The Effect of Artificial Intelligence Helpfulness and Uncertainty on Cognitive Interactions with Humans: A Randomized Controlled Trial

Chuan-Ching Tsai, Jin Yong Kim, Qiyuan Chen, Brigid Rowell, X Jessie Yang, Raed Al Kontar, Corey Lester

Submitted to: Journal of Medical Internet Research
on: May 16, 2024

Table of Contents

Original Manuscript 5
Supplementary Files 26
Figures 27
Figure 1 28
Figure 2 29
Figure 3 30
Figure 4 31
Figure 5 32
Figure 6 33

The Effect of Artificial Intelligence Helpfulness and Uncertainty on Cognitive Interactions with Humans: A Randomized Controlled Trial

Chuan-Ching Tsai[1] MHI; Jin Yong Kim[2] BEng; Qiyuan Chen[2] MSE; Brigid Rowell[1] MA; X Jessie Yang[2] PhD; Raed Al Kontar[2] PhD; Corey Lester[1] PhD, PharmD

[1]Department of Clinical Pharmacy College of Pharmacy University of Michigan Ann Arbor US
[2]Department of Industrial and Operations Engineering College of Engineering University of Michigan Ann Arbor US

Corresponding Author:
Corey Lester PhD, PharmD
Department of Clinical Pharmacy
College of Pharmacy
University of Michigan
428 Church Street
Ann Arbor
US

Abstract

Background: Clinical decision support systems (CDSS) leveraging artificial intelligence (AI) are increasingly integrated into healthcare practices, including pharmacy medication verification. Communicating uncertainty in an AI prediction is viewed as an important mechanism for boosting human collaboration and trust. Yet, little is known about the effects on human cognition as a result of interacting with such types of AI advice.

Objective: To evaluate the cognitive interaction patterns of pharmacists during medication product verification when using an AI prototype. Moreover, we examine the impact of AI's assistance — both helpful and unhelpful — and the communication of uncertainty of AI-generated results on pharmacists' cognitive interaction with the prototype and their performance.

Methods: In a randomized controlled trial, 30 pharmacists from professional networks each performed 200 medication verification tasks while their eye movements were recorded using a virtual eye tracker. Participants completed 100 verifications without AI assistance and 100 with AI assistance (either with simple help without uncertainty information, or advanced help which displays AI uncertainty). Fixation patterns (first and last areas fixated, number of fixations, fixation duration, and dwell times) were analyzed in relation to AI help type and helpfulness.

Results: Pharmacists shifted 19-26% of their total fixations to AI-generated regions when these were available, suggesting integration of AI advice in decision-making. AI assistance did not reduce the number of fixations on fill images, which remained the primary focus area. Unhelpful AI advice led to longer dwell times on reference and fill images, indicating increased cognitive processing. Displaying AI uncertainty led to longer cognitive processing times as measured by dwell times in original images.

Conclusions: Unhelpful AI increases cognitive processing time in the original images. Transparency in AI is needed in "black-box" systems, but showing more information can add cognitive burden. Therefore, the communication of uncertainty should be optimized and integrated into clinical workflows using user-centered design to avoid increasing cognitive load or impeding clinicians' original workflow.

(JMIR Preprints 16/05/2024:59946)
DOI: https://doi.org/10.2196/preprints.59946

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

The Effect of Artificial Intelligence Helpfulness and Uncertainty on Cognitive Interactions with Humans: A Randomized Controlled Trial

Chuan-Ching (Jonathan) Tsai, PharmD, MHI. College of Pharmacy, University of Michigan, Ann Arbor, MI, United States
Jin Yong Kim, BEng. Department of Industrial and Operations Engineering, College of Engineering, University of Michigan, Ann Arbor, MI, United States
Qiyuan Chen, MSE. Department of Industrial and Operations Engineering, College of Engineering, University of Michigan, Ann Arbor, MI, United States
Brigid Rowell, MA. Department of Clinical Pharmacy, College of Pharmacy, University of Michigan, Ann Arbor, MI, United States
X Jessie Yang, PhD. Department of Industrial and Operations Engineering, College of Engineering, University of Michigan, Ann Arbor, MI, United States
Raed Al Kontar, PhD. Department of Industrial and Operations Engineering, College of Engineering, University of Michigan, Ann Arbor, MI, United States
Corey Lester, PharmD, PhD. Department of Clinical Pharmacy, College of Pharmacy, University of Michigan, Ann Arbor, MI, United States

Abstract

Background: Clinical decision support systems (CDSS) leveraging artificial intelligence (AI) are increasingly integrated into healthcare practices, including pharmacy medication verification. Communicating uncertainty in an AI prediction is viewed as an important mechanism for boosting human collaboration and trust. Yet, little is known about the effects on human cognition as a result of interacting with such types of AI advice.

Objective: To evaluate the cognitive interaction patterns of pharmacists during medication product verification when using an AI prototype. Moreover, we examine the impact of AI's assistance — both helpful and unhelpful — and the communication of uncertainty of AI-generated results on pharmacists' cognitive interaction with the prototype and their performance.

Methods: In a randomized controlled trial, 30 pharmacists from professional networks each performed 200 medication verification tasks while their eye movements were recorded using a virtual eye tracker. Participants completed 100 verifications without AI assistance and 100 with AI assistance (either with simple help without uncertainty information, or advanced help which displays AI uncertainty). Fixation patterns (first and last areas fixated, number of fixations, fixation duration, and dwell times) were analyzed in relation to AI help type and helpfulness.

Results: Pharmacists shifted 19-26% of their total fixations to AI-generated regions when these were available, suggesting integration of AI advice in decision-making. AI assistance did not reduce the number of fixations on fill images, which remained the primary focus area. Unhelpful AI advice led to longer dwell times on reference and fill images, indicating increased cognitive processing. Displaying AI uncertainty led to longer cognitive processing times as measured by dwell times in original images.

Conclusions: Unhelpful AI increases cognitive processing time in the original images. Transparency in AI is needed in “black-box” systems, but showing more information can add cognitive burden. Therefore, the communication of uncertainty should be optimized and integrated into clinical workflows using user-centered design to avoid increasing cognitive load or impeding clinicians’ original workflow.

Keywords: clinical decision support system (CDSS); eye-tracking; medication verification; uncertainty visualization; AI helpfulness and accuracy

Introduction

Clinical decision support systems (CDSS) are tools that utilize medical knowledge and health information to aid clinicians’ decision-making to provide enhanced patient care [1,2]. CDSS can be classified into two types: knowledge-based and non-knowledge-based [2]. In knowledge-based CDSS relevant information is evaluated by a set of IF-THEN rules and recommendations are generated. Non-knowledge-based CDSS use artificial intelligence (AI) and/or machine learning (ML) methods rather than rules to evaluate information and generate recommendations [3]. Various CDSS have been designed to aid pharmacists’ clinical work, including checking for drug-drug interactions [4], antibiotics stewardship [5], drug utilization review [6], as well as medication product verification [7].

Pharmacists’ medication product verification is a pivotal yet time-consuming and vigilant process. Pharmacists spend 30% - 48% of their time verifying and dispensing medications [8,9]. Multiplying the time by the mean salary of a pharmacist [10], this equals to $ 38,823 - $ 62,117 monetary value per pharmacist per year. In addition to this task being time-consuming and costly, it also requires pharmacists to be vigilant, which can cause fatigue, cognitive overload, and even result in verification and dispensing errors. Medication dispensing error can be defined as “any deviations of the prescription order,” including but not limited to dispensing the wrong dosage, strength, dose form, or pharmaceutical ingredient [11]. The estimated rate of medication dispensing errors is 2.4 in 100 prescriptions [12] in community pharmacies, with the most common types being wrong ingredient, wrong strength, and labeling errors [12,13]. These errors can result in unfavorable therapeutic outcomes, which can have adverse effects on patient safety and outcomes.

Leveraging its capability to process large amounts of data, AI-based CDSS can help pharmacists reduce cognitive load and maintain vigilance. The term vigilance is used to describe the level of maintaining focus and alertness over a prolonged period of time [14]. Human vigilance tends to decrease over time, and there are multiple theories that explain the phenomenon, including the overload theory (reduction of information-processing abilities), underload theory (mindlessness due to repetitive nature), and fatigue [15–17]. AI-based tools may help users maintain vigilance in several mechanisms. Since AI has the power to process large amounts of information, it can reduce humans’ cognitive load and maintain their information-processing ability.

AI/ML-based CDSS are often "black-box" systems, where there is a lack of insight into the AI/ML prediction. As a result, there is an increasing call for greater AI transparency among medical professionals, emphasizing the need to display the uncertainty of AI-generated results [18,19]. While presenting this uncertainty may initially confuse users and require more cognitive effort to make decisions [20], it has significant benefits. Displaying uncertainty helps mitigate over-reliance on AI/ML and automation bias, fostering better human-AI collaboration and trust [20,21]. This transparency enables users to recognize that AI/ML predictions are not definitive, encouraging them to incorporate their judgment into decision-making. Additionally, showing uncertainty serves as a new stimulus that maintains user vigilance, preventing boredom or mindlessness as suggested by the underload theory [17].

AI/ML-based CDSS has the potential to improve medical care if the systems perform well and are appropriately implemented [22]. However, when the AI provides inaccurate, unreliable, and biased results, it can lead users to decision errors due to mechanisms such as automation bias and algorithmic aversion [23]. Automation bias is defined as people's heuristic belief that automation's performance is consistent when the system's performance is not always perfect, which can lead to users accepting incorrect advice or rejecting correct advice [23]. Conversely, algorithmic aversion reflects the tendency of users to not trust the algorithm's performance after seeing it make mistakes, even when its performance is better than human [24]. To prevent such human errors in clinical decision making with the help of AI, high-performing and reliable systems are pivotal. In this study, we assess the impact of helpful and unhelpful AI on user performance.

