Title

Usability and Clinician Acceptance of a Deep Learning-Based Clinical Decision Support Tool for Predicting Glaucomatous Visual Field Progression

Journal

Journal of Glaucoma, 32(3)

ISSN

1057-0829

Authors

Chen, Jimmy S

Baxter, Sally L

van den Brandt, Astrid

et al.

Publication Date

2023-03-01

DOI

10.1097/ijg.0000000000002163

Peer reviewed

HHS Public Access

Author manuscript

J Glaucoma. Author manuscript; available in PMC 2024 March 01.

Published in final edited form as:

J Glaucoma. 2023 March 01; 32(3): 151–158. doi:10.1097/IJG.0000000000002163.

Usability and Clinician Acceptance of a Deep Learning-based Clinical Decision Support Tool for Predicting Glaucomatous Visual Field Progression

Jimmy S. Chen, MD[1,2,*], **Sally L. Baxter, MD, MSc**[1,2,*], **Astrid van den Brandt, MSc**[3], **Alexander Lieu, BS**[1], **Andrew S. Camp, MD**[1], **Jiun L. Do, MD, PhD**[1], **Derek S. Welsbie, MD, PhD**[1], **Sasan Moghimi, MD**[1], **Mark Christopher, PhD**[1], **Robert N. Weinreb, MD**[1], **Linda Zangwill, PhD**[1]

[1]Division of Ophthalmology Informatics and Data Science, Viterbi Family Department of Ophthalmology and Shiley Eye Institute, University of California San Diego, La Jolla, CA

[2]UCSD Health Department of Biomedical Informatics, University of California San Diego, La Jolla, CA

[3]Eindhoven University of Technology, The Netherlands

Abstract

Purpose: To evaluate clinician perceptions of a prototyped clinical decision support (CDS) tool that integrates visual field (VF) metric predictions from artificial intelligence (AI) models.

Methods: 10 ophthalmologists and optometrists from University of California San Diego participated in 6 cases from 6 patients, consisting of 11 eyes, uploaded to a CDS tool ("GLANCE", designed to help clinicians "at a glance"). For each case, clinicians answered questions about management recommendations and attitudes towards GLANCE, particularly regarding utility and trustworthiness of the AI-predicted VF metrics, and willingness to decrease VF testing frequency.

Main Outcome(s) and Measure(s): Mean counts of management recommendations and mean Likert scale scores were calculated to assess overall management trends and attitudes towards the CDS tool for each case. Additionally, system usability scale (SUS) scores were calculated.

Results: The mean Likert scores for trust in and utility of the predicted VF metric, and clinician willingness to decrease VF testing frequency was 3.27, 3.42, and 2.64 respectively (1=strongly disagree, 5=strongly agree). When stratified by glaucoma severity, all mean Likert scores decreased as severity increased. The SUS score across all responders was 66.1 ± 16.0 (43rd percentile).

Conclusions: A CDS tool can be designed to present AI model outputs in a useful, trustworthy manner that clinicians were generally willing to integrate into their clinical decision making.

Corresponding Author: Linda M. Zangwill, PhD, 9415 Campus Point Drive, MC 0946, La Jolla, CA 92093-0946, lzangwill@health.ucsd.edu.

*JSC and SLB contributed to this work equally

Future work is needed to understand how to best develop explainable and trustworthy CDS tools integrating AI before clinical deployment.

PRECIS

We updated a clinical decision support tool integrating predicted visual field (VF) metrics from an artificial intelligence model and assessed clinician perceptions of the predicted VF metric in this usability study.

Keywords

artificial intelligence; glaucoma; clinical decision support; informatics

INTRODUCTION

Applications of artificial intelligence (AI) are rapidly advancing in ophthalmology. Deep learning (DL) methods in particular have facilitated the classification and interpretation of ophthalmic data.[1–4] Glaucoma remains a leading cause of global irreversible blindness[5] and is a key clinical domain for AI applications. Glaucoma management involves the integration of multiple testing and imaging modalities,[6] making it ripe for clinical decision support (CDS). Broadly, CDS "provides clinicians, staff, [or] patients… with knowledge and person-specific information, intelligently filtered or presented… to enhance health and healthcare."[7] The potential for AI tools to provide CDS via facilitating diagnosis of glaucoma and identification of patients at risk for progression has been widely touted.[8–11]

However, implementation of AI-based CDS for clinical practice is not straightforward, and best practices are still evolving.[12–14] Early efforts at creating knowledge-based systems have been considered complex and time-intensive,[15,16] and in recent years CDS has reverted to simpler forms, such as order sets or adverse drug event alerts. Although these types of CDS interventions have been integrated into electronic health record (EHR) systems, the emergence of advanced computational models (i.e. those associated with machine learning [ML] or DL) present unique challenges and considerations. Previous studies have described frameworks describing some of these challenges, including algorithm explainability and transparency, data standardization, and clinical workflow integration to provide meaningful and actionable decision support to clinicians.[14,17,18]

Prior studies have shown that obtaining support from organizational leadership and end-users early in the implementation of AI-based tools is critically important.[17,19] In a framework for implementing ML into healthcare published by Shaw and colleagues,[17] the authors stated that algorithms need to have a "meaningful entryway" into decision-making. Establishing what is "meaningful" inherently requires end-user input. Furthermore, usability has become a prominent concern in health information technology (IT), particularly within a field such as ophthalmology that incorporatse multiple technological modalities and has high patient volumes demanding usability and efficiency.[20–22] Several studies have also shown that health IT inflicts significant cognitive burden on clinicians, which can contribute to burnout.[23,24] Therefore, prioritizing usability and minimizing additional cognitive load with the integration of these tools is crucial.

In this study, we solicited feedback and performed usability assessments among practicing ophthalmologists and optometrists regarding an early prototyped CDS tool for glaucoma management.[25] This tool leverages a previously developed DL model predicting quantitative visual field (VF) measurements based on optical coherence tomography (OCT) scans.[26,27] Given the ease and speed of OCT testing, these models have the potential to tailor the frequency of VF testing for patients and may result in decreased need for VF testing, which is time-consuming and frequently difficult for patients.[28] Using a mixed-methods approach to solicit feedback from potential end-users of this tool early in the design process, we aimed to improve the feasibility of this tool for future clinical implementation and draw insights that may inform efforts in implementing AI-based tools in ophthalmology workflows more generally.

