in detecting lymphoma from medical imaging: a systematic review and meta-analysis. *BMC Medical Informatics and Decision Making*, 24:13.

[Bajaj and Nidumolu, 1998] Bajaj, A. and Nidumolu, S. R. (1998). A feedback model to understand information system usage. *Information & Management*, 33:213–224.

[Barber and House, 1985] Barber, B. and House, J. S. (1985). The logic and limits of trust. *Social Forces*, 64:219.

[Baroni et al., 2022] Baroni, I., Calegari, G. R., Scandolari, D., and Celino, I. (2022). Ai-tam: a model to investigate user acceptance and collaborative intention inhuman-in-the-loop ai applications. *Human Computation*, 9:1–21.

[Bayer et al., 2022] Bayer, S., Gimpel, H., and Markgraf, M. (2022). The role of domain expertise in trusting and following explainable ai decision support systems. *Journal of Decision Systems*, 32(1):110–138.

[Beauxis-Aussalet et al., 2021] Beauxis-Aussalet, E., Behrisch, M., Borgo, R., Chau, D. H., Collins, C., Ebert, D., El-Assady, M., Endert, A., Keim, D. A., Kohlhammer, J., Oelke, D., Peltonen, J., Riveiro, M., Schreck, T., Strobelt, H., and van Wijk, J. J. (2021). The role of interactive visualization in fostering trust in ai. *IEEE Computer Graphics and Applications*, 41:7–12.

[Benner, 1982] Benner, P. (1982). From novice to expert. *AJN The American Journal of Nursing*, 82(3):402–407.

[Benner et al., 1984] Benner, P. et al. (1984). From novice to expert. *Menlo Park*, 84(1480):10–1097.

[Berchuck et al., 2019] Berchuck, S. I., Mwanza, J.-C., and Warren, J. L. (2019). A spatially varying change points model for monitoring glaucoma progression using visual field data. *Spatial Statistics*, 30:1–26.

[Bettin and Matteo, 2013] Bettin, P. and Matteo, F. D. (2013). Glaucoma: Present challenges and future trends. *Ophthalmic Research*, 50:197–208.

[Bhatt and Bae, 2023] Bhatt, A. B. and Bae, J. (2023). Collaborative intelligence to catalyze the digital transformation of healthcare. *npj Digital Medicine*, 6:177.

[Bhattacherjee, 2001] Bhattacherjee, A. (2001). Understanding information systems continuance: An expectation-confirmation model. *MIS Quarterly*, 25:351.

[Bian et al., 2020] Bian, Y., Xiang, Y., Tong, B., Feng, B., and Weng, X. (2020). Artificial intelligence–assisted system in postoperative follow-up of orthopedic patients: Exploratory quantitative and qualitative study. *Journal of Medical Internet Research*, 22:e16896.

[Bozorgmehr and Weltermann, 2023] Bozorgmehr, A. and Weltermann, B. (2023). Prediction of chronic stress and protective factors in adults: Development of an interpretable prediction model based on xgboost and shap using national cross-sectional degs1 data. *JMIR AI*, 2:e41868.

[Brunyé et al., 2019] Brunyé, T. T., Drew, T., Weaver, D. L., and Elmore, J. G. (2019). A review of eye tracking for understanding and improving diagnostic interpretation. *Cognitive Research: Principles and Implications*, 4:7.

[Burns, 2019] Burns, C. M. (2019). Models of expertise. *The Oxford Handbook of Expertise*, page 451.

[Caddick et al., 2023] Caddick, Z. A., Fraundorf, S. H., Rottman, B. M., and Nokes-Malach, T. J. (2023). Cognitive perspectives on maintaining physicians' medical expertise: Ii. acquiring, maintaining, and updating cognitive skills. *Cognitive Research: Principles and Implications*, 8:47.

[Cai et al., 2019] Cai, C. J., Winter, S., Steiner, D., Wilcox, L., and Terry, M. (2019). "hello ai": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 3:1–24.

[Campbell et al., 2020] Campbell, C. G., Ting, D. S. W., Keane, P. A., and Foster, P. J. (2020). The potential application of artificial intelligence for diagnosis and management of glaucoma in adults. *British Medical Bulletin*, 134:21–33.

[Campitelli and Gobet, 2010] Campitelli, G. and Gobet, F. (2010). Herbert simon's decision-making approach: Investigation of cognitive processes in experts. *Review of General Psychology*, 14:354–364.

[Cao et al., 2009] Cao, Y., Theune, M., and Nijholt, A. (2009). Towards cognitive-aware multimodal presentation: the modality effects in high-load hci. In *Engineering Psychology and Cognitive Ergonomics: 8th International Conference, EPCE 2009, Held as Part of HCI International 2009, San Diego, CA, USA, July 19-24, 2009. Proceedings 8*, pages 3–12. Springer.

[Chai et al., 2018] Chai, Y., Liu, H., and Xu, J. (2018). Glaucoma diagnosis based on both hidden features and domain knowledge through deep learning models. *Knowledge-Based Systems*, 161:147–156.

[Chaiken, 1980] Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of personality and social psychology*, 39(5):752.

[Chao et al., 2022] Chao, X., Ran, Q., Chen, J., Li, T., Qian, Q., and Ergu, D. (2022). Regulatory technology (reg-tech) in financial stability supervision: Taxonomy, key methods, applications and future directions. *International Review of Financial Analysis*, 80:102023.

[Chaurasia et al., 2022] Chaurasia, A. K., Greatbatch, C. J., and Hewitt, A. W. (2022). Diagnostic accuracy of artificial intelligence in glaucoma screening and clinical practice. *Journal of Glaucoma*, 31:285–299.

[Chayan et al., 2022] Chayan, T. I., Islam, A., Rahman, E., Reza, M. T., Apon, T. S., and Alam, M. G. R. (2022). Explainable ai based glaucoma detection using transfer learning and lime. In *2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–6. IEEE.

[Chen et al., 2024] Chen, G., Fan, J., and Azam, M. (2024). Exploring artificial intelligence (ai) chatbots adoption among research scholars using unified theory of acceptance and use of technology (utaut). *Journal of Librarianship and Information Science*.

[Chen et al., 2022a] Chen, J. S., Lin, W.-C., Kaluzny, J. V., Chen, A., Chiang, M. F., and Hribar, M. R. (2022a). Discrepancies in ophthalmic medication documentation for glaucoma patients. *Ophthalmology Science*, 2:100091.

[Chen et al., 2022b] Chen, M., Zhang, B., Cai, Z., Seery, S., Gonzalez, M. J., Ali, N. M., Ren, R., Qiao, Y., Xue, P., and Jiang, Y. (2022b). Acceptance of clinical artificial intelligence among physicians and medical students: A systematic review with cross-sectional survey. *Frontiers in Medicine*, 9.

[Chen et al., 2023] Chen, Y., Wang, W., and Chen, X. M. (2023). Bibliometric methods in traffic flow prediction based on artificial intelligence. *Expert Systems with Applications*, 228:120421.

[Chi et al., 1981] Chi, M. T. H., Feltovich, P. J., and Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5:121–152.

[Choudhury et al., 2022] Choudhury, A., Saremi, M. L., and Urena, E. (2022). *Perception, Trust, and Accountability Affecting Acceptance of Artificial Intelligence*, pages 105–124.

[Coan et al., 2023] Coan, L. J., Williams, B. M., Adithya, V. K., Upadhyaya, S., Alkafri, A., Czanner, S., Venkatesh, R., Willoughby, C. E., Kavitha, S., and Czanner, G. (2023). Automatic detection of glaucoma via fundus imaging and artificial intelligence: A review. *Survey of ophthalmology*, 68(1):17–41.

[Cook, 1993] Cook, G. J. (1993). An empirical investigation of information search strategies with implications for decision support system design. *Decision Sciences*, 24:683–698.

[Croskerry et al., 2023] Croskerry, P., Campbell, S. G., and Petrie, D. A. (2023). The challenge of cognitive science for medical diagnosis. *Cognitive Research: Principles and Implications*, 8:13.

[Curran et al., 2022] Curran, K., Congdon, N., Peto, T., Dardis, C., Nguyen, Q. N., Hoang, T. T., Bannon, F., Luu, A., Mai, T. Q., Nguyen, V. T., Nguyen, H. T., Tran, H., Tran, H. H., and Lohfeld, L. (2022). Capturing the clinical decision-making processes of expert and novice diabetic retinal graders using a 'think-aloud' approach. *Eye*, 36:1019–1026.

[Dai et al., 2024] Dai, C.-P., Ke, F., Pan, Y., Moon, J., and Liu, Z. (2024). Effects of artificial intelligence-powered virtual agents on learning outcomes in computer-based simulations: A meta-analysis. *Educational Psychology Review*, 36:31.

[Das et al., 2023] Das, S., Kuht, H. J., De Silva, I., Deol, S. S., Osman, L., Burns, J., Sarvananthan, N., Sarodia, U., Kapoor, B., Islam, T., et al. (2023). Feasibility and clinical utility of handheld fundus cameras for retinal imaging. *Eye*, 37(2):274–279.

[Davis, 1989] Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13:319.

[de Voogd et al., 2006] de Voogd, S., Ikram, M. K., Wolfs, R. C., Jansonius, N. M., Witteman, J. C., Hofman, A., and de Jong, P. T. (2006). Is diabetes mellitus a risk factor for open-angle glaucoma? *Ophthalmology*, 113:1827–1831.

[Dietvorst et al., 2015] Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144:114–126.

[Dietvorst et al., 2018] Dietvorst, B. J., Simmons, J. P., and Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64:1155–1170.

[Dignum, 2019] Dignum, V. (2019). *Responsible Artificial Intelligence*. Springer International Publishing.

[Dillon and Morris, 1996] Dillon, A. and Morris, M. G. (1996). *User acceptance of new information technology: theories and models*. Medford, N.J.: Information Today.

[Dishaw and Strong, 1999] Dishaw, M. T. and Strong, D. M. (1999). Extending the technology acceptance model with task–technology fit constructs. *Information & Management*, 36:9–21.

[Doozandeh et al., 2022] Doozandeh, A., Yazdani, S., Pakravan, M., Ghasemi, Z., Hassanpour, K., Hatami, M., and Ansari, I. (2022). Risk of missed diagnosis of primary open-angle glaucoma by eye care providers. *Journal of Current Ophthalmology*, 34:404.

[Doshi-Velez and Kim, 2017] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

[Dreyfus and Dreyfus, 1980] Dreyfus, S. E. and Dreyfus, H. L. (1980). A five-stage model of the mental activities involved in directed skill acquisition.

[Du and Lv, 2024] Du, L. and Lv, B. (2024). Factors influencing students' acceptance and use generative artificial intelligence in elementary education: an expansion of the utaut model. *Education and Information Technologies*.

[Du et al., 2023] Du, Y., Li, T., and Gao, C. (2023). Why do designers in various fields have different attitude and behavioral intention towards ai painting tools? an extended utaut model. *Procedia Computer Science*, 221:1519–1526.