Previously, our group developed an AI prototype aiming to help pharmacists with medication product verification and reduce dispensing errors. Using the participatory design method, researchers conducted focus groups with pharmacists who had medication verification experience to understand the workflow, difficulties and concerns of the current process, and ideas about incorporating AI to aid this process. Based on pharmacists' feedback, a user-centered AI interface emphasizing clarity and accessibility was developed. The development of the AI help software is detailed in previous research by Zheng et al [7]. The purpose of this paper is to establish how pharmacists cognitively incorporate and utilize an AI prototype during medication product verification process by analyzing eye-tracking fixation data.

Materials and Methods

Participants and Trial Design

This study was a randomized controlled trial. We recruited 30 pharmacists as participants. Two listservs for professional pharmacists, the Minnesota Pharmacy Practice-Based Research Network and the University of Michigan College of Pharmacy Preceptor Network, were used to recruit participants. The listserv managers sent a recruitment email instructing interested pharmacists to contact the study team to schedule a screening phone call. The study's inclusion criteria were: 1) a licensed pharmacist in the United States, 2) at least 18 years old at the time of screening, and 3) access to a laptop or desktop computer with a webcam to complete the experiment. The exclusion criteria were: 1) require assistive technology to use the computer, 2) need eyeglasses with more than one power to complete the experiment, 3) have uncorrected cataracts, intraocular implants, glaucoma, or permanently dilated pupils, or 4) eye movement or alignment abnormalities. A random number generator created in R assigned each participant a number from 1-8. The probability of being assigned to simple help trials or advanced help trials was equal for all participants as was the probability of completing the first 100 trials with AI help or without AI help. In each trial, participants would check the images on the screen and decide whether it was a good fill (the medication in the fill image is identical to the medication in the reference image) or a bad fill (the medications in the fill image and the reference image are different) and hit either the reject or the accept button.

The Task: pharmacists' medication verification

Pharmacists are responsible for dispensing the correct medication to patients to ensure optimal therapeutic outcome. Before dispensing the medications to patients, pharmacists have to visually inspect the medications being filled and compare them to the prescribed drugs. This process is called medication product verification. In the study, each participant performed 200 medication product verification tasks, where their eye movement was recorded using an eye tracker. Each participant conducted 100 medication verifications without AI assistance and 100 verifications according to their assigned AI help group. A screenshot of the participants' view is in Figure 1. To avoid carry-over effect, the medications being verified in no AI help trials were different from those with AI help. However, the medications were identical in trials with simple AI help and trials with advanced AI help.

The AI help software

The AI software used a Bayesian neural network [25] to predict the prescribed and dispensed drug's National Drug Code (NDC) based on the reference and fill images, respectively. The AI software then compares the two NDCs and generates insights on the match status. The final interface had two (2) AI-generated regions that were designed to aid pharmacists' medication verification - the AI match plot and AI histogram. The AI match plot showed the match status of four (4) characteristics between the AI-predicted NDC and the expected NDC on the prescription: imprint, color, shape, and score. The unmatched characteristics were denoted by a red "X", while matched ones by a green check. Figure 1 shows a prediction with four (4) green checks.

The AI histogram diaplayed the probability distribution from 50 AI predictions, each representing the AI's estimated likelihood that its predicted NDC matches the expected NDC. These probabilities form the histogram, illustrating the AI software's "confidence" in the match status. A green peak in the histogram indicates consistent results across the 50 simulations, suggesting high confidence. Conversely, a flat, colorful distribution signifies low confidence. In Figure 1, a highly confident prediction is displayed.

Figure 1. The medication verification system interface with AI Help. In trials without AI assistance, pharmacists only have access to the reference image (refImage) and fill image. With simple AI help, pharmacists saw an AI match plot, which shows AI's stance on the match status for four key medication characteristics (imprint, color, shape, and score). In advanced AI help, in addition to the match plot, pharmacists also saw the AI histogram, which displays the probability distribution from 50 predictions, indicating the uncertainty of the prediction.

Trial Type

Each trial was configured with specific variables, including type of AI assistance, case type, and participant performance. There were three categories of AI assistance: no help, simple help, and advanced help. In the no help condition, participants utilized only the reference image, fill image, and text. Simple help added an AI-generated match plot, while advanced help also included an AI-generated probability histogram. The case type, relevant only when AI assistance was provided, referred to the alignment between the AI's recommendation and the correct answer. There were two scenarios: helpful advice, where the AI's guidance matched the correct action (accept or reject), and unhelpful advice, where the AI's suggestion opposed the correct answer.

Eye-tracking data

Eye-tracking data were collected using software from Labvanced (Paderborn, Germany). Labvanced uses deep learning models to process webcam videos, allowing virtual eye movement tracking. The accuracy and precision are comparable to in-laboratory "gold standard" eye trackers and are verified in a peer-reviewed paper [26]. Participants logged on to the Labvanced Trials website. First, they verified the webcam accurately captured their face and agreed to the recording and data use policy before starting the trial. Participants were then prompted to complete a demographics questionnaire that included questions such as age, gender, race, pharmacy practice experience level, practice setting, and their trust level in automated systems.

The first phase of the trial was the calibration. During calibration, participants were first asked to measure the distance of their face to the screen, and set the center pose that worked best. The actual calibration included two parts - the position and orientation, and fixation to certain points on the screen. Participants were asked to follow and fixate on two series of red dots on the screen while maintaining their face at the center pose. After the calibration was completed, participants started the actual trials. The system continuously monitored participants' facial position and orientation. If participants moved out of the center pose, they were prompted to re-align their face to the center pose and rerun the fixation calibration. The raw eye-tracking data were processed by the Labvanced algorithm [27] to fixation data. Our analyses were based on the fixation data to prevent counting in eye movement in the saccades, where participants were simply transitioning from one fixation to the next and not necessarily fixating in the region.

Data analysis & Preprocessing

Fixation analysis is the most common metric for analyzing eye-tracking data. The fixation rate, fixation duration, and dwell time have been proposed to reflect human cognitive interest in certain areas of interests (AOI) [28–30]. Higher fixation rate and longer dwell time indicate repeated interest in a certain area. The longer fixation duration indicates higher cognitive load. To characterize the utilization pattern, we are also interested in the order in which pharmacists use these images. Thus, we report the first fixation region, the last fixation region, the number of fixations in each region, the average fixation duration of each fixation, and the dwell time, which was calculated by the sum of the duration of all fixations in each region in each trial. We further stratified the data based on different AI help types (i.e. simple versus advanced help), and case types (helpful versus unhelpful advice).

To eliminate calibration failure and measurement error, we calculated the modified z score of each region's dwell time in each trial. The modified z-score method is more robust in detecting outliers since it uses median instead of mean in the calculation of z-scores. If the modified z score of the dwell time in that region is greater than 3.5, we labeled that observation as an outlier and excluded it from our analyses.

Results

First and Last Fixation Area

After outlier removal, the dataset comprises 2449 trials without AI assistance, 1365 with simple AI help, and 1391 with advanced AI help. In trials without AI Help, the first fixation area (Figure 2, left) was: reference image (n = 1822, 74.4%), fill image (n = 627, 25.6%). In simple help, reference image (n = 622, 45.6%), fill image (n = 174, 12.7%), AI Match plot (n = 569, 41.7%). In advanced help, reference image (n = 469, 33.7%), fill image (n = 270, 19.4%), AI Match plot (n = 581, 41.8%), AI Histogram (n = 71, 5.1%). For trials without AI help, the last fixation area (Figure 2, right) was: reference image (n = 822, 33.6%), fill image (n = 1627, 66.4%). In simple help, reference image (n = 276, 20.2%), fill image (n = 651, 47.7%), AI Match plot (n = 438, 32.1%). In advanced help, reference image (n = 146, 10.5%), fill image (n = 339, 24.4%), AI Match plot (n = 202, 14.5%), AI Histogram (n = 704, 50.6%).

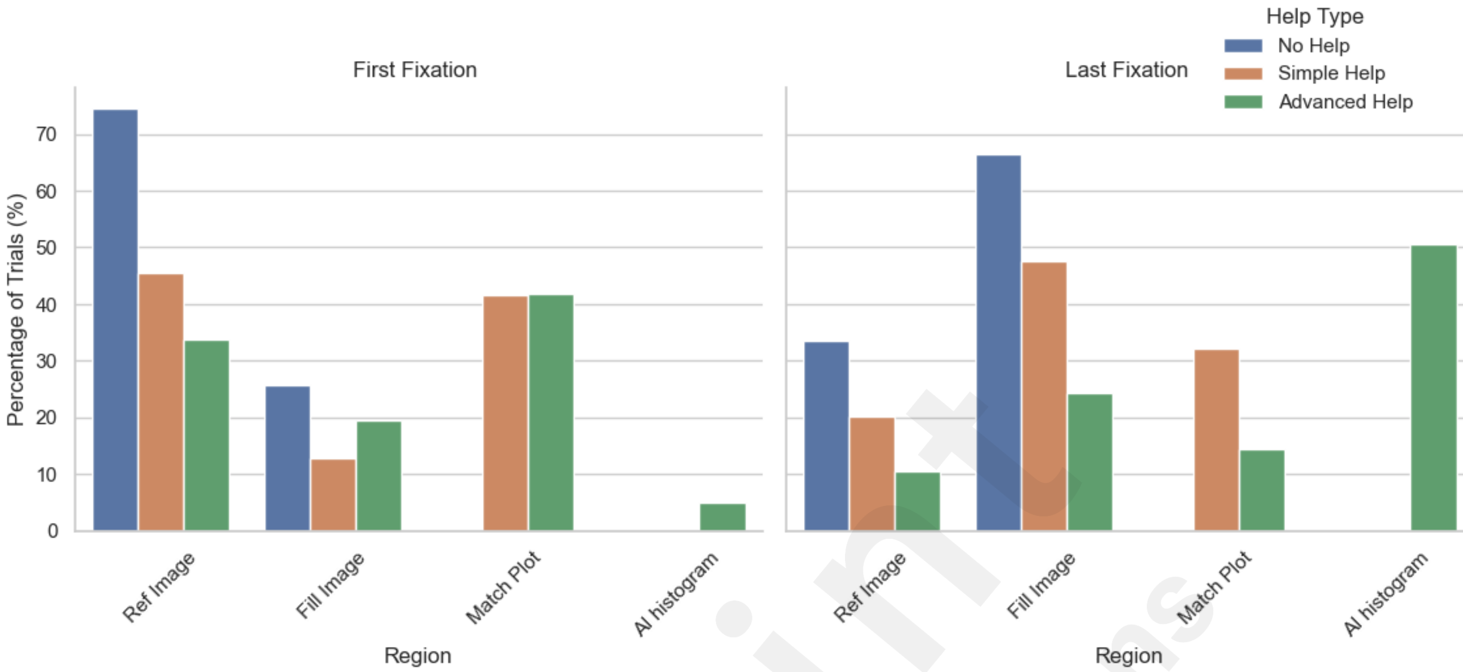


Figure 2. Percentage of Trials by Region of Interest and AI Help Type. Left: First Fixation, Right: Last Fixation