METHODS

Deep Learning Models for Predicting Visual Field Measurements

The output of a previously published DL model trained on optic nerve head OCT B-scans was chosen for implementation within a CDS tool.[27] This DL model was trained to predict corresponding VF outcomes, reported in mean deviation (MD). This model achieved an R^2 of roughly 0.7 in predicting MD, suggesting high correlation with patient-produced VF results.

Iterative Design of a Clinical Decision Support Tool for Visualizing Model Predictions

GLANCE was designed as a CDS tool to provide a graphical user interface (GUI) of the above DL model's predictions and relevant clinical data to assist physicians with evaluating glaucoma progression (helping clinicians "at a glance").[25] The design process consisted of a previously described user-centric process with clinician interviews and multiple rounds of prototyping interfaces.[25] This process demonstrated a need for a GUI that was 1) reliable, 2) showed why a model made a prediction, 3) highlighted imaging features relevant to a prediction, and 4) could guide future scheduling of VFs. The initial interface displayed the AI-predicted MD and heatmap visualization of relevant regions from OCT images used for model prediction. For this pilot study, we updated the interface to also include clinical data (age, race, refraction, pachymetry, ocular history and medications), longitudinal IOP data, historical VFs, and a visual comparison of previous AI-predicted MD to real MD results for each patient (Supplemental Figure 1).

Study Population

This study was approved by the University of California San Diego (UCSD) Institutional Review Board (IRB) as a quality improvement protocol and adhered to the principles of the Declaration of Helsinki. Clinicians invited to participate in the study were either ophthalmology residents (post-graduate year 2 or greater), glaucoma fellows, attending ophthalmologists, or optometrists. The evaluation was conducted anonymously online (Qualtrics, Provo, UT) from 10/2021–12/2021, with reminders emailed biweekly. The selection of the 6 patients included in our cases was previously described in a prior iteration of this study to represent a range of glaucoma severity,[25] and images from these patients were not previously seen by the DL model.

Evaluating Usability and Attitudes towards GLANCE

Clinicians were asked to provide age, race, ethnicity, and their clinical role (resident, fellow, attending, optometrist) at UCSD. For each case, which consisted of all ocular and demographic information associated with a glaucomatous eye, clinicians were shown the GLANCE interface (Figure 1). Clinicians were also asked to provide one of four recommendations for each case: 1) continuing present management with routine follow-up (no change in testing frequency), 2) longer follow-up (decreased testing frequency), 3) shorter follow-up (increased testing frequency), or 4) escalating therapy (increasing medications, recommending a laser procedure or surgery). Clinicians were also asked to rank 3 statements using a 5-point Likert scale (from 1=strongly disagree to 5=strongly agree) to assess their attitudes towards the AI-predicted MD in each case: 1) I trust the predicted MD enough to incorporate it into my decision-making, 2) the predicted MD provides additional useful information beyond the existing clinical information available, and 3) I would likely decrease the frequency of visual field testing for this patient if I had predicted MDs available from this algorithm. Users were given the option to comment on the interface or case. At the end of the cases, users were asked to complete the System Usability Scale (SUS)[29] regarding their overall experience with GLANCE. Finally, they were provided an opportunity to provide open-ended feedback on the design of GLANCE. These questions are provided in Supplemental Table 1.

Statistical Analyses

Statistical analysis was performed in R version 4.0.5 (R Foundation; Vienna, Austria). Users who completed >80% of the questions were included for analysis. For each question regarding glaucoma management, the mean number of responses for each management option was calculated across all users who responded. Mean Likert scale scores for each question assessing attitude towards GLANCE were calculated across all users for each case. Subgroup analyses of management responses and mean Likert scores were also performed by glaucoma disease severity, defined based on the glaucoma staging system (GSS)[30]: mild = −0 to −5.99 dB, moderate = −6 to −11.99 dB, advanced = −12 to −19.99 dB, and severe > −20 dB. For each user, we also calculated the SUS score, a previously validated scoring system out of 100 used commonly to evaluate ease of a particular interface in user experience design.[29]

RESULTS

Users

10 users completed > 80% of the cases and were included in this study. These users consisted of 3 ophthalmology residents (postgraduate year 2–4), 2 glaucoma fellows, 3 ophthalmology attendings (2 glaucoma, 1 comprehensive), and 2 optometrists. Four users were 25–34 years old, four were 35–44 years old, one was 45–54 years old, and one was 65–74 years old. One user (a glaucoma fellow) did not complete a case and the questions regarding attitudes towards the interface. This user was excluded from calculations of cases they did not participate in.

Cases

6 cases, consisting of data available for 6 patients and 11 eyes, were uploaded to GLANCE in randomized order of severity. The mean age of all patients was 76.3±6.9 years, and all patients included in this usability study were white. Based on GSS, 6 eyes had mild glaucoma, 2 eyes had moderate glaucoma, 2 eyes had advanced glaucoma, and 1 eye had severe glaucoma. All demographic, pachymetry, refraction, and visual field measurements for each case and eye are reported in Table 1.

User-Reported Management of Glaucoma

Clinicians generally favored no change in management (72.5%) for milder cases, and escalating care with more advanced and severe glaucoma (72.2% and 90% respectively), as shown in Figure 2. Clinicians also had more variation in their choices for managing milder glaucoma and generally agreed upon elevating care for more severe glaucoma. Subgroup analysis of decision making by age and role versus management showed no remarkable differences between management preferences (Supplemental Figure 2).

Attitudes Towards GLANCE

Clinicians generally perceived the predicted MD from the AI model as somewhat trustworthy (mean Likert score=3.27) and somewhat useful (mean Likert score=3.42), but did not feel that the predicted MD would decrease their visual field testing frequency (mean Likert score=2.64) [Supplemental Figure 3]. While mean Likert scores for all users demonstrated variation as users progressed through the cases, there was a general downtrend in all attitudes towards GLANCE (Supplemental Figure 4) despite randomization of case severity. The most notable decrease was in assessment of visual testing frequency, with the mean Likert score decreasing from 3 to 2.1 after 6 cases. The mean Likert score generally also decreased across all attitudes towards GLANCE as glaucoma severity increased (Figure 3).

System Usability Scale Scores

The mean ± standard deviation (SD) SUS score was 66.1 ± 16.0, translating roughly to the 43rd percentile.[29] Mean SUS scores ranged from 40–82.5 across the 9 users who completed this portion of the evaluation (Supplemental Figure 5).