[Dunford et al., 2011] Dunford, F., Reeve, J., and Larmer, P. (2011). Determining differences between novice and expert physiotherapists in undertaking emergency on-call duties. *New Zealand Journal of Physiotherapy*, 39(1):20.

[Dwivedi et al., 2021] Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., Medaglia, R., Meunier-FitzHugh, K. L., Meunier-FitzHugh, L. C. L., Misra, S., Mogaji, E., Sharma, S. K., Singh, J. B., Raghavan, V., Raman, R., Rana, N. P., Samothrakis, S., Spencer, J., Tamilmani, K., Tubadji, A., Walton, P., and Williams, M. D. (2021). Artificial intelligence (ai): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57:101994.

[Dzindolet et al., 2003] Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., and Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58:697–718.

[Ericsson et al., 1993] Ericsson, K. A., Krampe, R. T., and Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100:363–406.

[Esmaeilzadeh et al., 2021] Esmaeilzadeh, P., Mirzaei, T., and Dharanikota, S. (2021). Patients' perceptions toward human–artificial intelligence interaction in health care: Experimental study. *Journal of Medical Internet Research*, 23:e25856.

[Extance, 2023] Extance, A. (2023). Chatgpt has entered the classroom: how llms could transform education. *Nature*, 623:474–477.

[Faisal et al., 2025] Faisal, S., Baby, B., Samoth, D., Ghaffar, F., Aslam, Y., Stavropoulos, A., McCready-Branch, B., Elba, G., Gill, J. K., Choudhuri, P., et al. (2025). Designing a real-time medication intake and adherence dashboard: Features, functionality, and data display to meet the needs of patients, care partners and healthcare providers. *Journal of Medical Systems*, 49(1):1–18.

[Faulkner, 2003] Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, 35:379–383.

[fei Sun et al., 2024] fei Sun, P., jian Wang, C., Du, Y., Zhan, Y.-Q., pan Shen, P., and hui Ding, Y. (2024). Development and validation of a nomogram to predict risk of septic cardiomyopathy in the intensive care unit. *Scientific Reports*, 14:14114.

[Ferrario and Loi, 2022] Ferrario, A. and Loi, M. (2022). How explainability contributes to trust in ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1457–1466. ACM.

[Fu et al., 2018] Fu, H., Cheng, J., Xu, Y., Zhang, C., Wong, D. W. K., Liu, J., and Cao, X. (2018). Disc-aware ensemble network for glaucoma screening from fundus image. *IEEE Transactions on Medical Imaging*, 37:2493–2501.

[Fu et al., 2017] Fu, H., Xu, Y., Lin, S., Zhang, X., Wong, D. W. K., Liu, J., Frangi, A. F., Baskaran, M., and Aung, T. (2017). Segmentation and quantification for angle-closure glaucoma assessment in anterior segment oct. *IEEE Transactions on Medical Imaging*, 36:1930–1938.

[Fujimori et al., 2022] Fujimori, R., Liu, K., Soeno, S., Naraba, H., Ogura, K., Hara, K., Sonoo, T., Ogura, T., Nakamura, K., and Goto, T. (2022). Acceptance, barriers, and facilitators to implementing artificial intelligence–based decision support systems in emergency departments: Quantitative and qualitative evaluation. *JMIR Formative Research*, 6:e36501.

[Gall et al., 1977] Gall, J. E., Norwood, D. D., et al. (1977). *Demonstration and evaluation of a total hospital information system*. Dept. of Health, Education, and Welfare, Public Health Service, Health ....

[Gallimore et al., 2019] Gallimore, D., Lyons, J. B., Vo, T., Mahoney, S., and Wynne, K. T. (2019). Trusting robocop: Gender-based effects on trust of an autonomous robot. *Frontiers in Psychology*, 10:482.

[Ghaffar, 2024] Ghaffar, F. (2024). Diagnostic decision-making variability in glaucoma: Comparative analysis of novice and expert optometrists to inform ai system design.

[Ghaffar et al., 2024a] Ghaffar, F., Catherine, B., and Nadine, F. (2024a). Effect of AI system data modality on users trust. *Borealis*.

[Ghaffar et al., 2024b] Ghaffar, F., Furtado, N., Ali, I., and Burns, C. (2024b). Diagnostic decision-making variability in glaucoma: Comparative analysis of novice and expert optometrists to inform ai system design (preprint). *JMIR Medical Informatics*.

[Ghaffar et al., 2016] Ghaffar, F., Uyyanonvara, B., Sinthanayothin, C., Ali, L., and Kaneko, H. (2016). Detection of exudates from retinal images using morphological compact tree. In *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 1–5. IEEE.

[Ghazizadeh et al., 2012] Ghazizadeh, M., Lee, J. D., and Boyle, L. N. (2012). Extending the technology acceptance model to assess automation. *Cognition, Technology & Work*, 14:39–49.

[Gigerenzer and Gaissmaier, 2011] Gigerenzer, G. and Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62:451–482.

[Gille et al., 2020] Gille, F., Jobin, A., and Ienca, M. (2020). What we talk about when we talk about trust: Theory of trust for ai in healthcare. *Intelligence-Based Medicine*, 1-2:100001.

[Glikson and Woolley, 2020] Glikson, E. and Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14:627–660.

[Goceri, 2019] Goceri, E. (2019). *Skin Disease Diagnosis from Photographs Using Deep Learning*, pages 239–246. Springer.

[Goel et al., 2023] Goel, M., Tomar, P. K., Vinjamuri, L. P., Reddy, G. S., Al-Taee, M., and Alazzam, M. B. (2023). Using ai for predictive analytics in financial management. In

2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pages 963–967. IEEE.

[Goodfellow et al., 2020] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63:139–144.

[Goodhue and Thompson, 1995] Goodhue, D. L. and Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly*, 19:213.

[Gordon et al., 2002] Gordon, M. O., Beiser, J. A., Brandt, J. D., Heuer, D. K., Higginbotham, E. J., Johnson, C. A., Keltner, J. L., Miller, J. P., Parrish, R. K., Wilson, M. R., et al. (2002). The ocular hypertension treatment study: baseline factors that predict the onset of primary open-angle glaucoma. *Archives of ophthalmology*, 120(6):714–720.

[Group et al., 2007] Group, E. G. P. S. E. et al. (2007). Predictive factors for open-angle glaucoma among patients with ocular hypertension in the european glaucoma prevention study. *Ophthalmology*, 114(1):3–9.

[Gulshan et al., 2016] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410.

[Gursoy et al., 2019] Gursoy, D., Chi, O. H., Lu, L., and Nunkoo, R. (2019). Consumers acceptance of artificially intelligent (ai) device use in service delivery. *International Journal of Information Management*, 49:157–169.

[Hagerty and Rubinov, 2019] Hagerty, A. and Rubinov, I. (2019). Global ai ethics: a review of the social impacts and ethical implications of artificial intelligence. *arXiv preprint arXiv:1907.07892*.

[Hammond and R, 1981] Hammond and R, K. (1981). Principles of organization in intuitive and analytical cognition.

[Haselhuhn et al., 2015] Haselhuhn, M. P., Kennedy, J. A., Kray, L. J., Van Zant, A. B., and Schweitzer, M. E. (2015). Gender differences in trust dynamics: Women trust more than men following a trust violation. *Journal of Experimental Social Psychology*, 56:104–109.

[Hassija et al., 2024] Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., and Hussain, A. (2024). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 16:45–74.

[Hemelings et al., 2020] Hemelings, R., Elen, B., Barbosa-Breda, J., Lemmens, S., Meire, M., Pourjavan, S., Vandewalle, E., de Veire, S. V., Blaschko, M. B., Boever, P. D., and Stalmans, I. (2020). Accurate prediction of glaucoma from colour fundus images with a convolutional neural network that relies on active and transfer learning. *Acta Ophthalmologica*, 98.

[Hennink and Kaiser, 2022] Hennink, M. and Kaiser, B. N. (2022). Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social science & medicine*, 292:114523.

[Ho et al., 2022] Ho, S., Doig, G. S., and Ly, A. (2022). Attitudes of optometrists towards artificial intelligence for the diagnosis of retinal disease: A cross-sectional mail-out survey. *Ophthalmic and Physiological Optics*, 42:1170–1179.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9:1735–1780.

[Holbrook et al., 2024] Holbrook, C., Holman, D., Clingo, J., and Wagner, A. R. (2024). Overtrust in ai recommendations about whether or not to kill: Evidence from two human-robot interaction studies. *Scientific Reports*, 14:19751.

[Huang et al., 2024] Huang, W., Ong, W. C., Wong, M. K. F., Ng, E. Y. K., Koh, T., Chandramouli, C., Ng, C. T., Hummel, Y., Huang, F., Lam, C. S. P., and Tromp, J. (2024). Applying the utaut2 framework to patients' attitudes toward healthcare task shifting with artificial intelligence. *BMC Health Services Research*, 24:455.

[Hyodo et al., 2023] Hyodo, R., Saito, S., Nakano, T., Akabane, M., Kasuga, R., and Ogawa, T. (2023). Video surveillance system incorporating expert decision-making process: A case study on detecting calving signs in cattle. *arXiv preprint arXiv:2301.03926*.

[Inkpen et al., 2023] Inkpen, K., Chappidi, S., Mallari, K., Nushi, B., Ramesh, D., Michelucci, P., Mandava, V., Vepřek, L. H., and Quinn, G. (2023). Advancing human-ai complementarity: The impact of user expertise and algorithmic tuning on joint decision making. *ACM Transactions on Computer-Human Interaction*, 30:1–29.

[Inoki et al., 2020] Inoki, K., Sakamoto, T., Takamaru, H., Sekiguchi, M., Yamada, M., Matsuda, T., and Saito, Y. (2020). The diagnostic performance for colorectal neoplasms using magnified endoscopy differs between experts and novice endoscopists: A post hoc analysis. *Digestion*, 101:590–597.

[Ives et al., 1983] Ives, B., Olson, M. H., and Baroudi, J. J. (1983). The measurement of user information satisfaction. *Communications of the ACM*, 26:785–793.

[Jammal et al., 2020] Jammal, A. A., Thompson, A. C., Mariottoni, E. B., Berchuck, S. I., Urata, C. N., Estrela, T., Wakil, S. M., Costa, V. P., and Medeiros, F. A. (2020). Human versus machine: Comparing a deep learning algorithm to human gradings for detecting glaucoma on fundus photographs. *American Journal of Ophthalmology*, 211:123–131.

[Jarvis et al., 2020] Jarvis, T., Thornburg, D., Rebecca, A. M., and Teven, C. M. (2020). Artificial intelligence in plastic surgery. *Plastic and Reconstructive Surgery - Global Open*, 8:e3200.

[Jian et al., 2000a] Jian, J.-Y., Bisantz, A. M., and Drury, C. G. (2000a). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4:53–71.