Number of fixations

Irrespective of help type, fill images had the highest number of fixations in a trial, and the number of fixations per trial did not decrease after the introduction of AI generated areas (mean (S.D.) number of fixation: no help: 4.44 (2.92), simple help: 4.33 (2.90), advanced help: 5.14 (3.43)). The AI match plot and AI histogram received a lower number of fixations. The mean (S.D.) number of fixations in AI match plots in simple and advanced help trials are 1.91 (1.12) and 1.65 (0.82), respectively. The mean (S.D.) number of fixations in AI histogram in advanced help trials is 1.69 (0.87). The numbers are summarized in Table 1, and a box plot representation of the result is in Figure 3.

Help Type	Region of Interest	Number of Fixations (%)
No AI Help	Reference Image	5246 (35.0%)
	Fill Image	9756 (65.0%)
Simple Help	Reference Image	2344 (27.7%)
	Fill Image	4547 (53.7%)
	Match Plot	1583 (18.7%)
Advanced Help	Reference Image	2729 (23.9%)
	Fill Image	5729 (50.2%)
	Match Plot	1382 (12.1%)
	Histogram	1566 (13.7%)

Table 1. Number of Fixations (Percentages) For Each Region And Help Type.

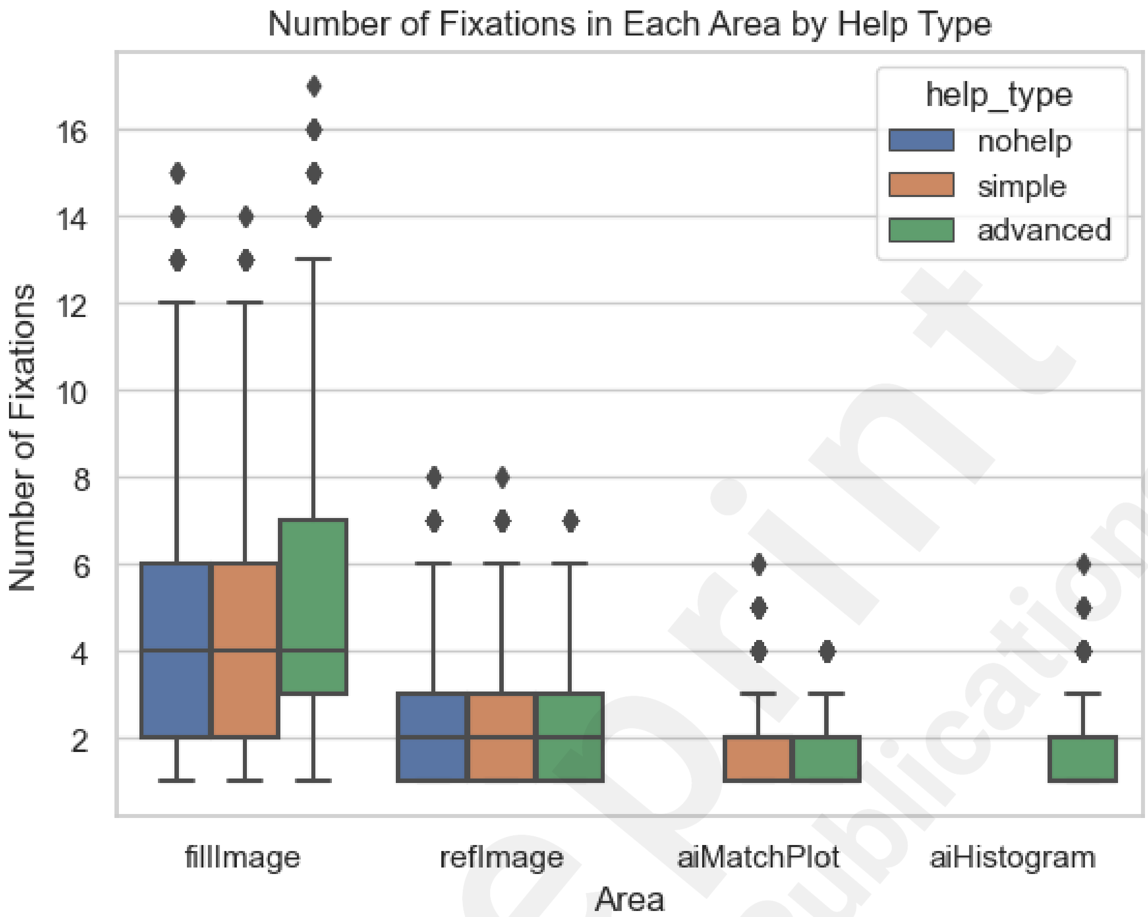


Figure 3. Boxplot Representation of Number of Fixations In Each Area Per Trial, Categorized by Help Type.

Fixation Duration

The fixation duration in each region was similar, regardless of help type. The mean duration of fixation was around 0.2 seconds, with a standard deviation of 0.055 seconds. This indicates that the cognitive load for processing information in each area was similar. The boxplot of the finding is in Figure 4.

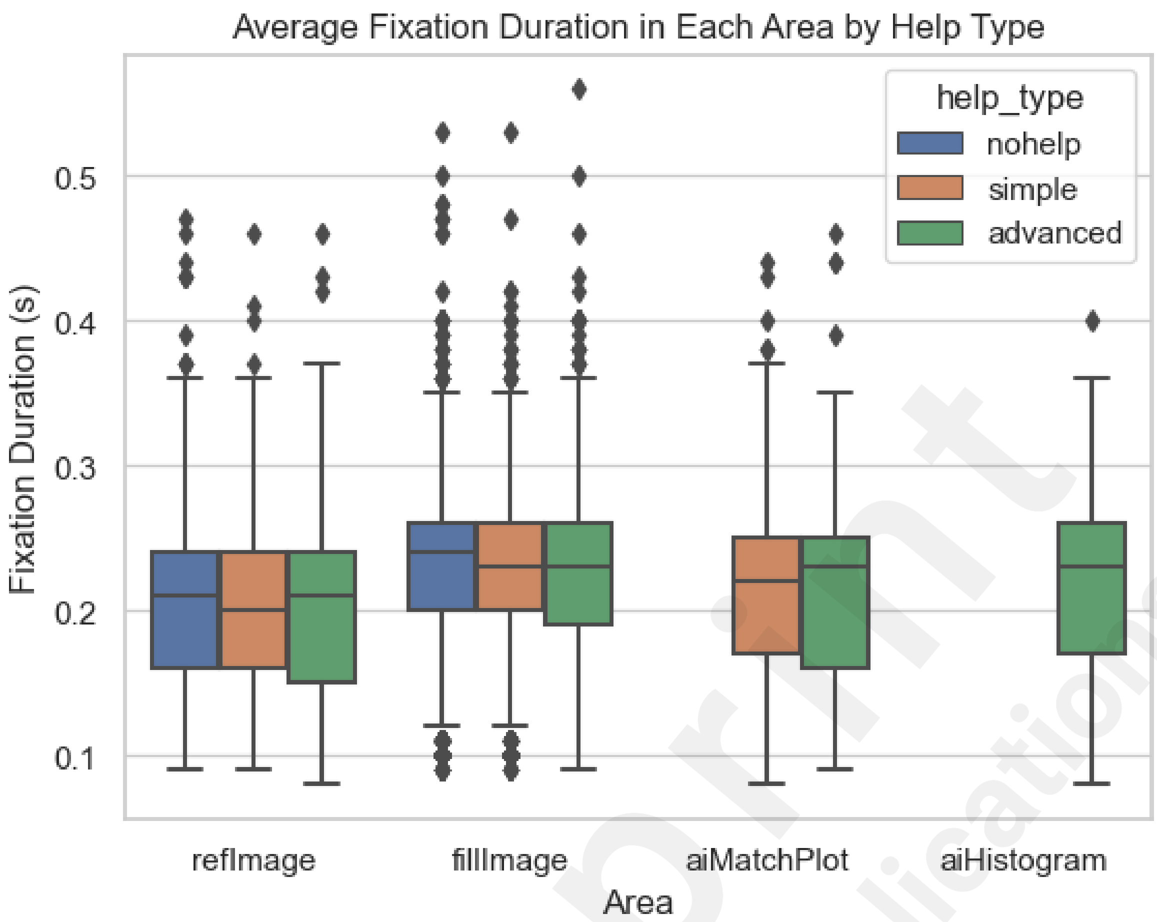


Figure 4. Boxplot Representation of Fixation Duration in Each Area, Categorized by Help Type

Dwell times

Participants consistently spent the most time on fill images, with mean (S.D.) dwell times (in seconds) of 0.99 (0.67) in no help trials, 0.96 (0.66) in simple help trials, and 1.13 (0.77) in advanced help trials. For the reference image, mean (S.D.) dwell times (in seconds) was 0.50 (0.29) in no help trials, 0.48 (0.29) in simple help trials, and 0.49 (0.28) in advanced help trials. The AI match plots had mean (S.D.) dwell times of 0.4 (0.25) and 0.35 (0.18) seconds in simple and advanced help trials, respectively. The AI histograms had mean (S.D.) dwell time of 0.36 (0.19) seconds in advanced help trials. (Figure 5)