Open-Ended Comments

Two clinicians responded to optional requests for comments on each case and overall comments regarding GLANCE. Overall, users expressed concern regarding how the AI model calculated MD, reliability of the heatmap, and how significantly elevated intraocular pressure or high MD would affect their clinical decision making. A full table of comments is available in Table 2.

DISCUSSION

In this study, we updated a GUI for GLANCE, a previously published AI-based CDS tool,[25] and conducted usability evaluations with clinicians to evaluate its utility in assisting

with glaucoma management. While literature regarding AI model development in glaucoma is advancing rapidly, there is a scarcity of studies examining end-user attitudes and design considerations for clinical implementation. This study addresses a significant gap in knowledge regarding how implementation of AI model outputs in a GUI could assist clinicians and potentially reduce the VF testing burden on patients.

There are two key findings in our study: 1) clinician perceptions were somewhat positive towards the trustworthiness and utility of AI-predicted VF metric, and 2) clinicians were less likely to use the AI output in their decision making as glaucoma severity increased. Overall, the mean Likert scale score for trustworthiness and utility of the predicted MD were 3.27 and 3.42 respectively (Supplemental Figure 3), signifying slightly positive sentiment. However, clinician perceptions of the predicted MD became increasingly unfavorable with more severe disease (Figure 3), particularly with decreasing VF testing. These findings altogether suggest that despite the AI model's published high performance,[27] clinicians were most hesitant in using the AI-predicted output as a surrogate for patient visual function to decrease VF frequency. Furthermore, their clinical decision-making process remains largely driven by the incorporation of all elements of data, which is reinforced by a clinician comment stating that the "IOP [was] too high for my comfort… regardless of the predicted MD" (Table 2). Although the AI output may be potentially helpful for glaucoma management, other reasons may explain why clinician trust in the predicted MD was not higher. For example, the predicted MDs were inherently noisy, especially for advanced disease, and may be difficult to compare to prior patient tests.[31] Additionally, while explainability methods such as heat maps were used to elucidate the AI model's rationale, clinician decision making may not always agree with the AI's reasoning, which may foster some distrust in the AI output. In our study, clinicians expressed concern towards the AI output when the regions of interest bounded the choroid in OCT scans, which is not conventionally assessed in glaucoma. While it is possible that the AI model identified the choroid as a potential biomarker,[32,33] more work is needed to improve the explainability of AI models' outputs to clinicians.[34] Other design choices may also assist in gaining physician trust in an AI model. As previously described, the GLANCE interface was designed as simply as possible, with the assumption that if past MD predictions were highly accurate for a specific patient (e.g. cases 1, 2 & 4), the clinician would be more likely to trust the AI algorithm. We also attempted to readily display all clinically relevant information (including IOP, age, pachymetry). Future studies may help clarify if more information or a change in user interface would affect clinician trust in the predicted MDs.

There remains a significant gap between algorithmic development AI integration into clinical workflows, a crucial step in realizing the benefits of AI for patients. Implementation of AI models into the EHR is a logical next step, as clinicians use the EHR routinely in their workflows.[35–38] A recent systematic review by Lee et al. revealed several challenges prohibiting widespread clinical implementation of predictive models in the EHR.[39] For example, CDS tools may increase time spent in the EHR, cognitive burden, and alert fatigue for clinicians who already experience high alert burden.[40] In our study, longitudinally decreasing Likert scores across all attitudes towards the AI-predicted MD (Supplemental Figure 4) may also indicate some element of "AI fatigue"[41] even across our small sample of cases. This has important implications for clinical implementation, especially

in high-volume glaucoma clinics where the majority of patients have glaucoma. Additional AI-specific challenges include: diagnostic drift, as well as transparency and explainability of its outputs.[14,42] Current predictive models embedded into EHRs have focused mostly on inpatient diseases such as deep vein thrombosis[43] and sepsis,[44] and none exist for ophthalmology. For glaucoma, reducing the frequency of VF testing without compromising clinical outcomes would be beneficial, particularly in light of social distancing due to the recent COVID-19 pandemic, and may decrease staff and patient time and expense. Additionally, GLANCE also has the potential to reduce clinician burden during chart review with integration of clinical and demographic data into a single interface. Although we focused mainly on trust and usability rather than other potential applications and benefits of this AI model, future research may focus on user-centered design workshops to elicit more nuanced feedback and improve future implementation.

Addressing user experience (UX) and usability of these CDS tools is an important next step to narrow the gap between development and implementation. Previously developed frameworks for usability evaluation such as User, Function, Representation, and Task (UFuRT)[45] exist for data management systems, and have been previously implemented in guiding CDS design.[46] Other work has focused on consistent design concepts, controlled terminology, and appropriate visual representation of data.[47] As part of our stakeholder interviews in our user design process, we focused on similar UX design elements, notably simplicity due to clinician time restraints and displaying key elements of information together (i.e. VF progression data with OCT scans). Previous usability studies for CDS tools utilizing defined UX principles have focused on management of diseases such as sepsis,[48] diabetes,[49,50] and depression,[51]and have mostly been evaluated in a pre-clinical context with the exception of the SepsisWatch model.[44] While our study represents a step towards clinician-centered design for AI outputs, more work is needed to understand how we can design user interfaces that support the needs of clinicians in context of their already hectic workflows. While GLANCE was perceived to have modest usability (SUS score in the 43rd percentile), previous work has shown that clinicians generally have negative perceptions of EHR usability when evaluated using SUS scores (mean scores < 10th percentile).[52] This stresses the challenges of creating usable CDS tools in the EHR and the need for ongoing work in this space. To achieve successful clinical implementation of AI-based CDS tools, these tools will need to undergo larger user studies with more clinicians and patient cases, with further iterations incorporating diverse patient factors such as race and comorbid disease pathology as well as AI-based CDS tools for predicting visual field progression.[47]

This study has several additional limitations that future work may address. First, our sample size was limited to 10 clinicians and 6 cases. Although prior studies have demonstrated that >90% of usability issues can be detected with 5 users,[53,54] it is imperative that these CDS tools are evaluated by more clinicians prior to clinical deployment, including those in other practice settings and those using different EHRs, as well as with cases consisting of more diverse patients both demographically and clinically. Second, we chose to display data in a single interface thought to be most relevant to glaucoma management. Other information (other data from VF testing such as pattern deviation, location of VF loss, and more medical history) may be helpful in glaucoma management and may be displayed in other ways such as dynamic user interfaces. Additionally, incorporation of point-wise

predictions of VF[55, 56] may assist in increasing clinician trust in the predicted VF. However, these will require future iterations of GLANCE and may be implemented as part of future stakeholder interviews and data review. Third, we conducted this usability pilot study online due to restrictions on research studies conducted during the COVID-19 pandemic, but future studies would benefit from in-person observations of clinicians using the tool and approaches such as “think-aloud” or “talk-aloud” protocols.