[Jian et al., 2000b] Jian, J.-Y., Bisantz, A. M., and Drury, C. G. (2000b). Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, 4(1):53–71.

[Judd et al., 2024] Judd, N., Aristodemou, M., Klingberg, T., and Kievit, R. (2024). Interindividual differences in cognitive variability are ubiquitous and distinct from mean performance in a battery of eleven tasks. *Journal of Cognition*, 7:45.

[Jumper et al., 2021] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O.,

Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589.

[Kaas and Habli, 2024] Kaas, M. H. L. and Habli, I. (2024). Assuring ai safety: fallible knowledge and the gricean maxims. *AI and Ethics*.

[Kahneman, 2011] Kahneman, D. (2011). *Thinking, fast and slow*. macmillan.

[Kalatzis et al., 2023] Kalatzis, A., Rahman, S., Girishan Prabhu, V., Stanley, L., and Wittie, M. (2023). A multimodal approach to investigate the role of cognitive workload and user interfaces in human-robot collaboration. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 5–14.

[Kamal et al., 2022a] Kamal, M. S., Dey, N., Chowdhury, L., Hasan, S. I., and Santosh, K. (2022a). Explainable ai for glaucoma prediction analysis to understand risk factors in treatment planning. *IEEE Transactions on Instrumentation and Measurement*, 71:1–9.

[Kamal et al., 2022b] Kamal, M. S., Dey, N., Chowdhury, L., Hasan, S. I., and Santosh, K. (2022b). Explainable ai for glaucoma prediction analysis to understand risk factors in treatment planning. *IEEE Transactions on Instrumentation and Measurement*, 71:1–9.

[Kaplan et al., 2023] Kaplan, A. D., Kessler, T. T., Brill, J. C., and Hancock, P. A. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 65:337–359.

[Kelly et al., 2023] Kelly, S., Kaye, S.-A., and Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? a systematic review. *Telematics and Informatics*, 77:101925.

[Khalifa and Albadawy, 2024] Khalifa, M. and Albadawy, M. (2024). Ai in diagnostic imaging: Revolutionising accuracy and efficiency. *Computer Methods and Programs in Biomedicine Update*, 5:100146.

[Khan et al., 2025] Khan, M. M., Shah, N., Shaikh, N., Thabet, A., alrabayah, T., and Belkhair, S. (2025). Towards secure and trusted ai in healthcare: A systematic review of emerging innovations and ethical challenges. *International Journal of Medical Informatics*, 195:105780.

[Kim et al., 2024] Kim, Y. J., Choi, J. H., and Fotso, G. M. N. (2024). Medical professionals' adoption of ai-based medical devices: Utaut model with trust mediation. *Journal of Open Innovation: Technology, Market, and Complexity*, 10:100220.

[Klein et al., 1989] Klein, G., Calderwood, R., and MacGregor, D. (1989). Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics*, 19:462–472.

[Klein, 1993] Klein, G. A. (1993). A recognition-primed decision (rpd) model of rapid decision making. In Klein, G. A., Orasanu, J., Calderwood, R., and Zsambok, C. E., editors, *Decision Making in Action: Models and Methods*, pages 138–147. Ablex Publishing, Norwood, NJ.

[Klein, 2017] Klein, G. A. (2017). *Sources of power: How people make decisions*. MIT press.

[Knop et al., 2022] Knop, M., Weber, S., Mueller, M., and Niehaves, B. (2022). Human factors and technological characteristics influencing the interaction of medical professionals with artificial intelligence–enabled clinical decision support systems: Literature review. *JMIR Human Factors*, 9:e28639.

[Kohn et al., 2021] Kohn, S. C., De Visser, E. J., Wiese, E., Lee, Y.-C., and Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in psychology*, 12:604977.

[Koklanis and Thorburn, 2021] Koklanis, K. and Thorburn, D. (2021). Orthoptist-led glaucoma monitoring: A pilot quality assurance audit of the reliability of orthoptists in glaucoma management decisions. *Australian Orthoptic Journal*, 53:9–15.

[Körber, 2019] Körber, M. (2019). Theoretical considerations and development of a questionnaire to measure trust in automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20*, pages 13–30. Springer.

[Kostick-Quenet et al., 2024] Kostick-Quenet, K., Lang, B. H., Smith, J., Hurley, M., and Blumenthal-Barby, J. (2024). Trust criteria for artificial intelligence in health: normative and epistemic considerations. *Journal of Medical Ethics*, 50:544–551.

[Krishnan et al., 2025] Krishnan, A., Dutta, A., Srivastava, A., Konda, N., and Prakasam, R. (2025). Artificial intelligence in optometry: Current and future perspectives. *Clinical Optometry*, Volume 17:83–114.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, pages 1097–1105. Curran Associates Inc.

[Kumar et al., 2023] Kumar, V. V. N. S., Reddy, G. H., and GiriPrasad, M. (2023). A novel glaucoma detection model using unet++-based segmentation and resnet with gru-based optimized deep learning. *Biomedical Signal Processing and Control*, 86:105069.

[Lai et al., 2023] Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., and Tan, C. (2023). Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1369–1385. ACM.

[Le et al., 2003] Le, A., Mukesh, B. N., McCarty, C. A., and Taylor, H. R. (2003). Risk factors associated with the incidence of open-angle glaucoma: The visual impairment project. *Investigative Opthalmology & Visual Science*, 44:3783.

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–444.

[LeCun et al., 1989] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551.

[Lecun et al., 1998] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324.

[LEE and MORAY, 1992] LEE, J. and MORAY, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35:1243–1270.

[Lee and Moray, 1994] Lee, J. D. and Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, 40:153–184.

[Lee and See, 2004] Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46:50–80.

[Lee and Chew, 2023] Lee, M. H. and Chew, C. J. (2023). Understanding the effect of counterfactual explanations on trust and reliance on ai for human-ai collaborative clinical decision making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–22.

[Leon, 2020] Leon, S. (2020). Ux in ai: Trust in algorithm-based investment decisions. *Junior Management Science*.

[Leske et al., 2008] Leske, M. C., Wu, S.-Y., Hennis, A., Honkanen, R., Nemesure, B., Group, B. S., et al. (2008). Risk factors for incident open-angle glaucoma: the barbados eye studies. *Ophthalmology*, 115(1):85–93.

[Li et al., 2018] Li, Z., He, Y., Keel, S., Meng, W., Chang, R. T., and He, M. (2018). Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*, 125:1199–1206.

[Lih et al., 2020] Lih, O. S., Jahmunah, V., San, T. R., Ciaccio, E. J., Yamakawa, T., Tanabe, M., Kobayashi, M., Faust, O., and Acharya, U. R. (2020). Comprehensive electrocardiographic diagnosis based on deep learning. *Artificial intelligence in medicine*, 103:101789.

[Lim et al., 2022] Lim, W. S., Ho, H.-Y., Ho, H.-C., Chen, Y.-W., Lee, C.-K., Chen, P.-J., Lai, F., Jang, J.-S. R., and Ko, M.-L. (2022). Use of multimodal dataset in ai for detecting glaucoma based on fundus photographs assessed with oct: focus group study on high prevalence of myopia. *BMC Medical Imaging*, 22:206.

[Lin et al., 2022] Lin, M., Hou, B., Liu, L., Gordon, M., Kass, M., Wang, F., Tassel, S. H. V., and Peng, Y. (2022). Automated diagnosing primary open-angle glaucoma from fundus image by simulating human's grading with deep learning. *Scientific Reports*, 12:14080.

[Lukyanenko et al., 2022] Lukyanenko, R., Maass, W., and Storey, V. C. (2022). Trust in artificial intelligence: From a foundational trust framework to emerging research opportunities. *Electronic Markets*, 32:1993–2020.

[MacFarlane et al., 2022] MacFarlane, E., Stevens, A. C., North, R., Ryan, B., and Acton, J. (2022). A mixed-methods characterisation of patient safety incidents by primary eye care practitioners. *Ophthalmic and Physiological Optics*, 42:1304–1315.

[Madhavan and Wiegmann, 2007] Madhavan, P. and Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, 8:277–301.

[Maish, 1979] Maish, A. M. (1979). A user's behavior toward his mis. *MIS Quarterly*, 3:39.

[Malle and Ullman, 2021] Malle, B. F. and Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In *Trust in human-robot interaction*, pages 3–25. Elsevier.

[Mann and Whitney, 1947] Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

[Marks et al., 2012] Marks, J. R., Harding, A. K., Harper, R. A., Williams, E., Haque, S., Spencer, A. F., and Fenerty, C. (2012). Agreement between specially trained and

accredited optometrists and glaucoma specialist consultant ophthalmologists in their management of glaucoma patients. *Eye*, 26:853–861.

[Mayer et al., 1995a] Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995a). An integrative model of organizational trust. *Academy of Management Review*, 20:709–734.

[Mayer et al., 1995b] Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995b). An integrative model of organizational trust. *Academy of management review*, 20(3):709–734.

[McCarthy et al., 1955] McCarthy, J., Minsky, M., Rochester, N., and Shannon, C. (1955). A proposal for the dartmouth summer research project on artificial intelligence, www-formal.

[McCartney, 2017] McCartney, M. (2017). Margaret mccartney: Nurses must be allowed to exercise professional judgment. *BMJ*, page j1548.

[Merriam, 2024] Merriam (2024). Trust [def. 1]. Retrieved November 23, 2024, from http://www.merriam-webster.com/dictionary/trust.

[Merriam-Webster, 2024] Merriam-Webster (2024). Acceptance [def. 2]. Retrieved December 6, 2024, from https://www.merriam-webster.com/dictionary/acceptance.

[Miller, 2019a] Miller, T. (2019a). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

[Miller, 2019b] Miller, T. (2019b). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

[Mishra et al., 2024] Mishra, A. K., Tyagi, A. K., and Arowolo, M. O. (2024). *Future Trends and Opportunities in Machine Learning and Artificial Intelligence for Banking and Finance*, pages 211–238. Springer.

[Mitchell, 1997] Mitchell, T. M. (1997). Does machine learning really work? *AI magazine*, 18(3):11–11.

[Mittal et al., 2024] Mittal, M., Gupta, V., Aamash, M., and Upadhyay, T. (2024). Machine learning for pest detection and infestation prediction: A comprehensive review. *WIREs Data Mining and Knowledge Discovery*, 14.

[Moreno-Montañés et al., 2017] Moreno-Montañés, J., Antón, V., Antón, A., Larrosa, J. M., de-la Casa, J. M. M., Rebolleda, G., Ussa, F., and García-Granero, M. (2017). Intraobserver and interobserver agreement of structural and functional software programs for measuring glaucoma progression. *JAMA Ophthalmology*, 135:313.

[MUIR, 1994] MUIR, B. M. (1994). Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37:1905–1922.

[MUIR and MORAY, 1996] MUIR, B. M. and MORAY, N. (1996). Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39:429–460.