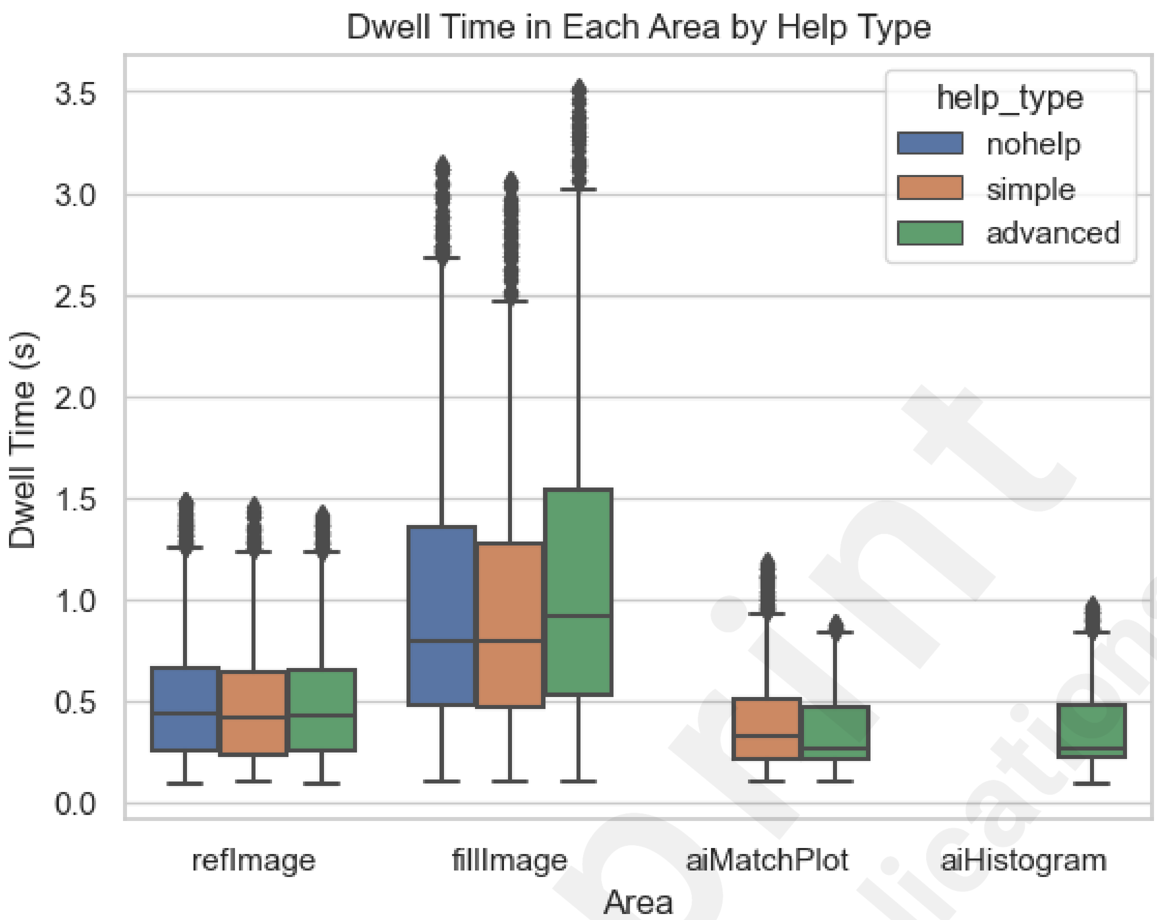


Figure 5. Boxplot Representation of Dwell Time in Each Area, Categorized by Help Type

The pairwise t-test result showed no significant difference between the AI help types and the dwell time in reference images. However, for fill images, there was a significant increase in dwell times for advanced help trials compared to no help trials (Bonferroni adjusted *P*<.001) or simple help trials (Bonferroni adjusted *P*<.001). The difference in fill image dwell times was non-significant between no help and simple help (Bonferroni adjusted *P*=.71) (Figure 5).

Dwell times in fill images and reference images were significantly longer in trials with unhelpful advice compared to helpful advice (fill image, mean: 1.36 vs 0.99 sec, *P*<.001; ref image, mean: 0.56 vs 0.47 sec, *P*<.001) (Figure 6).

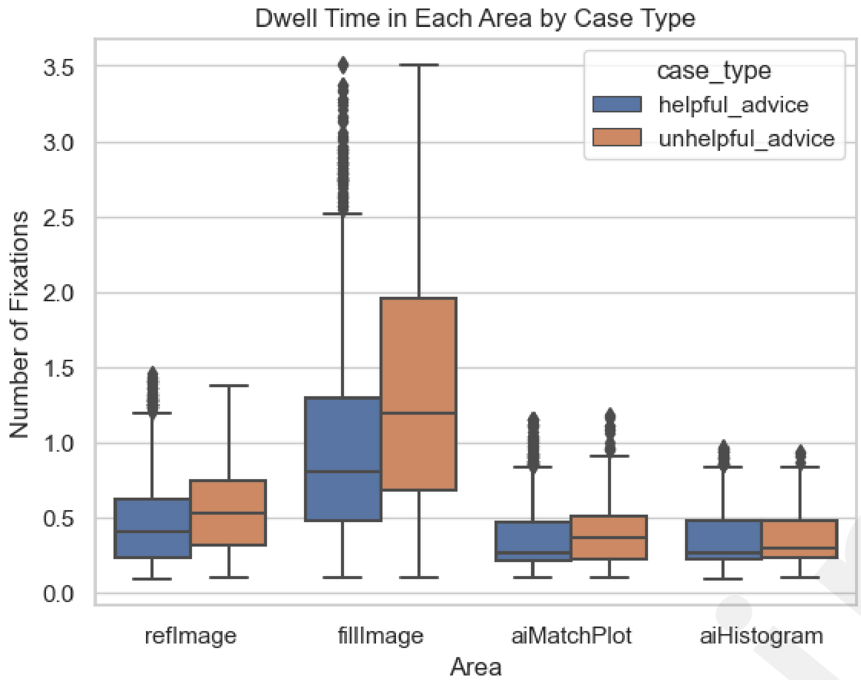


Figure 6. Dwell Time in Each Area By Case Type. Compared to helpful advice trials, those with unhelpful advice resulted in significantly longer dwell times in fill and reference images.

Discussion

Our study presents critical insights into the interaction between healthcare professionals and AI-based CDSS. Our findings highlight a shift in the visual and cognitive engagement of pharmacists when AI-driven tools are employed, resulting in a change of traditional medication product verification workflow. Although communicating the uncertainty of AI has been shown to have benefits such as preventing over-reliance and fostering trust in AI [18,20], our study found that it resulted in longer cognitive-processing time in the original images. The criticality of AI accuracy is underscored by our findings, as AI advice of varying correctness significantly impacts cognitive processing time of users. These findings suggest the potential human-AI collaboration to enhance healthcare delivery, if AI is accurate, reliable, and deployed properly without interfering with healthcare professionals' existing workflow or increasing their cognitive load.

AI-based CDSS influenced participants' drug verification processes. When participants had access to AI, 19% to 26% of total fixations were shifted to AI regions, and there was a decrease in fixation in the original (fill and reference) images. This indicates that the AI tool changed how participants process information, and participants incorporate AI-generated information in their decision-making process. The decrease in fixation counts in original images and the shift to AI-generated regions indicate that AI may support users in completing the task. When AI is available, users tend to look at the AI match plot first, and then move on to the fill and reference images. This suggests that providing a simple graphical "summary" of the AI advice may aid users in their decision-making and help them maintain vigilance. User-centered participatory design informed the development of this AI prototype [7]. In this use case, pharmacists emphasized the need for simple and accessible information to guide them in completing the task. To avoid overwhelming pharmacists with redundant information, the match plot only displays AI's stance on color, shape, imprint, and score, which is the most important information for pharmacists in the medication verification process [7]. By incorporating user feedback and preferences into the design process, developers can create systems that align more closely with user needs and preferences, ultimately enhancing usability and effectiveness. In conclusion, prioritizing user-centered design principles not only improves the performance of AI-based systems but also ensures that these systems effectively augment rather than disrupt user workflows.

Irrespective of AI intervention, participants allocated the majority of their fixation time to fill images. The fact that pharmacists need to inspect tablets from various angles to verify color, shape, and imprint for accurate identification may give rise to longer cognitive processing times. On the other hand, AI-generated regions have shorter dwell times. This might stem from the graphical simplicity of these images [31], or indicate that users correctly perceive the AI-generated regions as supportive in completing the task. Compared to not showing AI outputs' uncertainty (simple or advanced), we found that displaying uncertainty results in significantly longer dwell times in the original (fill) images. This may suggest that users are confused, or double-guess the AI's correctness when they see the uncertainty histogram, leading to a need to go back to the fill images and verify the correctness of the advice.

Since many AI algorithms are essentially a "black box", there is a demand for AI transparency [21,32]. Displaying more information about the advice can decrease user vigilance since more details need to be processed. In designing AI-based systems, there is a need to balance the transparency and explainability of AI output and the amount of information presented to the users. In this use case, a more graphically-simple alternative to the histogram may be considered to avoid overwhelming the users. According to Prabhudesai et al. [20], users develop their expertise in explaining the uncertainty plot over time, which resolved the initial confusion. A strategy to assist in users' learning is to provide training on how to interpret AI's uncertainty. The lack of benefit observed in this use case can also be due to the nature of the medication verification task since this is a relatively simple task. Too much complexity, such as an uncertainty histogram, might not be suitable. In the pharmacy, pharmacists are often tasked with additional responsibilities, which interfere with their vigilance in performing the verification task [33]. Simplifying the communication of the uncertainty might balance the benefits of this kind of advice from AI.