A CDS tool for glaucoma can be developed to display AI-based outputs for VF testing management in a useful and trustworthy manner when designed with user-centered principles. While more work is needed to understand how clinicians interact with the CDS tool and its outputs, this study represents an important step towards translating AI models to bedside. Continued interdisciplinary collaboration between informaticians, computer scientists, clinicians, and patients will be needed to make clinical implementation of AI a reality.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This study was supported by the National Institutes of Health (Bethesda, MD, Grants T15LM01127, DP5OD029610, P30EY022589, EY026574, K99EY030942, R01EY027510, R01EY029058).

Financial Support:

This study was supported by the National Institutes of Health (Bethesda, MD, Grants T15LM01127, DP5OD029610, P30EY022589, EY026574, K99EY030942 R01EY027510, R01EY029058). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The funders had no role in the design or conduct of this research.

Conflicts of Interest:

Linda Zangwill is a consultant for Abbvie Inc. Digital Diagnostics, receives research funding from the National Eye Institute, Carl Zeiss Meditec Inc.,Heidelberg Engineering GmbH, Optovue Inc, Topcon Medical Systems Inc, and has patented intellectual property with Zeiss Meditec. Robert N. Weinreb is a consultant for Abbvie, Aerie Pharmaceuticals, Allergan, Equinox, Iantrek, Implandata, Nicox, Topcon Medical, and receives research funding from National Eye Institute, National Institute on Minority Health and Health Disparities, Bausch & Lomb, Topcon Medical, Heidelberg Engineering, Carl Zeiss Meditec, Optovue, Centervue, and has patented intellectual property licensed by the University of California San Diego to Zeiss Meditec and Toromedes.

REFERENCES

1. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016;316(22):2402–2410. doi:10.1001/jama.2016.17216 [PubMed: 27898976]
2. Brown JM, Peter Campbell J, Beers A, et al. Fully automated disease severity assessment and treatment monitoring in retinopathy of prematurity using deep learning. In: Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications. Vol 10579. SPIE; 2018:149–155. doi:10.1117/12.2295942
3. Chen JS, Coyner AS, Ostmo S, et al. Deep Learning for the Diagnosis of Stage in Retinopathy of Prematurity: Accuracy and Generalizability across Populations and Cameras. Ophthalmol Retina. 2021;5(10):1027–1035. doi:10.1016/j.oret.2020.12.013 [PubMed: 33561545]

4. Fan R, Bowd C, Christopher M, et al. Detecting Glaucoma in the Ocular Hypertension Study Using Deep Learning. JAMA Ophthalmol. Published online March 17, 2022. doi:10.1001/jamaophthalmol.2022.0244
5. Steinmetz JD, Bourne RRA, Briant PS, et al. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. The Lancet Global Health. 2021;9(2):e144–e160. https://www.sciencedirect.com/science/article/pii/S2214109X20304897 [PubMed: 33275949]
6. de Moraes CG, Liebmann JM, Medeiros FA, Weinreb RN. Management of advanced glaucoma: Characterization and monitoring. Surv Ophthalmol. 2016;61(5):597–615. doi:10.1016/j.survophthal.2016.03.006 [PubMed: 27018149]
7. 10×10 with university of Utah: Course description. AMIA - American Medical Informatics Association. Accessed April 15, 2022. https://amia.org/education-events/education-catalog/10×10-university-utah/course-description
8. Zheng C, Johnson TV, Garg A, Boland MV. Artificial intelligence in glaucoma. Curr Opin Ophthalmol. 2019;30(2):97–103. doi:10.1097/ICU.0000000000000552 [PubMed: 30562242]
9. Salazar H, Misra V, Swaminathan SS. Artificial intelligence and complex statistical modeling in glaucoma diagnosis and management. Curr Opin Ophthalmol. 2021;32(2):105–117. doi:10.1097/ICU.0000000000000741 [PubMed: 33395111]
10. Devalla SK, Liang Z, Pham TH, et al. Glaucoma management in the era of artificial intelligence. British Journal of Ophthalmology. 2020;104(3):301–311. doi:10.1136/bjophthalmol-2019-315016 [PubMed: 31640973]
11. Girard MJA, Schmetterer L. Artificial intelligence and deep learning in glaucoma: Current state and future prospects. Prog Brain Res. 2020;257:37–64. doi:10.1016/bs.pbr.2020.07.002 [PubMed: 32988472]
12. Harris AH. Path From Predictive Analytics to Improved Patient Outcomes: A Framework to Guide Use, Implementation, and Evaluation of Accurate Surgical Predictive Models. Ann Surg. 2017;265(3):461–463. doi:10.1097/SLA.0000000000002023 [PubMed: 27735825]
13. Amarasingham R, Patzer RE, Huesch M, Nguyen NQ, Xie B. Implementing electronic health care predictive analytics: considerations and challenges. Health Aff. 2014;33(7):1148–1154. doi:10.1377/hlthaff.2014.0352
14. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med. 2019;25(1):30–36. doi:10.1038/s41591-018-0307-0 [PubMed: 30617336]
15. Coats PK. Why Expert Systems Fail. Financial Management. 1988;17(3):77–86. doi:10.2307/3666074
16. Heathfield H. The rise and “fall” of expert systems in medicine. Expert Systems. 1999;16(3):183–188. doi:10.1111/1468-0394.00107
17. Shaw J, Rudzicz F, Jamieson T, Goldfarb A. Artificial Intelligence and the Implementation Challenge. J Med Internet Res. 2019;21(7):e13659. doi:10.2196/13659
18. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. J Med Internet Res. 2017;19(11):e367. doi:10.2196/jmir.8775 [PubMed: 29092808]
19. Benda NC, Das LT, Abramson EL, et al. “How did you get to this number?” Stakeholder needs for implementing predictive analytics: a pre-implementation qualitative study. Journal of the American Medical Informatics Association. 2020;27(5):709–716. doi:10.1093/jamia/ocaa021 [PubMed: 32159774]
20. Baxter SL, Gali HE, Chiang MF, et al. Promoting Quality Face-to-Face Communication during Ophthalmology Encounters in the Electronic Health Record Era. Appl Clin Inform. 2020;11(1):130–141. doi:10.1055/s-0040-1701255 [PubMed: 32074650]
21. Chiang MF, Boland MV, Brewer A, et al. Special requirements for electronic health record systems in ophthalmology. Ophthalmology. 2011;118(8):1681–1687. doi:10.1016/j.ophtha.2011.04.015 [PubMed: 21680023]