[Mursch-Edlmayr et al., 2020] Mursch-Edlmayr, A. S., Ng, W. S., Diniz-Filho, A., Sousa, D. C., Arnould, L., Schlenker, M. B., Duenas-Angeles, K., Keane, P. A., Crowston, J. G., and Jayaram, H. (2020). Artificial intelligence algorithms to diagnose glaucoma and detect glaucoma progression: Translation to clinical practice. *Translational Vision Science & Technology*, 9:55.

[Muzahid et al., 2023] Muzahid, A. J. M., Kamarulzaman, S. F., Rahman, M. A., Murad, S. A., Kamal, M. A. S., and Alenezi, A. H. (2023). Multiple vehicle cooperation and collision avoidance in automated vehicles: survey and an ai-enabled conceptual framework. *Scientific Reports*, 13:603.

[Natarajan and Augustin, 2024] Natarajan, E. and Augustin, F. (2024). A bipolar fuzzy decision-making system for assessing high-risk coexisting tuberculosis disease in pregnant women. *Heliyon*, 10:e31416.

[Nicolaou et al., 2013] Nicolaou, A. I., Ibrahim, M., and van Heck, E. (2013). Information quality, trust, and risk perceptions in electronic data exchanges. *Decision Support Systems*, 54:986–996.

[Nida et al., 2021] Nida, E. K., Bekele, S., Geurts, L., and Abeele, V. V. (2021). Acceptance of a smartphone-based visual field screening platform for glaucoma: Pre-post study. *JMIR Formative Research*, 5:e26602.

[Nilsson, 2009] Nilsson, N. J. (2009). *The Quest for Artificial Intelligence*. Cambridge University Press.

[Nirala et al., 2022] Nirala, K. K., Singh, N. K., and Purani, V. S. (2022). A survey on providing customer and public administration based services using ai: chatbot. *Multimedia Tools and Applications*, 81:22215–22246.

[Nouhan et al., 2021] Nouhan, C., Scott, N., and Womack, J. (2021). Emergent role of artificial intelligence in higher education. *IEEE Technology Policy and Ethics*, 6:1–5.

[Nourani et al., 2020] Nourani, M., King, J., and Ragan, E. (2020). The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In

Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, volume 8, pages 112–121.

[Ntoutsi et al., 2020] Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T., and Staab, S. (2020). Bias in data-driven artificial intelligence systems—an introductory survey. *WIREs Data Mining and Knowledge Discovery*, 10.

[Ogawa et al., 2023] Ogawa, R., Shima, S., Takemura, T., and ichi Fukuzumi, S. (2023). *A Study on Trust Building in AI Systems Through User Commitment*, pages 557–567.

[Oliver, 1980] Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research*, 17:460.

[Organization, 2019] Organization, W. H. (2019). World report on vision.

[Ostrom et al., 2019] Ostrom, A. L., Fotheringham, D., and Bitner, M. J. (2019). *Customer Acceptance of AI in Service Encounters: Understanding Antecedents and Consequences*, pages 77–103.

[Oxford, 2024a] Oxford (2024a). Acceptance [def. 1]. Retrieved December 6, 2024, from https://www.oxfordlearnersdictionaries.com/definition/english/acceptance?q=acceptance.

[Oxford, 2024b] Oxford (2024b). Trust [def. 2]. Retrieved November 23, 2024, from https://www.oed.com/search/dictionary/?scope=Entries&q=trust.

[Pan et al., 2023] Pan, M., Wang, J., and Wang, J. (2023). Application of artificial intelligence in education: Opportunities, challenges, and suggestions. In *2023 13th International Conference on Information Technology in Medicine and Education (ITME)*, pages 623–627. IEEE.

[Papagni et al., 2023] Papagni, G., de Pagter, J., Zafari, S., Filzmoser, M., and Koeszegi, S. T. (2023). Artificial agents' explainability to support trust: considerations on timing and context. *AI & SOCIETY*, 38:947–960.

[Parasuraman and Manzey, 2010] Parasuraman, R. and Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 52:381–410.

[Parasuraman and Riley, 1997] Parasuraman, R. and Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39:230–253.

[Pareek et al., 2024] Pareek, S., Velloso, E., and Goncalves, J. (2024). Trust development and repair in ai-assisted decision-making during complementary expertise. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 546–561.

[Parimbelli et al., 2018] Parimbelli, E., Marini, S., Sacchi, L., and Bellazzi, R. (2018). Patient similarity for precision medicine: A systematic review. *Journal of Biomedical Informatics*, 83:87–96.

[Park et al., 2020] Park, H. S., Kim, K. I., Soh, J. Y., Hyun, Y. H., Jang, S. K., Lee, S., Hwang, G. Y., and Kim, H. S. (2020). Factors influencing acceptance of personal health

record apps for workplace health promotion: Cross-sectional questionnaire study. *JMIR mHealth and uHealth*, 8:e16723.

[Pierre et al., 2023] Pierre, K., Haneberg, A. G., Kwak, S., Peters, K. R., Hochhegger, B., Sananmuang, T., Tunlayadechanont, P., Tighe, P. J., Mancuso, A., and Forghani, R. (2023). Applications of artificial intelligence in the radiology roundtrip: Process streamlining, workflow optimization, and beyond. *Seminars in Roentgenology*, 58:158–169.

[Pourjavan et al., 2024] Pourjavan, S., Bourguignon, G.-H., Marinescu, C., Otjacques, L., and Boschi, A. (2024). Evaluating the influence of clinical data on inter-observer variability in optic disc analysis for ai-assisted glaucoma screening. *Clinical Ophthalmology*, Volume 18:3999–4009.

[Poursabzi-Sangdeh et al., 2021a] Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W. W., and Wallach, H. (2021a). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–52. ACM.

[Poursabzi-Sangdeh et al., 2021b] Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W. W., and Wallach, H. (2021b). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–52. ACM.

[Qidwai et al., 2023] Qidwai, U., Qidwai, U., Sivapalan, T., and Ratnarajan, G. (2023). imigs: An innovative ai based prediction system for selecting the best patient-specific glaucoma treatment. *MethodsX*, 10:102209.

[Raja et al., 2025] Raja, H., Huang, X., Delsoz, M., Madadi, Y., Poursoroush, A., Munawar, A., Kahook, M. Y., and Yousefi, S. (2025). Diagnosing glaucoma based on the ocular hypertension treatment study dataset using chat generative pre-trained transformer as a large language model. *Ophthalmology Science*, 5(1):100599.

[Rasmussen, 1983] Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13:257–266.

[Rau, 2011] Rau, H. A. (2011). Trust and trustworthiness: A survey of gender differences. *Psychology of Gender Differences, Forthcoming*.

[Resnikoff et al., 2012] Resnikoff, S., Felch, W., Gauthier, T.-M., and Spivey, B. (2012). The number of ophthalmologists in practice and training worldwide: a growing gap despite more than 200000 practitioners. *British Journal of Ophthalmology*, 96:783–787.

[Reus et al., 2010] Reus, N. J., Lemij, H. G., Garway-Heath, D. F., Airaksinen, P. J., Anton, A., Bron, A. M., Faschinger, C., Holló, G., Iester, M., Jonas, J. B., Mistlberger, A., Topouzis, F., and Zeyen, T. G. (2010). Clinical assessment of stereoscopic optic disc photographs for glaucoma: The european optic disc assessment trial. *Ophthalmology*, 117:717–723.

[Rogers et al., 2019] Rogers, T. W., Jaccard, N., Carbonaro, F., Lemij, H. G., Vermeer, K. A., Reus, N. J., and Trikha, S. (2019). Evaluation of an ai system for the automated detection of glaucoma from stereoscopic optic disc photographs: the european optic disc assessment study. *Eye*, 33:1791–1797.

[Rousseau et al., 1998] Rousseau, D. M., Sitkin, S. B., Burt, R. S., and Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, 23:393–404.

[Russell and Norvig, 2021] Russell, S. and Norvig, P. (2021). Artificial intelligence: A modern approach, 4th, global edition pearson. *Harlow, UK*.

[Sachan et al., 2024] Sachan, S., Almaghrabi, F., Yang, J.-B., and Xu, D.-L. (2024). Human-ai collaboration to mitigate decision noise in financial underwriting: A study on fintech innovation in a lending firm. *International Review of Financial Analysis*, 93:103149.

[Sarhan et al., 2019] Sarhan, A., Rokne, J., and Alhajj, R. (2019). Glaucoma detection using image processing techniques: A literature review. *Computerized Medical Imaging and Graphics*, 78:101657.

[Sarsfield, 2014] Sarsfield, E. (2014). Differences between novices' and experts' solving ill-structured problems. *Public Health Nursing*, 31(5):444–453.

[Saxena et al., 2020] Saxena, A., Vyas, A., Parashar, L., and Singh, U. (2020). A glaucoma detection using convolutional neural network. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 815–820. IEEE.

[Scanzera et al., 2022] Scanzera, A. C., Shorter, E., Kinnaird, C., Valikodath, N., Al-Khaled, T., Cole, E., Kravets, S., Hallak, J. A., McMahon, T., and Chan, R. P. (2022). Optometrist's perspectives of artificial intelligence in eye care. *Journal of Optometry*, 15:S91–S97.

[Scheetz et al., 2019] Scheetz, J., Koklanis, K., McGuinness, M., Long, M., and Morris, M. E. (2019). A randomized trial to increase the assessment accuracy of glaucoma and

optic disc characteristics by orthoptists. *Journal of Continuing Education in the Health Professions*, 39:161–167.

[Schunk, 2012] Schunk, D. H. (2012). *Learning theories an educational perspective*. Pearson Education, Inc.

[Schuster et al., 2020] Schuster, A. K., Erb, C., Hoffmann, E. M., Dietlein, T., and Pfeiffer, N. (2020). The diagnosis and treatment of glaucoma. *Deutsches Ärzteblatt international*.

[Shalaby et al., 2022] Shalaby, W. S., Ahmed, O. M., Waisbourd, M., and Katz, L. J. (2022). A review of potential novel glaucoma therapeutic options independent of intraocular pressure. *Survey of Ophthalmology*, 67:1062–1080.

[Shevtsova et al., 2024] Shevtsova, D., Ahmed, A., Boot, I. W. A., Sanges, C., Hudecek, M., Jacobs, J. J. L., Hort, S., and Vrijhoef, H. J. M. (2024). Trust in and acceptance of artificial intelligence applications in medicine: Mixed methods study. *JMIR Human Factors*, 11:e47031.

[Shneiderman, 2020] Shneiderman, B. (2020). Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction*, pages 109–124.

[Silver et al., 2016] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489.

[Silver et al., 2017] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F.,

Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, 550:354–359.

[Singh et al., 2022] Singh, L. K., Pooja, Garg, H., and Khanna, M. (2022). Deep learning system applicability for rapid glaucoma prediction from fundus images across various data sets. *Evolving Systems*, 13:807–836.

[Singh et al., 2020] Singh, S. P., Wang, L., Gupta, S., Goli, H., Padmanabhan, P., and Gulyás, B. (2020). 3d deep learning on medical images: A review. *Sensors*, 20:5097.