The results revealed that AI's correctness (i.e., helpfulness) had a significant impact on users' cognitive processing time. Unhelpful advice resulted in significantly longer dwell times in reference images and fill images compared to helpful advice. This finding supports that users are susceptible to automation bias, the heuristic tendency of users to favor the suggestion made by automated systems [23], and underlines the importance of AI's correctness and performance. Developers should validate AI's accuracy and reliability before implementing it in the real world, for the underperformance of AI can introduce new user errors through mechanisms such as automation bias and might harm humans' trust in AI, slowing users down. The International Medical Device Regulators Forum guideline mandates a thorough clinical evaluation of Software as a Medical Device (SaMD) to ensure it meets its intended purpose and delivers clinically-meaningful results [34]. This includes assessing accuracy, reliability, and other performance metrics. Park et al. [35] introduced a framework for AI-based medical products, emphasizing phases akin to drug trials: balancing benefits and risks, confirming usability through methods like A/B testing, conducting large-scale trials to validate effectiveness, and ongoing monitoring for self-improvement. Like medications, AI-driven CDSS can profoundly impact patient care, necessitating rigorous pre and post-deployment validation.

Limitations and Future Work

Several other eye-tracking metrics have been proposed to assess cognitive load. Voluntary movements such as saccade length and saccade velocity, and involuntary movements such as pupil dilation can reflect cognitive load. Due to the lack of data availability, we did not include these metrics in our analysis. Also, both qualitative methods, such as interviews, and quantitative methods, such as surveys, can provide insights into how participants perceive these AI-generated regions. This may help inform the design and implementation of AI tools in augmenting healthcare professionals' workflow. Future studies can include these eye-tracking metrics and qualitative and/or quantitative methods to further understand user experience and mental workload.

Conclusions

The goal of this study was to assess how pharmacists cognitively incorporate and utilize an AI prototype during the medication product verification process, using eye-tracking data to analyze their interaction with the system. The study revealed that AI-based CDSS can alter traditional workflow and cognitive engagement, with pharmacists allocating significant focus to AI-generated regions, which indicates the integration of AI advice in decision-making. However, communicating AI uncertainty by displaying a probability histogram increased cognitive processing time for original images. The correctness of AI suggestions directly affects cognitive processing, with helpful AI advice reducing and unhelpful AI advice increasing it. The findings underscore the importance of accurate, reliable AI in healthcare and suggest that user-centered design and AI transparency are crucial for effective human-AI collaboration.

Reference

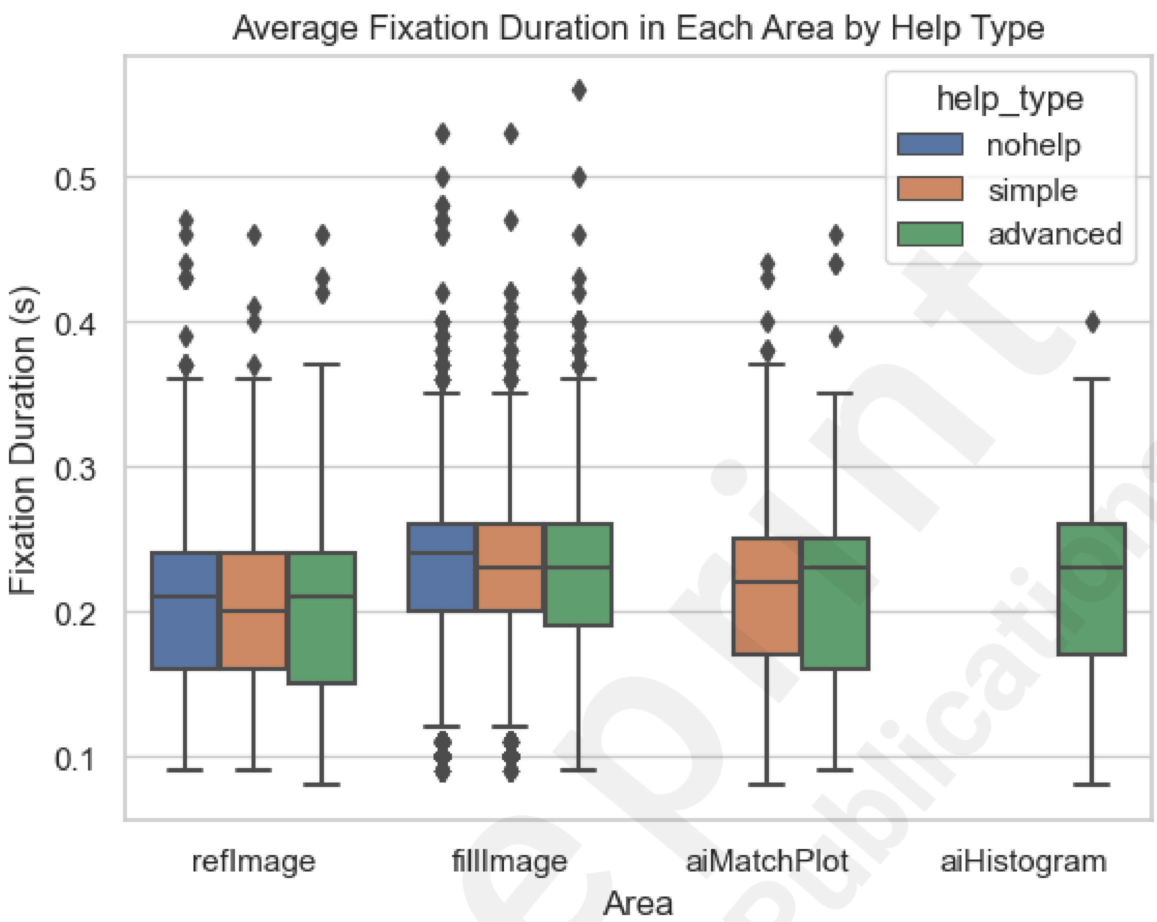
1. Moon J, Chladek JS, Wilson P, Chui MA. Clinical decision support systems in community pharmacies: a scoping review. J Am Med Inform Assoc 2024 Jan 1;31(1):231–239. doi: 10.1093/jamia/ocad208
2. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. Npj Digit Med Nature Publishing Group; 2020 Feb 6;3(1):1–10. doi: 10.1038/s41746-020-0221-y
3. Berner ES, editor. Clinical decision support systems: theory and practice. 2. ed. New York, NY [Heidelberg]: Springer; 2007. ISBN:978-0-387-33914-6
4. Smithburger PL, Buckley MS, Bejian S, Burenheide K, Kane-Gill SL. A critical evaluation of clinical decision support for the detection of drug–drug interactions. Expert Opin Drug Saf Taylor & Francis; 2011 Nov 1;10(6):871–882. PMID:21542665
5. Madaras-Kelly KJ, Hannah EL, Bateman K, Samore MH. Experience with a clinical decision support system in community pharmacies to recommend narrow-spectrum antimicrobials, nonantimicrobial prescriptions, and OTC products to decrease broad-spectrum antimicrobial use. J Manag Care Pharm JMCP 2006 Jun;12(5):390–397. PMID:16792446
6. Armstrong Edward P, Markson Tracy J. On-Line OUR Messages: Pharmacists' Attitudes and Actions in Response: Of the DUR messages studied, pharmacists found Drug-Drug Interactions most useful. J Am Pharm Assoc 1996 1997 May 1;37(3):315–320. doi: 10.1016/S1086-5802(16)30215-7
7. Zheng Y, Rowell B, Chen Q, Kim JY, Kontar RA, Yang XJ, Lester CA. Designing Human-Centered AI to Prevent Medication Dispensing Errors: Focus Group Study With Pharmacists. JMIR Form Res 2023 Dec 25;7(1):e51921. doi: 10.2196/51921
8. Karia A, Norman R, Robinson S, Lehnbom E, Laba T-L, Durakovic I, Balane C, Joshi R, Webster R. Pharmacist's time spent: Space for Pharmacy-based Interventions and Consultation TimE (SPICE)-an observational time and motion study. BMJ Open 2022 Mar 2;12(3):e055597. PMID:35236731
9. Hohmeier KC, Wheeler J, Heintz K, Gatwood J. Community pharmacist workflow and medication therapy management delegation: An assessment of preferences and barriers. J Am