22. Chiang MF, Read-Brown S, Tu DC, et al. Evaluation of electronic health record implementation in ophthalmology at an academic medical center (an American Ophthalmological Society thesis). Trans Am Ophthalmol Soc. 2013;111:70–92. https://www.ncbi.nlm.nih.gov/pubmed/24167326 [PubMed: 24167326]
23. Downing NL, Lance Downing N, Bates DW, Longhurst CA. Physician Burnout in the Electronic Health Record Era: Are We Ignoring the Real Cause? Annals of Internal Medicine. 2018;169(1):50. doi:10.7326/m18-0139 [PubMed: 29801050]
24. Kruse CS, Mileski M, Dray G, Johnson Z, Shaw C, Shirodkar H. Physician Burnout and the Electronic Health Record Leading Up to and During the First Year of COVID-19: Systematic Review. J Med Internet Res. 2022;24(3):e36200. doi:10.2196/36200
25. van den Brandt A, Christopher M, Zangwill LM, et al. GLANCE: Visual Analytics for Monitoring Glaucoma Progression. The Eurographics Association; 2020. doi:10.2312/vcbm.20201175
26. Christopher M, Bowd C, Proudfoot JA, et al. Deep Learning Estimation of 10–2 and 24–2 Visual Field Metrics Based on Thickness Maps from Macula OCT. Ophthalmology. 2021;128(11):1534–1548. doi:10.1016/j.ophtha.2021.04.022 [PubMed: 33901527]
27. Christopher M, Bowd C, Belghith A, et al. Deep Learning Approaches Predict Glaucomatous Visual Field Damage from OCT Optic Nerve Head En Face Images and Retinal Nerve Fiber Layer Thickness Maps. Ophthalmology. 2020;127(3):346–356. doi:10.1016/j.ophtha.2019.09.036 [PubMed: 31718841]
28. Broadway DC. Visual field testing for glaucoma - a practical guide. Community Eye Health. 2012;25(79–80):66–70. https://www.ncbi.nlm.nih.gov/pubmed/23520423 [PubMed: 23520423]
29. Brooke J. SUS: A “Quick and Dirty” Usability Scale. In: Usability Evaluation In Industry. CRC Press; 1996:207–212. doi:10.1201/9781498710411-35
30. Mills RP, Budenz DL, Lee PP, et al. Categorizing the stage of glaucoma from pre-diagnosis to end-stage disease. Am J Ophthalmol. 2006;141(1):24–30. doi:10.1016/j.ajo.2005.07.044 [PubMed: 16386972]
31. Artes PH, Chauhan BC. Signal/noise analysis to compare tests for measuring visual field loss and its progression. Invest Ophthalmol Vis Sci. 2009;50(10):4700–4708. doi:10.1167/iovs.09-3601 [PubMed: 19458326]
32. Verticchio Vercellin A, Harris A, Stoner AM, Oddone F, Mendoza KA, Siesky B. Choroidal Thickness and Primary Open-Angle Glaucoma-A Narrative Review. J Clin Med Res. 2022;11(5). doi:10.3390/jcm11051209
33. Breher K, Terry L, Bower T, Wahl S. Choroidal Biomarkers: A Repeatability and Topographical Comparison of Choroidal Thickness and Choroidal Vascularity Index in Healthy Eyes. Transl Vis Sci Technol. 2020;9(11):8. doi:10.1167/tvst.9.11.8
34. Arun N, Gaw N, Singh P, et al. Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. Radiol Artif Intell. 2021;3(6):e200267. doi:10.1148/ryai.2021200267
35. Chen JS, Hribar MR, Goldstein IH, et al. Electronic health record note review in an outpatient specialty clinic: who is looking? JAMIA Open. 2021;4(3):ooab044. doi:10.1093/jamiaopen/ooab044
36. Hribar MR, Read-Brown S, Goldstein IH, et al. Secondary use of electronic health record data for clinical workflow analysis. Journal of the American Medical Informatics Association. 2018;25(1):40–46. doi:10.1093/jamia/ocx098 [PubMed: 29036581]
37. Henriksen BS, Goldstein IH, Rule A, et al. Electronic Health Records in Ophthalmology: Source and Method of Documentation. Am J Ophthalmol. 2020;211:191–199. doi:10.1016/j.ajo.2019.11.030 [PubMed: 31811860]
38. Baxter SL, Apathy NC, Cross DA, Sinsky C, Hribar MR. Measures of electronic health record use in outpatient settings across vendors. J Am Med Inform Assoc. 2021;28(5):955–959. doi:10.1093/jamia/ocaa266 [PubMed: 33211862]
39. Lee TC, Shah NU, Haack A, Baxter SL. Clinical Implementation of Predictive Models Embedded within Electronic Health Record Systems: A Systematic Review. Informatics. 2020;7(3):25. doi:10.3390/informatics7030025 [PubMed: 33274178]