[Sivaraman et al., 2023] Sivaraman, V., Bukowski, L. A., Levin, J., Kahn, J. M., and Perer, A. (2023). 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*, pages 1–18. ACM.

[Soh et al., 2021] Soh, Z., Yu, M., Betzler, B. K., Majithia, S., Thakur, S., Tham, Y. C., Wong, T. Y., Aung, T., Friedman, D. S., and Cheng, C.-Y. (2021). The global extent of undetected glaucoma in adults. *Ophthalmology*, 128:1393–1404.

[St. Pierre and Nyce, 2020] St. Pierre, M. and Nyce, J. M. (2020). How novice and expert anaesthetists understand expertise in anaesthesia: a qualitative study. *BMC medical education*, 20:1–11.

[Stainer et al., 2015] Stainer, M. J., Anderson, A. J., and Denniss, J. (2015). Examination strategies of experienced and novice clinicians viewing the retina. *Ophthalmic and Physiological Optics*, 35:424–432.

[Stevens and Stetson, 2023] Stevens, A. F. and Stetson, P. (2023). Theory of trust and acceptance of artificial intelligence technology (traait): An instrument to assess clini-

cian trust and acceptance of artificial intelligence. *Journal of Biomedical Informatics*, 148:104550.

[Sun et al., 2016] Sun, W., Zheng, B., and Qian, W. (2016). Computer aided lung cancer diagnosis with deep learning algorithms. In *Medical imaging 2016: computer-aided diagnosis*, volume 9785, page 97850Z. International Society for Optics and Photonics.

[Sun et al., 2022] Sun, Y., Chen, A., Zou, M., Zhang, Y., Jin, L., Li, Y., Zheng, D., Jin, G., and Congdon, N. (2022). Time trends, associations and prevalence of blindness and vision loss due to glaucoma: an analysis of observational data from the global burden of disease study 2017. *BMJ Open*, 12:e053805.

[Sun et al., 2023] Sun, Z., Lin, M., Zhu, Q., Xie, Q., Wang, F., Lu, Z., and Peng, Y. (2023). A scoping review on multimodal deep learning in biomedical images and texts. *Journal of Biomedical Informatics*, 146:104482.

[Swanson, 1974] Swanson, E. B. (1974). Management information systems: Appreciation and involvement. *Management Science*, 21:178–188.

[Swapna et al., 2018] Swapna, G., Vinayakumar, R., and Soman, K. (2018). Diabetes detection using deep learning algorithms. *ICT express*, 4(4):243–246.

[Sweller, 1988] Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12:257–285.

[Taddeo and Floridi, 2018] Taddeo, M. and Floridi, L. (2018). How ai can be a force for good. *Science*, 361:751–752.

[Taylor and Todd, 1995] Taylor, S. and Todd, P. A. (1995). Understanding information technology usage: A test of competing models. *Information Systems Research*, 6:144–176.

[Textor et al., 2022] Textor, C., Zhang, R., Lopez, J., Schelble, B. G., McNeese, N. J., Freeman, G., Pak, R., Tossell, C., and de Visser, E. J. (2022). Exploring the relationship between ethics and trust in human–artificial intelligence teaming: A mixed methods approach. *Journal of Cognitive Engineering and Decision Making*, 16:252–281.

[Tilden and Tilden, 1985] Tilden, V. P. and Tilden, S. (1985). Review of *from novice to expert: Excellence and power in clinical nursing practice* by patricia benner. *Research in Nursing and Health*, 8(1):95–97. Book review.

[Ting et al., 2019] Ting, D. S. W., Pasquale, L. R., Peng, L., Campbell, J. P., Lee, A. Y., Raman, R., Tan, G. S. W., Schmetterer, L., Keane, P. A., and Wong, T. Y. (2019). Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*, 103:167–175.

[Tomsett et al., 2020] Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., and Kaplan, L. (2020). Rapid trust calibration through interpretable and uncertainty-aware ai. *Patterns*, 1:100049.

[Topouzis and Anastasopoulos, 2007] Topouzis, F. and Anastasopoulos, E. (2007). Glaucoma — the importance of early detection and early treatment. *European Ophthalmic Review*, 00:13.

[Tran et al., 2021] Tran, A. Q., Nguyen, L. H., Nguyen, H. S. A., Nguyen, C. T., Vu, L. G., Zhang, M., Vu, T. M. T., Nguyen, S. H., Tran, B. X., Latkin, C. A., Ho, R. C. M.,

and Ho, C. S. H. (2021). Determinants of intention to use artificial intelligence-based diagnosis support system among prospective physicians. *Frontiers in Public Health*, 9.

[Tsang et al., 2023] Tsang, M., Martin, L., Blissett, S., Gauthier, S., Ahmed, Z., Muhammed, D., and Sibbald, M. (2023). What do clinicians mean by "good clinical judgment": A qualitative study. *International Medical Education*, 2:1–10.

[Tsatsos et al., 2024] Tsatsos, M., Rodafinos, A., Athanasiadis, I. K., and Mavropoulou, D. (2024). Patients' intentions to use telemedicine for ophthalmic medical care during the covid-19 pandemic. *Cureus*.

[Tucci et al., 2022a] Tucci, V., Saary, J., and Doyle, T. E. (2022a). Factors influencing trust in medical artificial intelligence for healthcare professionals: a narrative review. *Journal of Medical Artificial Intelligence*, 5:4–4.

[Tucci et al., 2022b] Tucci, V., Saary, J., and Doyle, T. E. (2022b). Factors influencing trust in medical artificial intelligence for healthcare professionals: a narrative review. *Journal of Medical Artificial Intelligence*, 5.

[Turing, 2009] Turing, A. M. (2009). *Computing machinery and intelligence*. Springer.

[Uereten, 2023] Uereten, E. (2023). Models of cognitive work analysis in neurocritical care and perspectives on expertise–an ecological interface design approach.

[Valik et al., 2023] Valik, J. K., Ward, L., Tanushi, H., Johansson, A. F., Färnert, A., Mogensen, M. L., Pickering, B. W., Herasevich, V., Dalianis, H., Henriksson, A., and Nauclér, P. (2023). Predicting sepsis onset using a machine learned causal probabilistic network algorithm based on electronic health records data. *Scientific Reports*, 13:11760.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010. Curran Associates Inc.

[Venkatesh et al., 2003] Venkatesh, Morris, Davis, and Davis (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27:425.

[Venkatesh et al., 2012] Venkatesh, Thong, and Xu (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, 36:157.

[Venkatesh and Davis, 2000] Venkatesh, V. and Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, 46:186–204.

[Vessey, 1991] Vessey, I. (1991). Cognitive fit: A theory-based analysis of the graphs versus tables literature*. *Decision Sciences*, 22:219–240.

[Wakchaure et al., 2023] Wakchaure, M., Patle, B., and Mahindrakar, A. (2023). Application of ai techniques and robotics in agriculture: A review. *Artificial Intelligence in the Life Sciences*, 3:100057.

[Wang et al., 2024] Wang, C., Chen, D.-F., Shang, X., Wang, X., Chu, X., Hu, C., Huang, Q., Cheng, G., Li, J., Ren, R., and Liang, Y. (2024). Evaluating diagnostic concordance in primary open-angle glaucoma among academic glaucoma subspecialists. *Diagnostics*, 14:2460.

[Wang et al., 2021] Wang, G., Liu, X., Shen, J., Wang, C., Li, Z., Ye, L., Wu, X., Chen, T., Wang, K., Zhang, X., Zhou, Z., Yang, J., Sang, Y., Deng, R., Liang, W., Yu, T.,

Gao, M., Wang, J., Yang, Z., Cai, H., Lu, G., Zhang, L., Yang, L., Xu, W., Wang, W., Olvera, A., Ziyar, I., Zhang, C., Li, O., Liao, W., Liu, J., Chen, W., Chen, W., Shi, J., Zheng, L., Zhang, L., Yan, Z., Zou, X., Lin, G., Cao, G., Lau, L. L., Mo, L., Liang, Y., Roberts, M., Sala, E., Schönlieb, C.-B., Fok, M., Lau, J. Y.-N., Xu, T., He, J., Zhang, K., Li, W., and Lin, T. (2021). A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and covid-19 pneumonia from chest x-ray images. *Nature Biomedical Engineering*, 5:509–521.

[Wang and Strong, 1996] Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33.

[Wanner et al., 2022] Wanner, J., Herm, L.-V., Heinrich, K., and Janiesch, C. (2022). The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study. *Electronic Markets*, 32:2079–2102.

[Welch, 1947] Welch, B. L. (1947). The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34:28–35.

[Wilde, 1982] Wilde, G. J. S. (1982). The theory of risk homeostasis: Implications for safety and health. *Risk Analysis*, 2:209–225.

[Wojton et al., 2020] Wojton, H. M., Porter, D., T. Lane, S., Bieber, C., and Madhavan, P. (2020). Initial validation of the trust of automated systems test (toast). *The Journal of social psychology*, 160(6):735–750.

[Xiong et al., 2022] Xiong, J., Li, F., Song, D., Tang, G., He, J., Gao, K., Zhang, H., Cheng, W., Song, Y., Lin, F., Hu, K., Wang, P., Li, J.-P. O., Aung, T., Qiao, Y., Zhang,

X., and Ting, D. (2022). Multimodal machine learning using visual fields and peripapillary circular oct scans in detection of glaucomatous optic neuropathy. *Ophthalmology*, 129:171–180.

[Xue et al., 2022] Xue, Y., Zhu, J., Huang, X., Xu, X., Li, X., Zheng, Y., Zhu, Z., Jin, K., Ye, J., Gong, W., and Si, K. (2022). A multi-feature deep learning system to enhance glaucoma severity diagnosis with high accuracy and fast speed. *Journal of Biomedical Informatics*, 136:104233.

[Yang et al., 2012] Yang, H., Thompson, C., and Bland, M. (2012). The effect of clinical experience, judgment task difficulty and time pressure on nurses' confidence calibration in a high fidelity clinical simulation. *BMC Medical Informatics and Decision Making*, 12:113.

[Yang et al., 2020] Yang, Q., Steinfeld, A., Rosé, C., and Zimmerman, J. (2020). Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13. ACM.

[Yang et al., 2023] Yang, X. J., Schemanske, C., and Searle, C. (2023). Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 65:862–878.

[Yi et al., 2006] Yi, M. Y., Jackson, J. D., Park, J. S., and Probst, J. C. (2006). Understanding information technology acceptance by individual professionals: Toward an integrative view. *Information & Management*, 43:350–363.

[Yoshioka et al., 2015] Yoshioka, N., Wong, E., Kalloniatis, M., Yapp, M., Hennessy, M. P., Agar, A., Healey, P. R., Hayen, A., and Zangerl, B. (2015). Influence of education and diagnostic modes on glaucoma assessment by optometrists. *Ophthalmic and Physiological Optics*, 35:682–698.