Pharm Assoc JAPhA 2020;60(6):e215–e223. PMID:32811750
10. U.S. Bureau of Labor Statistics. Occupational Employment and Wage Statistics. Bur Labor Stat. 2019. Available from: https://www.bls.gov/oes/2019/may/oes291051.htm [accessed Apr 22, 2024]
11. Szeinbach S, Seoane-Vazquez E, Parekh A, Herderick M. Dispensing errors in community pharmacy: perceived influence of sociotechnical factors. Int J Qual Health Care J Int Soc Qual Health Care 2007 Aug;19(4):203–209. PMID:17567597
12. Um IS, Clough A, Tan ECK. Dispensing error rates in pharmacy: A systematic review and meta-analysis. Res Soc Adm Pharm 2024 Jan 1;20(1):1–9. doi: 10.1016/j.sapharm.2023.10.003
13. Aldhwaihi K, Schifano F, Pezzolesi C, Umaru N. A systematic review of the nature of dispensing errors in hospital pharmacies. Integr Pharm Res Pract 2016;5:1–10. PMID:29354533
14. Parasuraman R, Mouloua M. Interaction of signal discriminability and task type in vigilance decrement. Percept Psychophys 1987 Jan 1;41(1):17–22. doi: 10.3758/BF03208208
15. Warm JS, Dember WN, Hancock PA. Vigilance and workload in automated systems. Autom Hum Perform Theory Appl Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc; 1996. p. 183–200.
16. Manly T, Robertson IH, Galloway M, Hawkins K. The absent mind: further investigations of sustained attention to response. Neuropsychologia 1999 Jun;37(6):661–670. PMID:10390027
17. Thomson DR, Besner D, Smilek D. A resource-control account of sustained attention: evidence from mind-wandering and vigilance paradigms. Perspect Psychol Sci J Assoc Psychol Sci 2015 Jan;10(1):82–96. PMID:25910383
18. Cai CJ, Winter S, Steiner D, Wilcox L, Terry M. “Hello AI”: Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. Proc ACM Hum-Comput Interact 2019 Nov 7;3(CSCW):104:1-104:24. doi: 10.1145/3359206
19. Jussupow E, Spohrer K, Heinzl A, Gawlitza J. Augmenting Medical Diagnosis Decisions? An Investigation into Physicians’ Decision-Making Process with Artificial Intelligence. Inf Syst Res INFORMS; 2021 Sep;32(3):713–735. doi: 10.1287/isre.2020.0980
20. Prabhudesai S, Yang L, Asthana S, Huan X, Liao QV, Banovic N. Understanding Uncertainty: How Lay Decision-makers Perceive and Interpret Uncertainty in Human-AI Decision Making. Proc 28th Int Conf Intell User Interfaces New York, NY, USA: Association for Computing Machinery; 2023. p. 379–396. doi: 10.1145/3581641.3584033
21. von Eschenbach WJ. Transparency and the Black Box Problem: Why We Do Not Trust AI. Philos Technol 2021 Dec 1;34(4):1607–1622. doi: 10.1007/s13347-021-00477-0
22. Wolfstadt JI, Gurwitz JH, Field TS, Lee M, Kalkar S, Wu W, Rochon PA. The Effect of Computerized Physician Order Entry with Clinical Decision Support on the Rates of Adverse Drug Events: A Systematic Review. J Gen Intern Med 2008 Apr 1;23(4):451–458. doi: 10.1007/s11606-008-0504-5
23. Jones-Jang SM, Park YJ. How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. J Comput-Mediat Commun 2023 Jan 1;28(1):zmac029. doi: 10.1093/jcmc/zmac029
24. Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: People erroneously avoid algorithms after seeing them err. J Exp Psychol Gen US: American Psychological Association; 2015;144(1):114–126. doi: 10.1037/xge0000033
25. Wang H, Yeung D-Y. A Survey on Bayesian Deep Learning. ACM Comput Surv 2020 Sep 28;53(5):108:1-108:37. doi: 10.1145/3409383
26. Kaduk T, Goeke C, Finger H, König P. Webcam eye tracking close to laboratory standards: Comparing a new webcam-based system and the EyeLink 1000. Behav Res Methods 2023 Oct 11; doi: 10.3758/s13428-023-02237-8
27. Labvanced. Labvanced/eye-tracking-fixation-detection. 2023. Available from: https://github.com/Labvanced/eye-tracking-fixation-detection [accessed Apr 22, 2024]

28. Zu T, Hutson J, Loschky LC, Rebello NS. Use of Eye-Tracking Technology to Investigate Cognitive Load Theory. arXiv; 2018. doi: 10.48550/arXiv.1803.02499
29. Chen S, Epps J, Ruiz N, Chen F. Eye activity as a measure of human mental effort in HCI. Proc 16th Int Conf Intell User Interfaces New York, NY, USA: Association for Computing Machinery; 2011. p. 315–318. doi: 10.1145/1943403.1943454
30. Joseph AW, Murugesh R. Potential Eye Tracking Metrics and Indicators to Measure Cognitive Load in Human-Computer Interaction Research. J Sci Res 2020;64(01):168–175. doi: 10.37398/JSR.2020.640137
31. Wang Q, Yang S, Liu M, Cao Z, Ma Q. An eye-tracking study of website complexity from cognitive load perspective. Decis Support Syst 2014 Jun 1;62:1–10. doi: 10.1016/j.dss.2014.02.007
32. Pedreschi D, Giannotti F, Guidotti R, Monreale A, Ruggieri S, Turini F. Meaningful Explanations of Black Box AI Decision Systems. Proc AAAI Conf Artif Intell 2019 Jul 17;33(01):9780–9784. doi: 10.1609/aaai.v33i01.33019780
33. Arya V, Bakken BK, Doucette WR, Gaither CA, Kreling DH, Mott DA, Schommer JC, Witry MJ. 2022 National Pharmacist Workforce Study Final Report. American Association of Colleges of Pharmacy; 2023. Available from: https://www.aacp.org/sites/default/files/2023-08/2022-npws-final-report.pdf
34. Stewart JP. Software as a Medical Device (SaMD): Clinical Evaluation. International Medical Device Regulators Forum; 2017. Available from: https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-170921-samd-n41-clinical-evaluation_1.pdf
35. Park Y, Jackson GP, Foreman MA, Gruen D, Hu J, Das AK. Evaluating artificial intelligence in medicine: phases of clinical research. JAMIA Open 2020 Oct;3(3):326–331. PMID:33215066