40. Embi PJ, Leonard AC. Evaluating alert fatigue over time to EHR-based clinical trial alerts: findings from a randomized controlled study. J Am Med Inform Assoc. 2012;19(e1):e145–e148. doi:10.1136/amiajnl-2011-000743 [PubMed: 22534081]
41. Wong A, Cao J, Lyons PG, et al. Quantification of Sepsis Model Alerts in 24 US Hospitals Before and During the COVID-19 Pandemic. JAMA Netw Open. 2021;4(11):e2135286. doi:10.1001/jamanetworkopen.2021.35286
42. Abràmoff MD, Cunningham B, Patel B, et al. Foundational Considerations for Artificial Intelligence Using Ophthalmic Images. Ophthalmology. 2022;129(2):e14–e32. doi:10.1016/j.ophtha.2021.08.023 [PubMed: 34478784]
43. Novis SJ, Havelka GE, Ostrowski D, et al. Prevention of thromboembolic events in surgical patients through the creation and implementation of a computerized risk assessment program. Journal of Vascular Surgery. 2010;51(3):648–654. doi:10.1016/j.jvs.2009.08.097 [PubMed: 20022209]
44. Theiling BJ, Donohoe R, Sendak M, et al. 2 Sepsis Watch: A Successful Deployment of a Deep Learning Sepsis Detection and Treatment Platform. Annals of Emergency Medicine. 2019;74(4):S1–S2. doi:10.1016/j.annemergmed.2019.08.005 [PubMed: 31655663]
45. Nahm M, Zhang J. Operationalization of the UFuRT methodology for usability analysis in the clinical research data management domain. J Biomed Inform. 2009;42(2):327–333. doi:10.1016/j.jbi.2008.10.004 [PubMed: 19026765]
46. Yuan MJ, Finley GM, Long J, Mills C, Johnson RK. Evaluation of user interface and workflow design of a bedside nursing clinical decision support system. Interact J Med Res. 2013;2(1):e4. doi:10.2196/ijmr.2402 [PubMed: 23612350]
47. Horsky J, Schiff GD, Johnston D, Mercincavage L, Bell D, Middleton B. Interface design principles for usable decision support: a targeted review of best practices for clinical prescribing interventions. J Biomed Inform. 2012;45(6):1202–1216. doi:10.1016/j.jbi.2012.09.002 [PubMed: 22995208]
48. Sendak MP, Ratliff W, Sarro D, et al. Real-World Integration of a Sepsis Deep Learning Technology Into Routine Clinical Care: Implementation Study. JMIR Med Inform. 2020;8(7):e15182. doi:10.2196/15182
49. Larsen K, Akindele B, Head H, et al. Developing a User-Centered Digital Clinical Decision Support App for Evidence-Based Medication Recommendations for Type 2 Diabetes Mellitus: Prototype User Testing and Validation Study. JMIR Hum Factors. 2022;9(1):e33470. doi:10.2196/33470
50. Rodbard D, Vigersky RA. Design of a decision support system to help clinicians manage glycemia in patients with type 2 diabetes mellitus. J Diabetes Sci Technol. 2011;5(2):402–411. doi:10.1177/193229681100500230 [PubMed: 21527112]
51. Zeier Z, Carpenter LL, Kalin NH, et al. Clinical Implementation of Pharmacogenetic Decision Support Tools for Antidepressant Drug Prescribing. Am J Psychiatry. 2018;175(9):873–886. doi:10.1176/appi.ajp.2018.17111282 [PubMed: 29690793]
52. Melnick ER, Dyrbye LN, Sinsky CA, et al. The Association Between Perceived Electronic Health Record Usability and Professional Burnout Among US Physicians. Mayo Clin Proc. 2020;95(3):476–487. doi:10.1016/j.mayocp.2019.09.024 [PubMed: 31735343]
53. Corrao NJ, Robinson AG, Swiernik MA, Naeim A. Importance of Testing for Usability When Selecting and Implementing an Electronic Health or Medical Record System. Journal of Oncology Practice. 2010;6(3):120–124. doi:10.1200/jop.200017 [PubMed: 20808553]
54. Nielsen J. Usability Testing. Usability Engineering. Published online 1993:165–206. doi:10.1016/b978-0-08-052029-2.50009-7
55. Kihara Y, Montesano G, Chen A, et al. J. Policy-Driven, Multimodal Deep Learning for Predicting Visual Fields from the Optic Disc and OCT Imaging. Ophthalmology. Published online 2022:781–791. doi:10.1016/j.ophtha.2022.02.017 [PubMed: 35202616]
56. Hemelings R, Elen B, Barbosa-Breda J, et al. Pointwise Visual Field Estimation From Optical Coherence Tomography in Glaucoma Using Deep Learning. TVST. Published online 2022. doi:10.1167/tvst.11.8.22

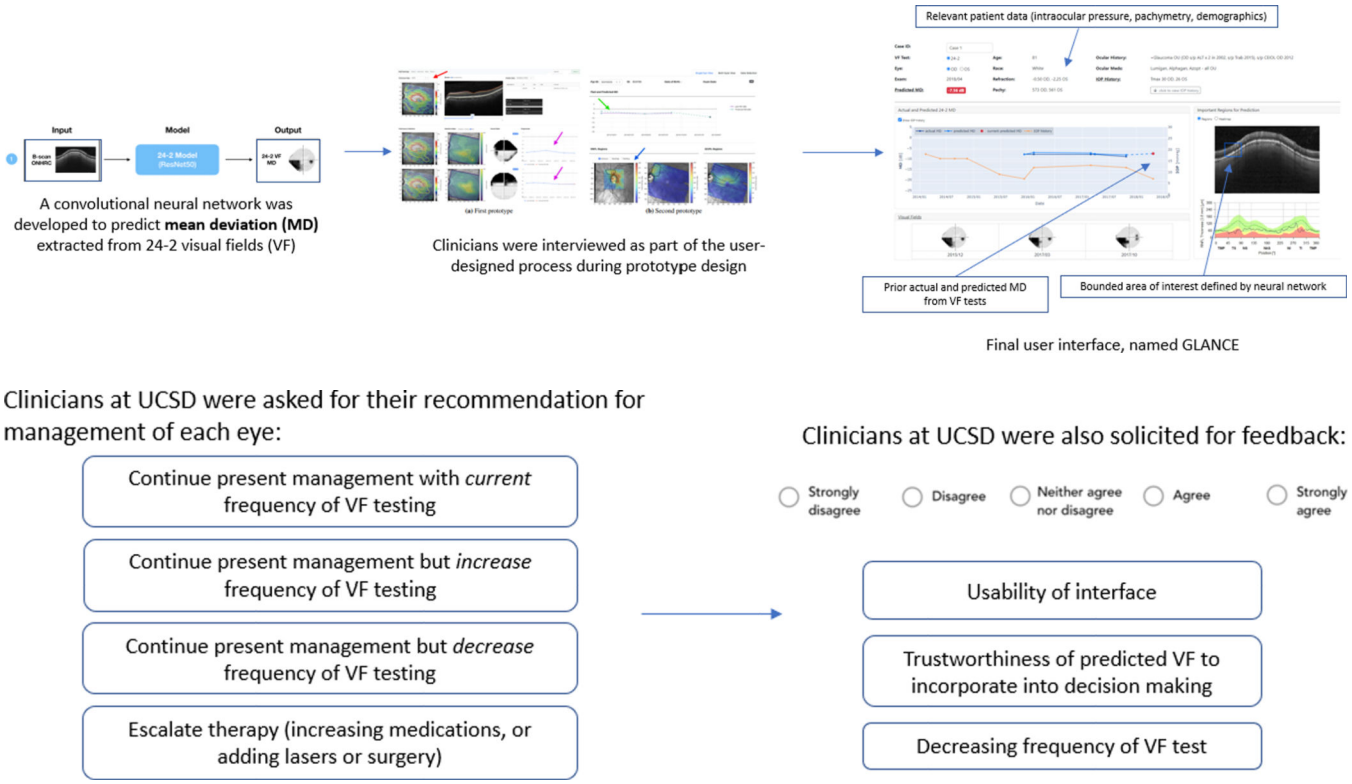

Figure 1. Design and implementation of the GLANCE Usability Study.