[Zerilli et al., 2022] Zerilli, J., Bhatt, U., and Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns*, 3. doi: 10.1016/j.patter.2022.100455.

[Zhang et al., 2021] Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J. C., and Sellitto, M. (2021). The ai index 2021 annual report. *arXiv preprint arXiv:2103.06312*.

[Zhang et al., 2024] Zhang, S., Yu, J., Xu, X., Yin, C., Lu, Y., Yao, B., Tory, M., Padilla, L. M., Caterino, J., Zhang, P., and Wang, D. (2024). Rethinking human-ai collaboration in complex medical decision making: A case study in sepsis diagnosis. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18. ACM.

[Zhu et al., 2024] Zhu, Y., Salowe, R., Chow, C., Li, S., Bastani, O., and O'Brien, J. M. (2024). Advancing glaucoma care: Integrating artificial intelligence in diagnosis, management, and progression detection. *Bioengineering*, 11:122.

[Zolanvari et al., 2023] Zolanvari, M., Yang, Z., Khan, K., Jain, R., and Meskin, N. (2023). Trust xai: Model-agnostic explanations for ai with a case study on iiot security. *IEEE Internet of Things Journal*, 10:2967–2978.

APPENDICES

Appendix A

Variabilty study additional data

A.1 Ethics application for varibility study

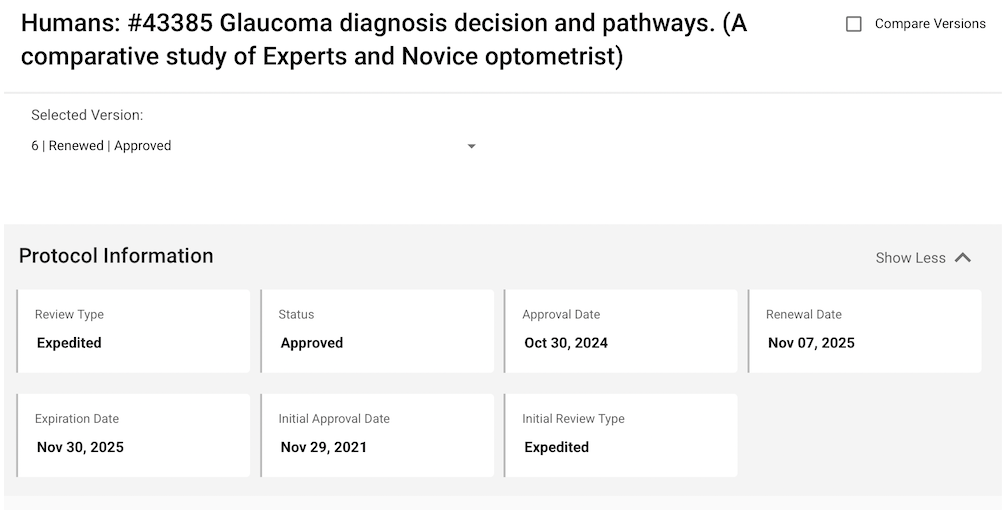


Figure A.1: The figure shows the accepted ethics application by Delegated Research Ethics Committee (DERC) of University of Waterloo

A.2 Participants consent letter for variability study



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 (A comparative study of Novice and Expert Optometrists) *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 #43385). If you have questions for the Board, contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 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: _____

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 (A comparative study of Novice and Expert Optometrists) *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 #43385). If you have questions for the Board, contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 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: _____

Figure A.2: Enter Caption

A.3 Demographic questionnaire

You may decline to answer any of the questions and/or share your personal information. You may also withdraw from this study at any time by notifying the researcher. If you choose to withdraw, your responses will be erased.

1. **What gender do you identify as?**

    A. Man

    B. Woman

    C. Trans* / Non-binary

    D. Prefer to self-identify (please specify): _____

    E. Prefer not to answer

2. **What is your age?**

    A. 20–30 years old

    B. 31–40 years old

    C. 41–50 years old

    D. 51–60 years old

    E. 61+

    F. Prefer not to answer

3. **Please specify your ethnicity.**

    A. Indigenous (First Nations, Inuit, Métis)

B. European

C. Asian (South, East, West)

D. African

E. Latin American, Central or South American

F. Caribbean

G. Oceania (New Zealand, Pacific Islands)

H. Multiracial / Multiethnic (two or more)

I. Prefer to self-identify (please specify): _____

J. Prefer not to answer

4. **What is the highest level of education you have completed?**

A. Bachelor's Degree

B. Master's Degree

C. Ph.D. or higher

D. Prefer not to answer

5. **Do you have any specific glaucoma-related degree, certificate, or training?**

A. Yes

B. No

C. Prefer not to answer

D. If yes, please provide details: _____

A.4 Questions asked in the interview

1. Do you think that this patient has glaucoma, is a glaucoma suspect, or doesn't have glaucoma at all?
2. Why do you think that this patient has/does not have glaucoma?
3. Why do you think this patient is/is not a glaucoma suspect?
4. How certain are you of your diagnosis on a scale of 1 to 5, with 5 representing a high level of certainty and 1 being a low level of certainty?
5. How would you stage this patient's level of glaucoma – early/mild, moderate, or severe/advanced?
6. Why do you think that the patient has this stage of glaucoma?
7. List the attributes based on which you have taken this decision, in order of priority.
8. When do you think this patient should be seen for a follow-up appointment?
9. Why did you choose that time frame? Why not have the patient visit in a different number of days?
10. Will there be any change in your decision if a specific test value changes?
11. How would you treat and manage this patient?
12. If we develop an AI system, what data should it use?
13. If we develop an AI system, what output should it provide to physicians?
14. If we develop an AI system, what data should it display to assist physicians in decision-making?

A.5 Thematic categorization of factors in glaucoma assessment

Category	Details
Family History	Patients having or not having any family history
Medical History	Blood pressure, diabetes, laser treatments, myopia, sleep apnea, any other treatments
Patient Background	Age, ethnicity
Clinical Findings	Anterior segment, cataracts, iris, pupil, VA aided/unaided
Optic Nerve Exam	Fundus photo analysis, CD ratio, cupping, posterior segment, hemorrhage analysis, ISNT rule, rim width analysis, OCT findings, RNFL analysis, symmetry findings
Optic Nerve Function Exam	Visual field findings
Structure-Function Correlation	OCT and VF correlation, structural and functional vision correlation
Progression and Change Analysis	Progression on OCT, changes in the eyes, color change analysis, defect repeatability analysis, previous visit comparisons
Other Risk Factors	Corneal thickness analysis, gonioscopy analysis, intraocular pressure
Data Factors	No data available for comparison, unreliable data, not enough data points
External Factors	Age-related considerations, comorbidity considerations, family history considerations, race and ethnicity considerations

A.6 Descriptive statistics for thematic assessments across visits

Assessment	Visit number	Mean	Std. Error of Mean	95% CI Mean Upper	95% CI Mean Lower	Std. Deviation
Family history	First Visit	0.788	0.072	0.93	0.646	0.415
	Subsequent Visit	0.364	0.07	0.502	0.226	0.522
Medical history	First Visit	1.061	0.123	1.301	0.82	0.704
	Subsequent Visit	0.444	0.086	0.614	0.275	0.634
Patient background	First Visit	0.788	0.072	0.93	0.646	0.415
	Subsequent Visit	0.296	0.063	0.419	0.173	0.461
Clinical findings	First Visit	0.97	0.192	1.346	0.593	1.104
	Subsequent Visit	0.509	0.088	0.682	0.337	0.639
Optic nerve exam	First Visit	3.065	0.42	3.887	2.242	2.337
	Subsequent Visit	1.72	0.121	1.958	1.482	0.858
Optic nerve function exam	First Visit	1.613	0.178	1.961	1.265	0.989
	Subsequent Visit	1.407	0.158	1.716	1.099	1.158
Correlation	First Visit	0.061	0.042	0.143	-0.022	0.242
	Subsequent Visit	0.127	0.058	0.242	0.013	0.433
Progression and Change	First Visit	0.152	0.077	0.302	8.223×10^{-4}	0.442
	Subsequent Visit	0.352	0.084	0.517	0.187	0.619
Other risk factors	First Visit	2.375	0.14	2.65	2.1	0.793
	Subsequent Visit	1.512	0.1	1.707	1.317	0.637

Figure A.3: Novices' descriptive statistics for first visits and subsequent visits

Assessment	Visit number	Mean	Std. Error of Mean	95% CI Mean Upper	95% CI Mean Lower	Std. Deviation
Family history	First Visit	0.778	0.101	0.975	0.58	0.428
	Subsequent Visit	0.323	0.085	0.49	0.155	0.475
Medical history	First Visit	1.167	0.167	1.493	0.84	0.707
	Subsequent Visit	0.452	0.121	0.689	0.214	0.675
Patient background	First Visit	0.944	0.056	1.053	0.836	0.236
	Subsequent Visit	0.552	0.094	0.736	0.368	0.506
Clinical findings	First Visit	1.118	0.169	1.449	0.786	0.697
	Subsequent Visit	0.897	0.135	1.16	0.633	0.724
Optic nerve exam	First Visit	2.765	0.504	3.752	1.777	2.078
	Subsequent Visit	1.926	0.261	2.438	1.414	1.357
Optic nerve function exam	First Visit	1.529	0.244	2.008	1.051	1.007
	Subsequent Visit	1.179	0.212	1.595	0.762	1.124
Correlation	First Visit	0.111	0.076	0.261	-0.038	0.323
	Subsequent Visit	0.226	0.076	0.375	0.076	0.425
Progression and Change	First Visit	0.167	0.121	0.404	-0.071	0.514
	Subsequent Visit	1.29	0.192	1.667	0.913	1.071
Other risk factors	First Visit	2.222	0.191	2.596	1.849	0.808
	Subsequent Visit	1.667	0.162	1.984	1.35	0.686

Figure A.4: Experts' descriptive statistics for first visits and subsequent visits

Appendix B

Modality study additional data

B.1 Ethics application for modality study

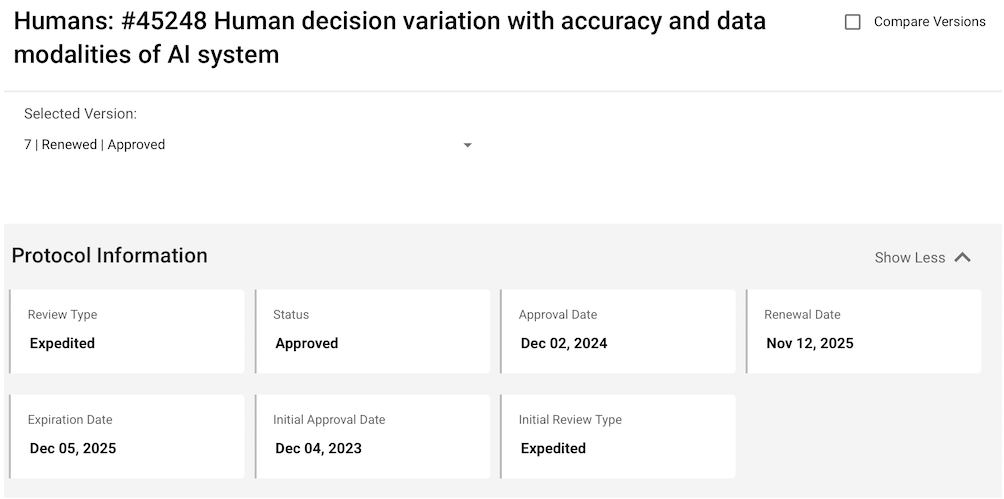


Figure B.1: The figure shows the accepted ethics application by Delegated Research Ethics Committee (DERC) of University of Waterloo for the Modality study