Appendix

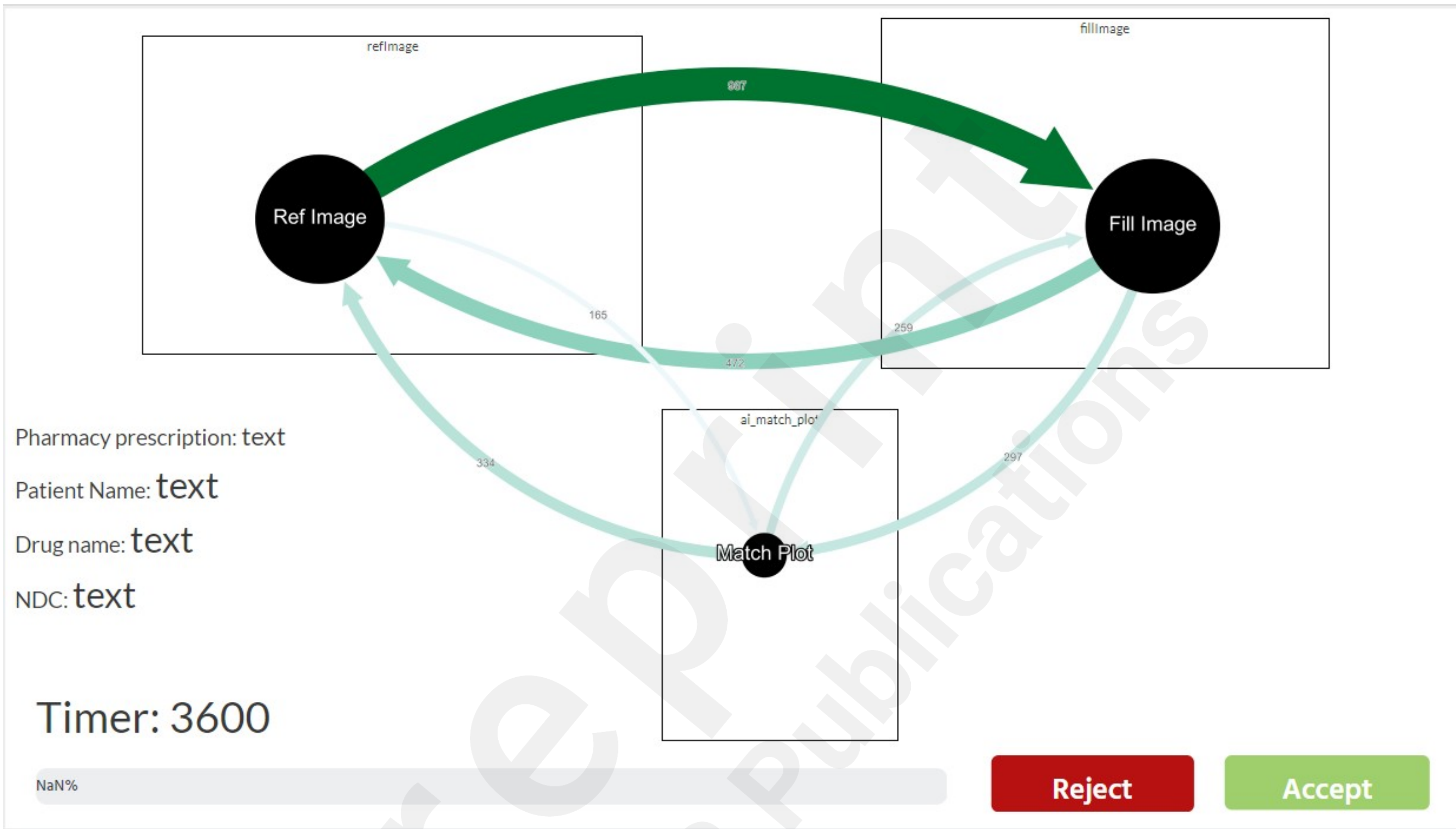
Average fixation duration

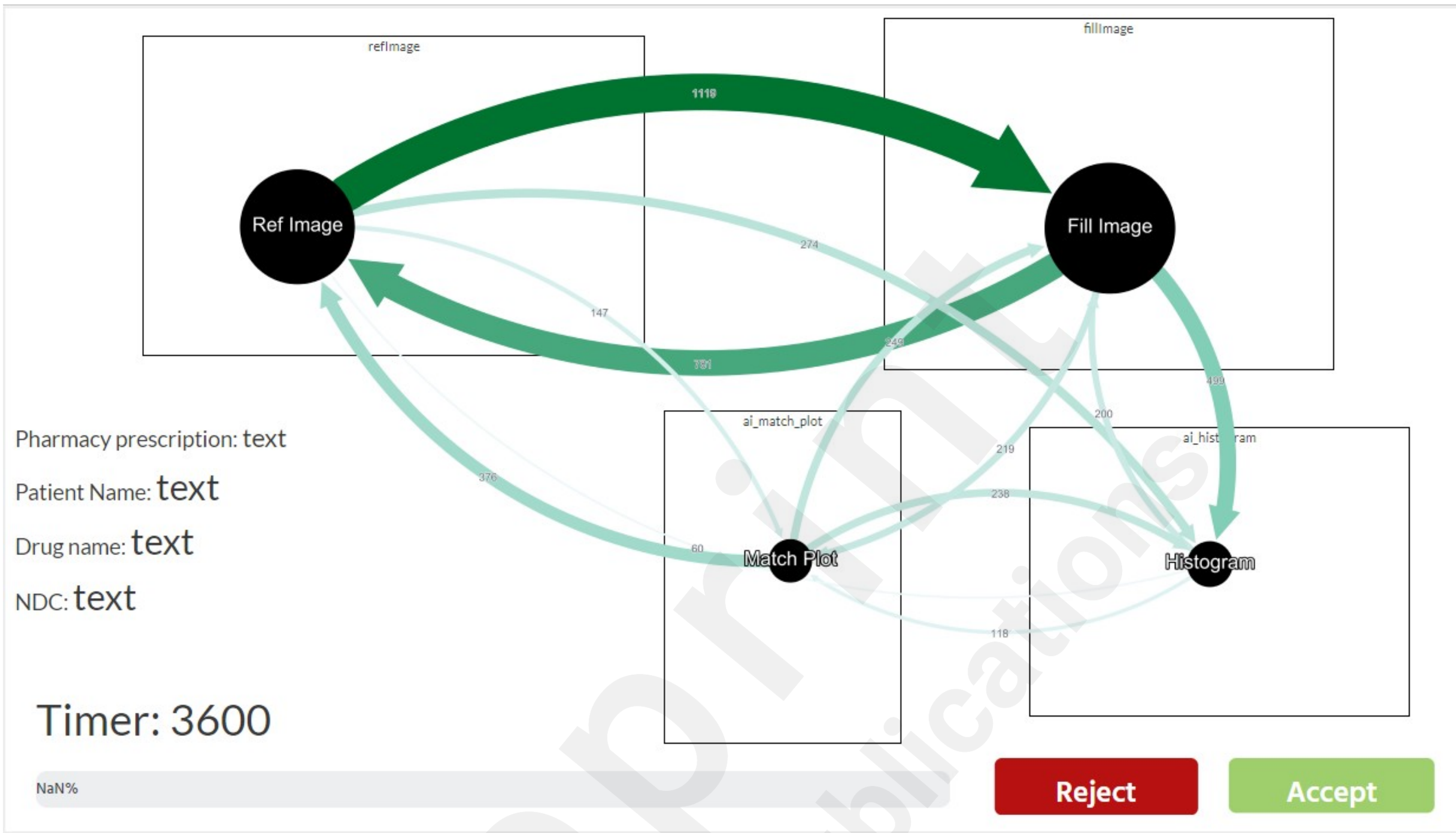


Average Fixation Duration						
Help Type	**Area**	**Mean**	**Standard Dev.**	**Min**	**Max**	**n**
No Help	**Fill Image**	0.22	0.054	0.09	0.53	9756
	Ref Image	0.20	0.057	0.09	0.47	5246
Simple	**Fill Image**	0.22	0.053	0.09	0.53	4547
	Ref Image	0.20	0.055	0.09	0.46	2344
	Match Plot	0.21	0.057	0.08	0.44	1583
Advanced	**Fill Image**	0.22	0.057	0.09	0.56	5729
	Ref Image	0.20	0.058	0.08	0.46	2729
	Match Plot	0.21	0.059	0.09	0.46	1382
	Histogram	0.22	0.056	0.08	0.4	1566

The average fixation duration in each region is similar, regardless of help type. On average, duration of each fixation is around 0.2 seconds, with the standard deviation of 0.055 seconds.

Gaze pattern & Link analysis

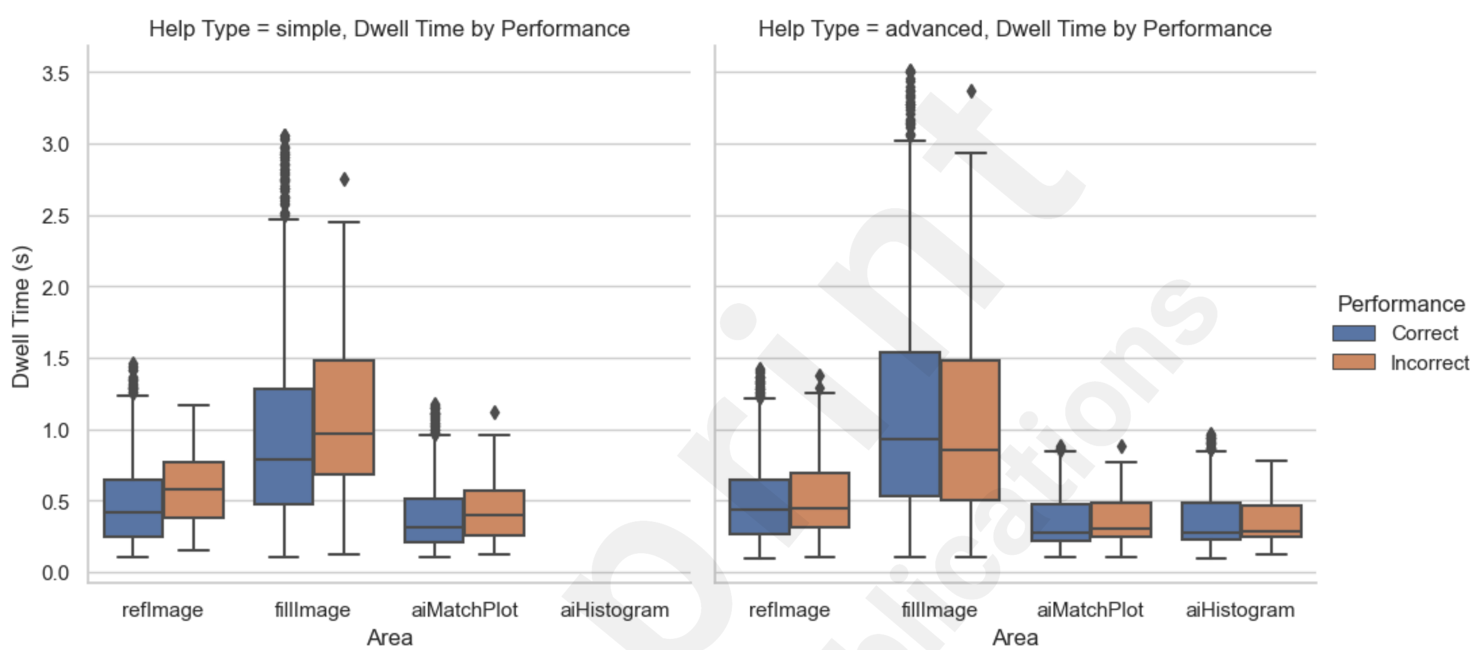




Dwell Time Difference by Performance

Dwell Time by Help type and Performance							
help_type	**Performance**	**area**	**mean**	**std**	**min**	**max**	**count**
Advanced	**Incorrect**	**aiHistogram**	0.35	0.15	0.12	0.78	50
		aiMatchPlot	0.36	0.17	0.1	0.88	54
		fillImage	1.10	0.77	0.1	3.37	76
		refImage	0.52	0.31	0.1	1.38	81
	Correct	**aiHistogram**	0.36	0.19	0.09	0.97	877
		aiMatchPlot	0.34	0.18	0.1	0.88	784
		fillImage	1.13	0.77	0.1	3.51	1039
		refImage	0.49	0.28	0.09	1.42	1031
Simple	**Incorrect**	**aiMatchPlot**	0.45	0.25	0.12	1.12	51
		fillImage	1.08	0.62	0.12	2.75	45

		refImage	0.58	0.26	0.15	1.17	52
	Correct	**aiMatchPlot**	0.40	0.25	0.1	1.18	776
		fillImage	0.96	0.66	0.1	3.06	1006
		refImage	0.47	0.29	0.1	1.46	933



For correct and incorrect trials, there is not a significant difference in dwell time of each region.