GLANCE is a graphical user interface designed to implement the output of an artificial intelligence (AI) model trained to predict visual field (VF) metrics such as mean deviation (MD) from Optical Coherence Tomography scans. **(A)**. The final interface was then evaluated by multiple clinicians surveyed at University of California San Diego for their management recommendations and their impressions of the AI model's predicted MD in terms of utility, trustworthiness, and its impact on their recommendations for VF testing frequency **(B)**.

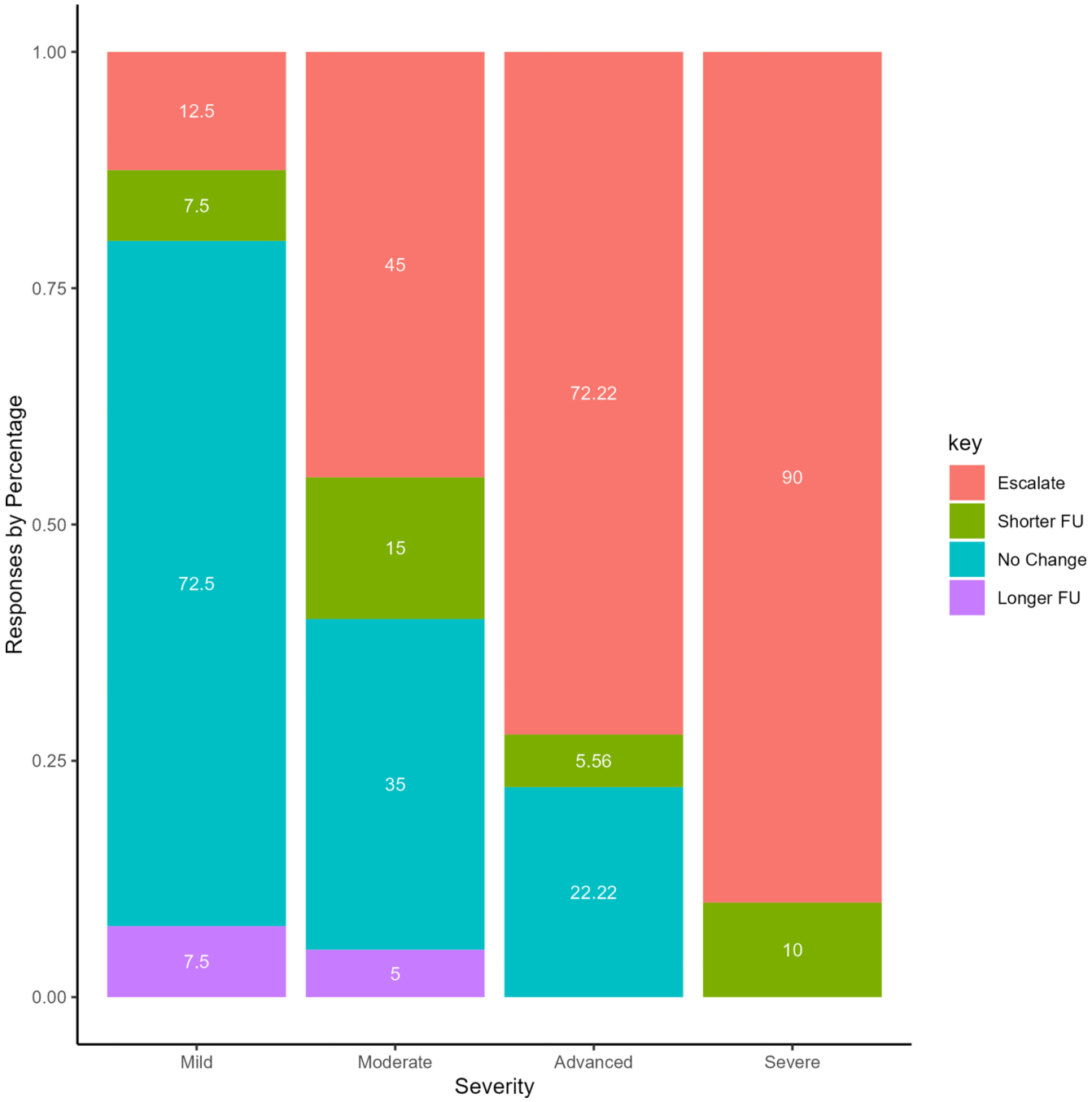

Figure 2. Severity of Glaucomatous Eye vs. Management Recommendations.

Clinicians generally chose to continue present management (i.e. no change to follow-up [FU] and visual field testing frequency) for milder disease and chose either shorter follow-up or escalating care for more severe disease.