B.2 Participants consent letter for modality study



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: _____

B.3 Questionnaire for assessing general AI acceptance

UNIVERSITY OF WATERLOO

Question Tour Block 1

Please enter your Participant ID

To what extent do you agree with following statements.
Use of unimodal data (Fundus imaging only) vs Comprehensive multimodal data (OCT, Visual Field, Fundus imaging, clinical history, and other relevant data types) by AI system

	Strongly	Fairly	Somewhat	Slightly	Not at all
Adoption and integrating of an AI system into your diagnostic process when it relies on comprehensive data will be higher than the AI system relying on Fundus imaging data.	○	○	○	○	○
Compared to traditional diagnostic methods, AI system which uses comprehensive data will be advantageous than the AI system using only Fundus imaging data?	○	○	○	○	○
Considering the potential for increased diagnostic accuracy and efficiency with the AI system's use of Fundus imaging data, how willing are you to adopt the technology into your clinical practice?	○	○	○	○	○
Considering the potential for increased diagnostic accuracy and efficiency with the AI system's use of comprehensive data, how willing are you to adopt the technology into your clinical practice?	○	○	○	○	○

Performance Expectancy

	Strongly	Fairly	Somewhat	Slightly	Not at all
Do you agree that an AI system will perform tasks at the same level or better than a human?	○	○	○	○	○
How confident are you that an AI system will produce accurate	○	○	○	○	○

results?

Do you agree that an AI system will save you time and effort compared to traditional methods?

Effort Expectancy

	Strongly	Fairly	Somewhat	Slightly	Not at all
How easy do you believe it is to interact with the AI system?	○	○	○	○	○
How confident are you that you can understand and use the AI system without any difficulty?	○	○	○	○	○
To what extent do you believe an AI system is user-friendly and intuitive to use?	○	○	○	○	○

Social influence

	Strongly	Fairly	Somewhat	Slightly	Not at all
To what extent do you believe that others in your organization use or will use an AI systems?	○	○	○	○	○
How important is it to you to use AI systems to fit in with others in your organization?	○	○	○	○	○
To what extent do you believe that the use of AI systems is becoming more common in your organization?	○	○	○	○	○

Facilitating conditions

	Strongly	Fairly	Somewhat	Slightly	Not at all
To what extent do you believe that the necessary resources, such as training and support, are available to help you use an AI system effectively?	○	○	○	○	○
How confident are you that the organization will provide the necessary resources for the successful implementation of the AI system?	○	○	○	○	○
To what extent do you believe that the organizational culture supports the use of AI systems?	○	○	○	○	○

Questions asked after each visit

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 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 feel confident in the clinical diagnosis?

Questionnaire for assessing trust

UNIVERSITY OF WATERLOO

Default Question Block

Please enter participant ID.

Please select the case you just reviewed

- ○ Patient 1 Unimodal
- ○ Patient 2 Unimodal
- ○ Patient 1 Multimodal
- ○ Patient 2 Multimodal

Use of unimodal data (Fundus imaging only) by AI system

	Strongly	Fairly	Somewhat	Slightly	Not at all
How confident are you that the AI system's reliance on Fundus imaging data alone results in accurate and reliable diagnoses?	○	○	○	○	○
To what extent do you trust the AI system's recommendations when they are solely based on Fundus imaging data	○	○	○	○	○
To what extent do you believe that Fundus imaging data alone is sufficient for the AI system to assist optometrist in diagnostic decisions?	○	○	○	○	○

Use of Comprehensive data (OCT, Visual Field, Fundus imaging, clinical history, and other relevant data types) AI system

	Strongly	Fairly	Somewhat	Slightly	Not at all
How confident are you that the comprehensive data used by the AI system leads to more accurate and reliable diagnostic decisions.	○	○	○	○	○
To what extent do you trust the AI system's recommendations based on the comprehensive data?	○	○	○	○	○
To what extent do you believe that the comprehensive data used by AI system system is sufficient for the AI system to assist optometrist in diagnostic decisions?	○	○	○	○	○

Information quality of AI system

	Strongly	Fairly	Somewhat	Slightly	Not at all
Do you trust the accuracy of the AI system's data output?	○	○	○	○	○
To what extent do you believe that the AI system provides complete and comprehensive information on a patient's condition?	○	○	○	○	○
To what extent do you believe AI system's conveys its decision, reasoning and recommendation in clear and understandable manner?	○	○	○	○	○

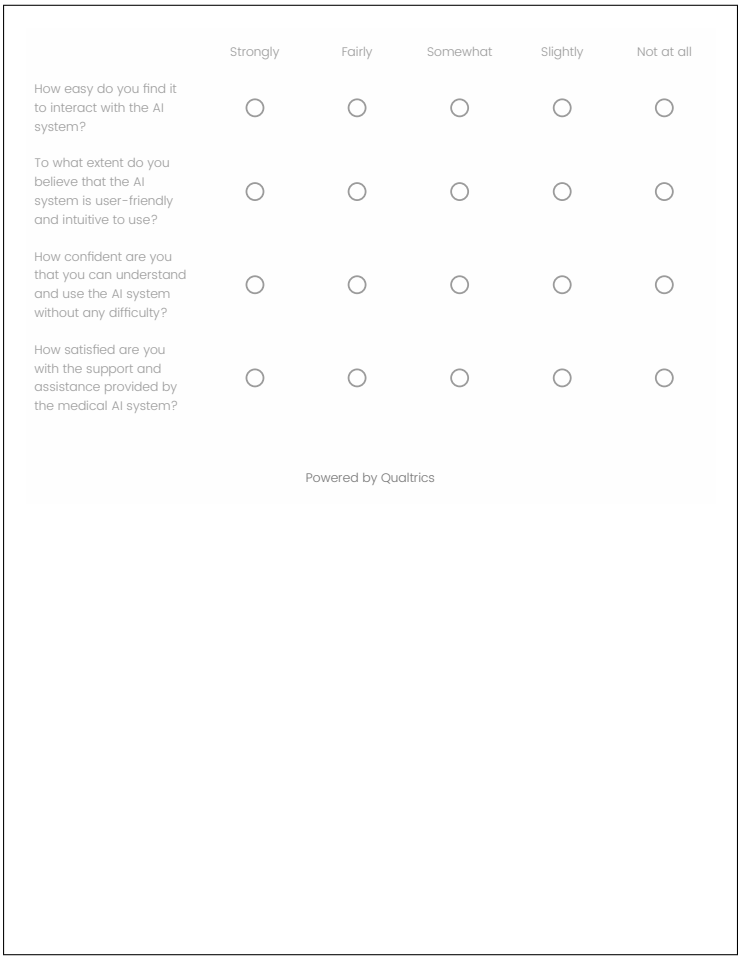
AI System Technical Accuracy

	Strongly	Fairly	Somewhat	Slightly	Not at all
To what extent do you trust the accuracy of the AI system's diagnosis?	○	○	○	○	○
To what extent do you believe that the AI system accurately detects glaucoma level?	○	○	○	○	○
To what extent do you believe that the level of accuracy provided by the this AI system is sufficient for making decisions or recommendations?	○	○	○	○	○
To what extent do you trust the accuracy of the AI system's explanations for its recommendations or decisions?	○	○	○	○	○

AI system Explainability

	Strongly	Fairly	Somewhat	Slightly	Not at all
To what extent do you trust the explainability of the AI system's?	○	○	○	○	○
How confident are you in the transparency and interpretability of the AI system's decision-making process?	○	○	○	○	○
To what extent do you trust the consistency and repeatability of the AI system's results?	○	○	○	○	○
How satisfied are you with the level of detail provided by the medical AI system to explain its decisions?	○	○	○	○	○

AI System Quality



	Strongly	Fairly	Somewhat	Slightly	Not at all
How easy do you find it to interact with the AI system?	○	○	○	○	○
To what extent do you believe that the AI system is user-friendly and intuitive to use?	○	○	○	○	○
How confident are you that you can understand and use the AI system without any difficulty?	○	○	○	○	○
How satisfied are you with the support and assistance provided by the medical AI system?	○	○	○	○	○

Powered by Qualtrics

Questionnaire for specific acceptance

UNIVERSITY OF WATERLOO

Default Question Block

Please enter your participant ID

Please select the case you just reviewed

- ○ Patient 1 Unimodal
- ○ Patient 2 Unimodal
- ○ Patient 1 Multimodal
- ○ Patient 2 Multimodal

Use of unimodal data (Fundus imaging only) by AI system

	Strongly	Fairly	Somewhat	Slightly	Not at all
How easy do you find it to adopt and integrate this AI system into your diagnostic process when it relied solely on Fundus imaging data?	○	○	○	○	○

	Strongly	Fairly	Somewhat	Slightly	Not at all
Compared to traditional diagnostic methods, how much more advantageous do you find this AI system that relied exclusively on Fundus imaging data?	○	○	○	○	○
Considering the potential for increased diagnostic accuracy and efficiency with the AI system's use of Fundus imaging data, how willing are you to adopt this technology into your clinical practice?	○	○	○	○	○

Use of Comprehensive multimodal data (OCT, Visual Field, Fundus imaging, clinical history, and other relevant data types) AI system

	Strongly	Fairly	Somewhat	slightly	Not at all
How easy do you find it to adopt and integrate this AI system into your diagnostic process when it relies on comprehensive multimodal data?	○	○	○	○	○

	Strongly	Fairly	Somewhat	slightly	Not at all
Compared to traditional diagnostic methods, how much more advantageous do you find this AI system that relies exclusively on comprehensive multimodal data?	○	○	○	○	○
Considering the potential for increased diagnostic accuracy and efficiency with this AI system's use of comprehensive multimodal data, how willing are you to adopt this technology into your clinical practice?	○	○	○	○	○

Performance Expectancy

	Strongly	Fairly	Somewhat	Slightly	Not at all
To what extent do you believe that this AI system will improve your ability to accurately diagnose and manage patients with glaucoma?	○	○	○	○	○
How confident are you that this AI system will produce accurate results in diagnosing and managing glaucoma patients ?	○	○	○	○	○

	Strongly	Fairly	Somewhat	Slightly	Not at all
To what extent do you believe that this AI system will save you time and effort?	○	○	○	○	○

Effort Expectancy

	Strongly	Fairly	Somewhat	Slightly	Not at all
How easy do you believe it will be to learn how to use this AI system?	○	○	○	○	○
To what extent do you believe that this AI system is be user-friendly and intuitive?	○	○	○	○	○
How confident are you that you can use AI system without experiencing significant difficulties or challenges?	○	○	○	○	○

Social Influence

	Strongly	Fairly	Somewhat	Slightly	Not at all
To what extent do you believe that your colleagues or peers would recommend using this AI system?	○	○	○	○	○

	Strongly	Fairly	Somewhat	Slightly	Not at all
How important is it to you that your supervisors or organizational policies support the use of this AI system?	○	○	○	○	○
To what extent do you believe that the use of this AI system will be common and accepted in the healthcare industry?	○	○	○	○	○

Facilitating Conditions

	strongly	Fairly	Somewhat	S,ightly	Not at all
To what extent do you believe that there are sufficient resources available (e.g., training, technical support) to help you use this AI system?	○	○	○	○	○
How confident are you that the organization will provide the necessary resources for the successful implementation of this AI system?	○	○	○	○	○

	strongly	Fairly	Somewhat	S,ightly	Not at all
To what extent do you believe that the organizational will supports the use of this AI systems?	○	○	○	○	○

Powered by Qualtrics

B.4 Descriptive statistics of Individual items of each dimension in trust

This section summarizes the results from different modalities (Uni and Multi) across various questions in each target. it includes statistical measures such as the mode, median, mean, standard deviation and providing insights into the response distributions and variability for each question and overall mean of the target.