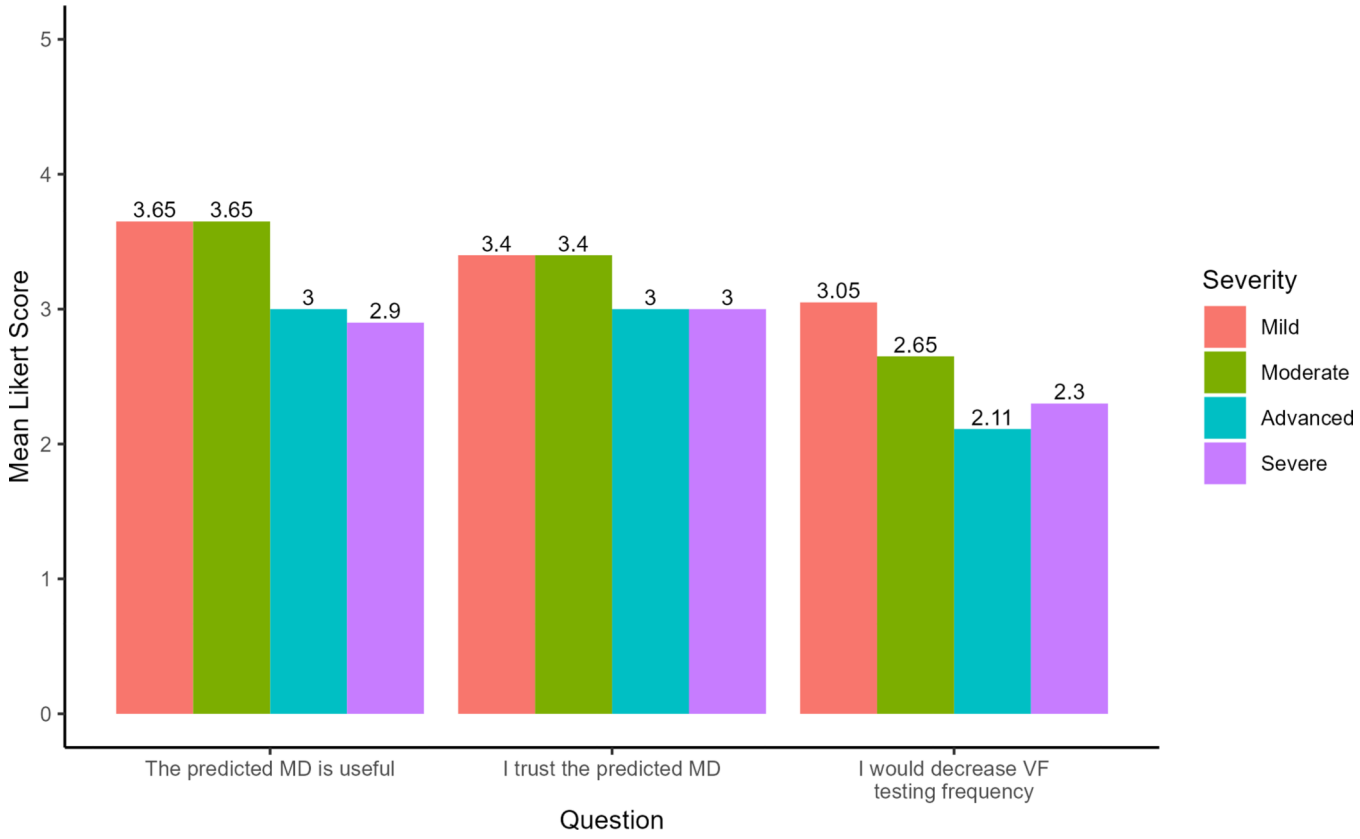

Figure 3. Mean Likert Scores for Trust and Usefulness of Interface and Decreasing Frequency of Testing, Stratified by Severity.

Clinicians generally trusted the predicted visual field output from the artificial intelligence model and found the output useful. However, their willingness to decrease frequency of testing was inversely correlated with severity of disease.

Table 1.

Cases Included in the GLANCE Usability Study.

Prior and predicted visual field metrics from the AI model, reported in mean deviation (MD), demographic data (i.e. age, race), and objective data relevant to glaucoma management included (refraction, pachymetry) were available to the clinician. Other data including optical coherence tomography (OCT) scans and ocular history/medications are not shown in this table, but are available in the full evaluation instrument provided in the supplement.

Case	Predicted Mean Deviation (Right Eye)	Predicted Mean Deviation (Left Eye)	Prior Mean Deviation (Right Eye)	Prior Mean Deviation (Left Eye)	Age	Race	Gender	Right Eye Spherical Equivalent (Diopters)	Left Eye Spherical Equivalent (Diopters)	Left Eye Pachymetry (μm)	Right Eye Pachymetry (μm)
1	−7.56	−2.47	−8.99	1.2	81	White	Male	−0.4	−2.25	573	561
2	−2.88	−2.97	−2.43	−3.2	82	White	Female	0	0	551	561
3	−2.83	−1.83	−6.28	−0.7	77	White	Female	−0.5	0	537	536
4	−22.62	−17.76	−21.84	−20.6	76	White	Male	1.125	0.625	495	498
5	−6.21	−3.76	−7.04	−6.05	63	White	Female	0.75	0.5	N/A	N/A
6	−12.75	N/A	−21	N/A	79	White	Female	−1.375	−1.375	476	475

Table 2.

Selected Comments from the Usability Study.

Clinicians were invited to give comments regarding the case or the GLANCE interface.

Case	User	Comment
1	1	Need additional info about validation of predicted MD. I like the concept but do not know to what patients it can be applied.
1	2	1) I don't like how the visits are displayed (year first and month second)... For whatever reason it is less clear to me 2) I am hesitant to add emphasis to the predicted MD as it is based upon some uncertain input. How do I know that the input is good? Is there a metric for the OCT quality that can be displayed. I know we have the one B-scan, but it is only 1 B-scan. The other scans could be riddled with artifacts. 3) I like the heatmap. This may reflect my lack of experience with AI, but when the heat map focuses on the choroid, I tend to want to consider the AI less.
2	1	Older patient who will outlive VF loss and loss of vision
2	2	I am influenced by how the heatmaps in Case 2 direct attention to the whole retina (not just the choroid) in the areas of attention. This makes me feel more positive about the predicted MD.
3	2	This case is tricky as the actual and predicted MDs appear to diverge with even further future improvement predicted. The gray scale alone OD is concerning but we don't have more info. Then the heat map is strange as there is much emphasis on non glaucoma regions.
4	2	IOP too high for my comfort level in this case regardless of prediction. Even if nothing happens in the short term (as per the prediction), over a longer scale (76 is still youngish), I prefer lower iop
5	2	OS: I am uncomfortable with IOP and IOP trend given paracentral defect. Predicted MD influences me although the heat map is again focused on the choroid OD: I'm also uncomfortable but less so than OS because there is no paracentral involvement at this time. Probably escalate therapy but OS is first and would discuss this with patient.
6	2	Confusing case... mostly only OD data. Is this a monocular patient? OS testing not possible?
Overall comment	2	I have never seen this before. It took 1–2 cases to click around and figure things out. Then it was ok. Decision making on glaucoma is also complex. I caught myself asking myself a lot about patients overall health. In the decision to "go" (escalate Rx or test more) vs "stay"... part of the rubrik is patient overall health and risk tolerance that isn't here. Having said that, I assume this all presents itself in the setting of the patient so that those factors can be uncovered by discussion.
Overall comment	3	Is there a better way to tell RNFL progression from this?
Overall comment	4	Helpful tool. I don't like that every time you touch the graph, the MD and IOP lines change.