Table B.1: Summary of input data modality across different modalities and questions.

	IDM_Q1		**IDM_Q2**		**IDM_Q3**		**Mean IDM**	
Modality	**Uni**	**Multi**	**Uni**	**Multi**	**Uni**	**Multi**	**Uni**	**Multi**
Valid	39	39	39	39	39	39	39	39
Mode	2.0	4.0	2.0	4.0	1.0	4.0	2	4
Median	2.0	4.0	2.0	4.0	2.0	4.0	2.0	4.0
Mean	2.3	3.9	2.4	3.7	1.7	4.0	2.1	3.9
Std. Deviation	1.0	0.9	0.8	0.8	0.7	1.0	0.7	0.8
Minimum	1.0	2.0	1.0	2.0	1.0	2.0	1	2
Maximum	4.0	5.0	4.0	5.0	4.0	5.0	3.7	5.0

Table B.2: Summary of information quality across different modalities and questions.

	IQ_Q1		**IQ_Q2**		**IQ_Q3**		**Mean IQ**	
Modality	**Uni**	**Multi**	**Uni**	**Multi**	**Uni**	**Multi**	**Uni**	**Multi**
Valid	39.0	39.0	39.0	39.0	39.0	39.0	39.0	39.0
Mode	3.0	4.0	1.0	4.0	4.0	5.0	3.0	3.7
Median	3.0	4.0	2.0	4.0	4.0	4.0	3.0	4.0
Mean	2.9	3.6	2.1	3.7	3.6	4.2	2.9	3.8
Std. Deviation	0.9	0.8	1.1	0.9	1.1	0.9	0.8	0.7
Minimum	1.0	2.0	1.0	2.0	1.0	2.0	1.0	2.0
Maximum	5.0	5.0	4.0	5.0	5.0	5.0	4.3	4.7

Table B.3: Summary of technical accuracy across different modalities and questions.

	TA_Q1		**TA_Q2**		**TA_Q3**		**TA_Q4**		**Mean TA**	
Modality	**Uni**	**Multi**	**Uni**	**Multi**	**Uni**	**Multi**	**Uni**	**Multi**	**Uni**	**Multi**
Valid	39.0	39.0	39.0	39.0	39.0	39.0	39.0	39.0	39.0	39.0
Mode	3.0	4.0	2.0	4.0	3.0	4.0	3.0	4.0	2.0	4.0
Median	3.0	4.0	3.0	4.0	2.0	4.0	3.0	4.0	2.8	3.8
Mean	2.8	3.5	2.8	3.5	2.2	3.4	2.6	3.7	2.6	3.5
Std. Deviation	0.9	0.8	1.1	0.9	1.0	1.0	0.9	0.9	0.9	0.8
Minimum	1.0	2.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	1.75
Maximum	5.0	5.0	5.0	5.0	5.0	5.0	4.0	5.0	4.5	4.8

Table B.4: Summary of output explainability across different modalities and questions.

	OE_Q1		OE_Q2		OE_Q3		OE_Q4		Mean OE	
Modality	**Uni**	**Multi**	**Uni**	**Multi**	**Uni**	**Multi**	**Uni**	**Multi**	**Uni**	**Multi**
Valid	39.0	39.0	39.0	39.0	39.0	39.0	39.0	38.0	39.0	39.0
Mode	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.3
Median	4.0	4.0	4.0	4.0	3.0	4.0	3.0	4.0	3.5	4.0
Mean	3.3	4.0	3.1	3.8	2.9	3.8	3.3	3.9	3.2	3.9
Std. Deviation	1.0	0.8	1.2	1.0	1.1	0.8	1.2	0.8	1.0	0.7
Minimum	1.0	2.0	1.0	1.0	1.0	2.0	1.0	2.0	1.0	2.3
Maximum	5.0	5.0	5.0	5.0	4.0	5.0	5.0	5.0	4.8	5.0

Table B.5: Summary of system quality across different modalities and questions.

	SQ_Q1		SQ_Q2		SQ_Q3		SQ_Q4		Mean SQ	
Modality	**Uni**	**Multi**	**Uni**	**Multi**	**Uni**	**Multi**	**Uni**	**Multi**	**Uni**	**Multi**
Valid	39.0	39.0	39.0	39.0	39.0	39.0	39.0	39.0	39.0	39.0
Mode	5.0	5.0	5.0	5.0	5.0	5.0	4.0	5.0	4.3	5.0
Median	5.0	5.0	5.0	5.0	5.0	5.0	4.0	5.0	4.3	5.0
Mean	4.4	4.7	4.5	4.7	4.4	4.6	3.8	4.5	4.3	4.6
Std. Deviation	1.0	0.5	0.8	0.5	0.9	0.7	1.0	0.6	0.8	0.5
Minimum	1.0	3.0	1.0	4.0	1.0	2.0	1.0	3.0	1.0	3.5
Maximum	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0

B.5 Descriptive statistics of confidence score

	Unimodal (Mean±SD)			Multimodal (Mean±SD)		
	Novice	Experts	Overall	Novice	Experts	Overall
Correct	3.2 ± 1.1	3.1 ± 1.5	3.1 ± 1.2	4.1 ± 0.7	4.0 ± 0.9	4.1 ± 0.7
Incorrect	3.0 ± 0.8	2.9 ± 1.2	2.9 ± 0.9	3.7 ± 0.9	3.1 ± 1.3	3.5 ± 1.1
Overall	3.1 ± 1.0	3.0 ± 1.4	3.1 ± 1.2	3.9 ± 0.8	3.8 ± 1.1	3.8 ± 1.1

Table B.6: Mean ± SD confidence scores for novice and expert optometrists under Unimodal and multimodal conditions, separated by AI decision correctness.

	Correct AI system		Incorrect AI system	
	Unimodal	Multimodal	Unimodal	Multimodal
Accuracy	69.8%	87.5%	54.1%	15.1%
Precision	58.8%	76.7%	50.5%	48.1%
Recall	34.5%	72.2%	36.7%	9%
F1-Score	32.4%	74.1%	31.3%	14%

Table B.7: Performance of optometrists while intercating with unimodal and multimodal systems, separated by AI decision correctness.

B.6 Additional statistics for user acceptance

Table B.8: Descriptive statistics of modified UTAUT items

UTAUT Item	Mean	SD	SE	Coeff. of Variation
Uni-Modality Q1	2.462	1.140	0.223	0.463
Multi-Modality Q1	1.667	0.679	0.131	0.408
Uni-Modality Q2	3.231	1.032	0.202	0.319
Multi-Modality Q2	2.000	0.620	0.119	0.310
Uni-Modality Q3	2.731	1.041	0.204	0.381
Multi-Modality Q3	1.630	0.742	0.143	0.455
Uni-Performance Expectancy - Q1	3.000	1.000	0.160	0.333
Multi-Performance Expectancy - Q1	2.231	0.902	0.144	0.404
Uni-Performance Expectancy - Q2	3.154	0.933	0.149	0.296
Multi-Performance Expectancy - Q2	2.538	0.822	0.132	0.324
Uni-Performance Expectancy - Q3	2.872	1.056	0.169	0.368
Multi-Performance Expectancy - Q3	2.205	1.031	0.165	0.467
Uni-Effort Expectancy - Q1	1.615	0.590	0.094	0.365
Multi-Effort Expectancy - Q1	1.359	0.537	0.086	0.395
Uni-Effort Expectancy - Q2	1.590	0.751	0.120	0.472
Multi-Effort Expectancy - Q2	1.359	0.486	0.078	0.358
Uni-Effort Expectancy - Q3	1.821	0.914	0.146	0.502
Multi-Effort Expectancy - Q3	1.462	0.682	0.109	0.467

Table B.9: Pairwise comparison of modified UTAUT items for unimodal and multimodal systems

Measure 1	**Measure 2**	**t**	**df**	**p**
Uni-Modality Q1	Multi-Modality Q1	3.339	25	0.003
Uni-Modality Q2	Multi-Modality Q2	5.050	25	$< .001$
Uni-Modality Q3	Multi-Modality Q3	5.312	25	$< .001$
Uni-Performance Expectancy - Q1	Multi-Performance Expectancy - Q1	4.234	38	$< .001$
Uni-Performance Expectancy - Q2	Multi-Performance Expectancy - Q2	3.991	38	$< .001$
Uni-Performance Expectancy - Q3	Multi-Performance Expectancy - Q3	4.127	38	$< .001$
Uni-Effort Expectancy - Q1	Multi-Effort Expectancy - Q1	2.693	38	0.010
Uni-Effort Expectancy - Q2	Multi-Effort Expectancy - Q2	2.688	38	0.011
Uni-Effort Expectancy - Q3	Multi-Effort Expectancy - Q3	3.837	38	$< .001$
Uni-Social Influence - Q1	Multi-Social Influence - Q1	3.261	38	0.002
Uni-Social Influence - Q2	Multi-Social Influence - Q2	2.667	38	0.011
Uni-Social Influence - Q3	Multi-Social Influence - Q3	2.022	38	0.050
Uni-Facilitating Conditions - Q1	Multi-Facilitating Conditions - Q1	3.650	38	$< .001$
Uni-Facilitating Conditions - Q2	Multi-Facilitating Conditions - Q2	2.823	38	0.008
Uni-Facilitating Conditions - Q3	Multi-Facilitating Conditions - Q3	3.376	38	0.002