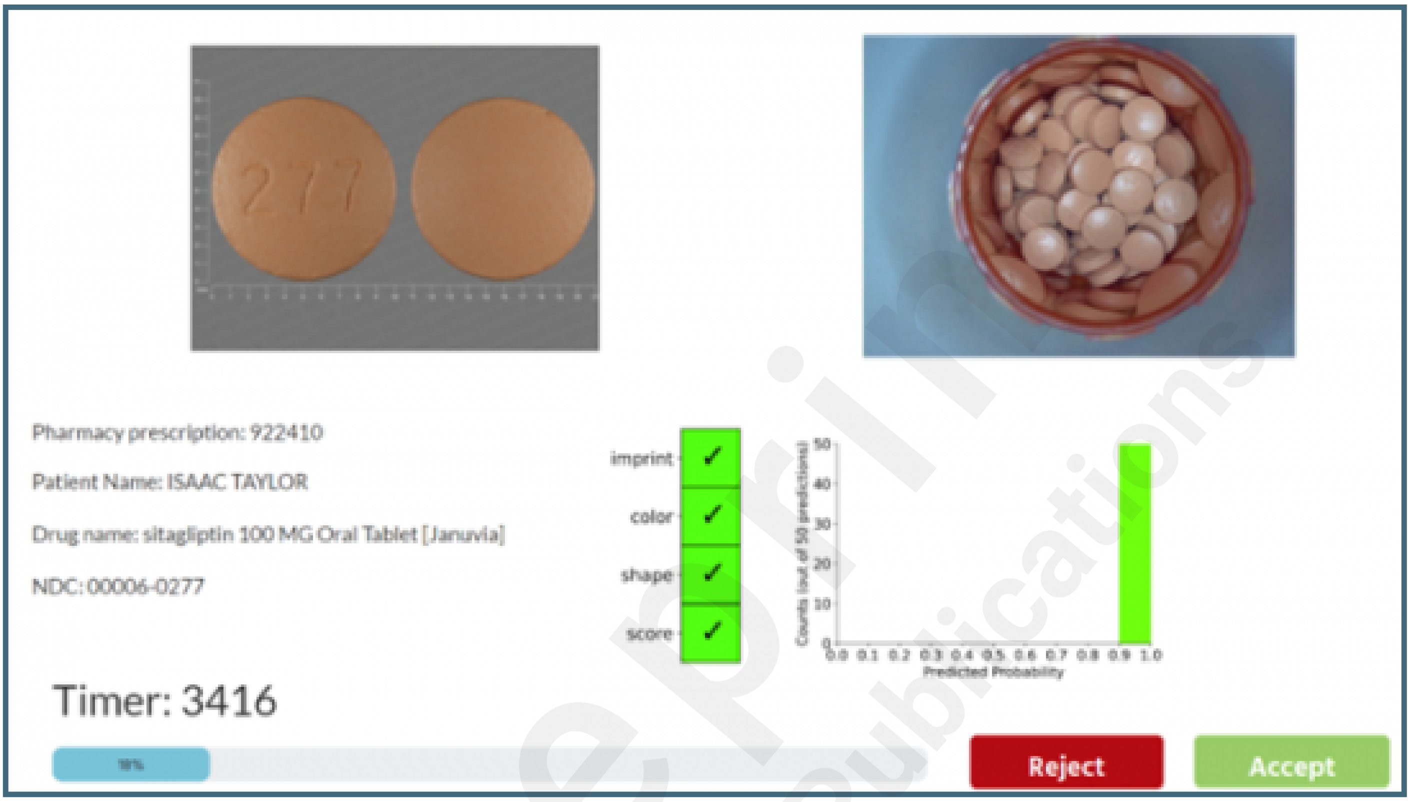
Dwell Time by Case Type							
help_type	**case_type**	**area**	**mean**	**std**	**min**	**max**	**count**
Advanced	**Helpful advice**	**aiHistogram**	0.36	0.19	0.09	0.97	757
		aiMatchPlot	0.34	0.18	0.1	0.88	685
		fillImage	1.05	0.71	0.1	3.51	922
		refImage	0.47	0.27	0.09	1.42	913
	Unhelpful advice	**aiHistogram**	0.37	0.19	0.1	0.94	170
		aiMatchPlot	0.37	0.18	0.1	0.81	153
		fillImage	1.52	0.93	0.1	3.5	193
		refImage	0.58	0.31	0.1	1.38	199
Simple	**Helpful advice**	**aiMatchPlot**	0.38	0.23	0.1	1.15	652
		fillImage	0.92	0.64	0.1	3.06	882
		refImage	0.46	0.28	0.1	1.46	806

	Unhelpful advice	**aiMatchPlot**	0.46	0.28	0.1	1.18	175
		fillImage	1.19	0.71	0.12	3.06	169
		refImage	0.55	0.31	0.1	1.35	179

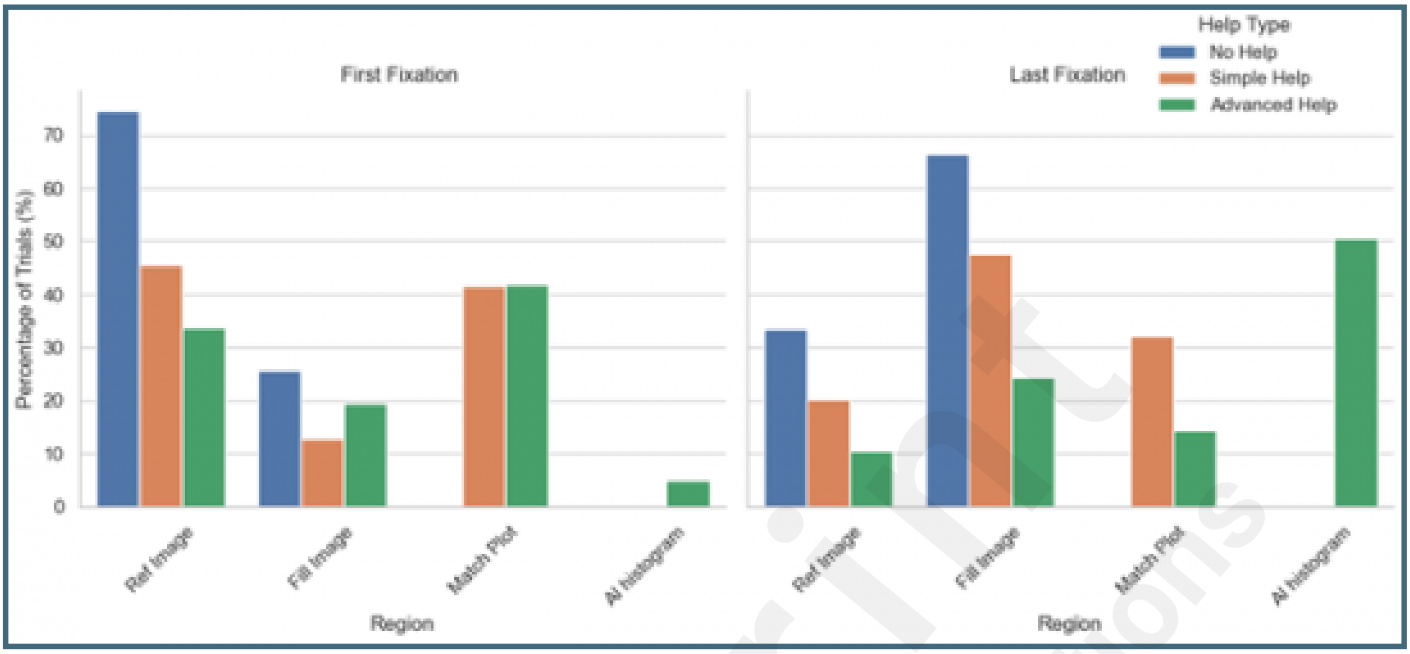
Supplementary Files

Figures

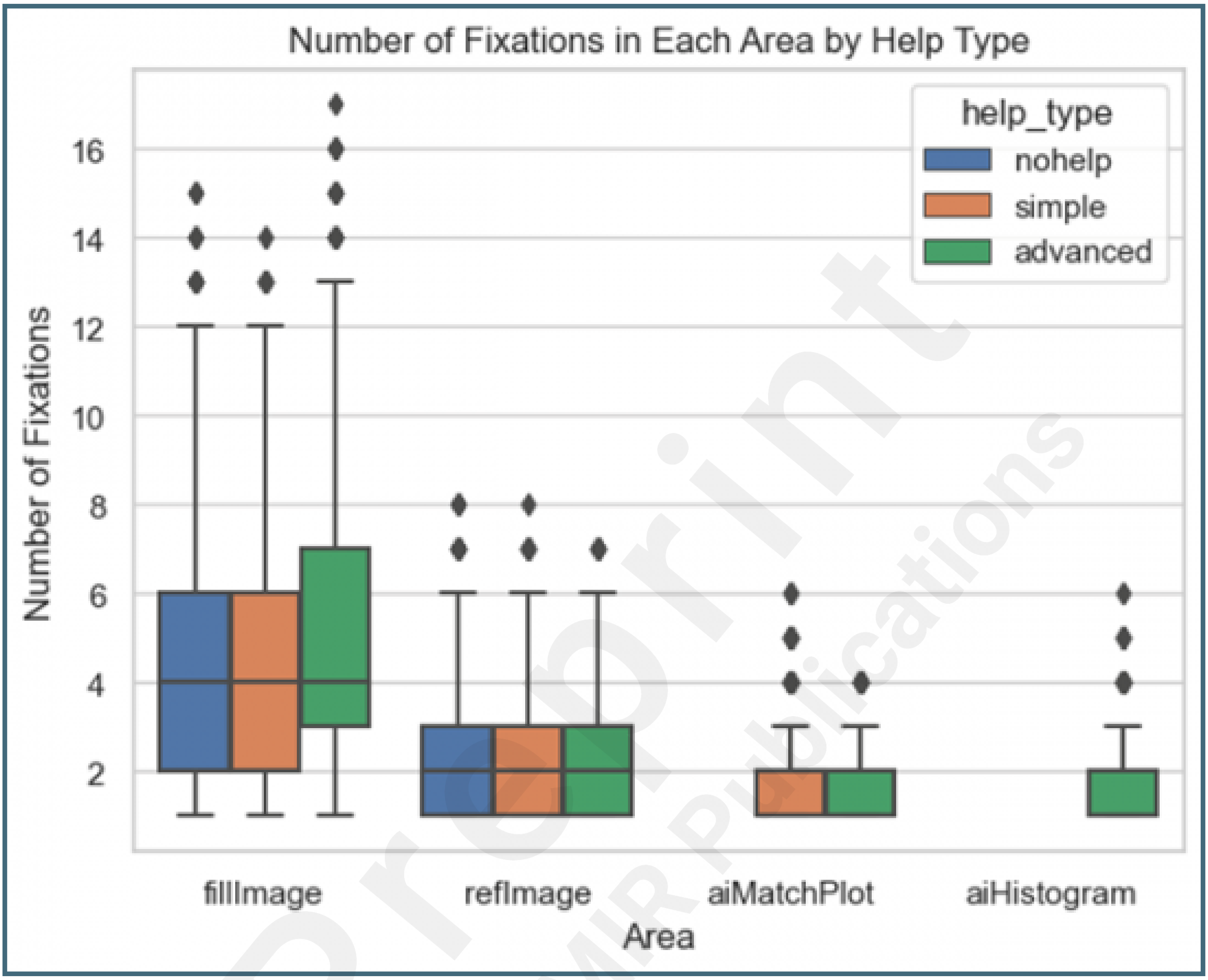
The medication verification system interface with AI Help. In trials without AI assistance, pharmacists only have access to the reference image (refImage) and fill image. With simple AI help, pharmacists saw an AI match plot, which shows AI's stance on the match status for four key medication characteristics (imprint, color, shape, and score). In advanced AI help, in addition to the match plot, pharmacists also saw the AI histogram, which displays the probability distribution from 50 predictions, indicating the uncertainty of the prediction.



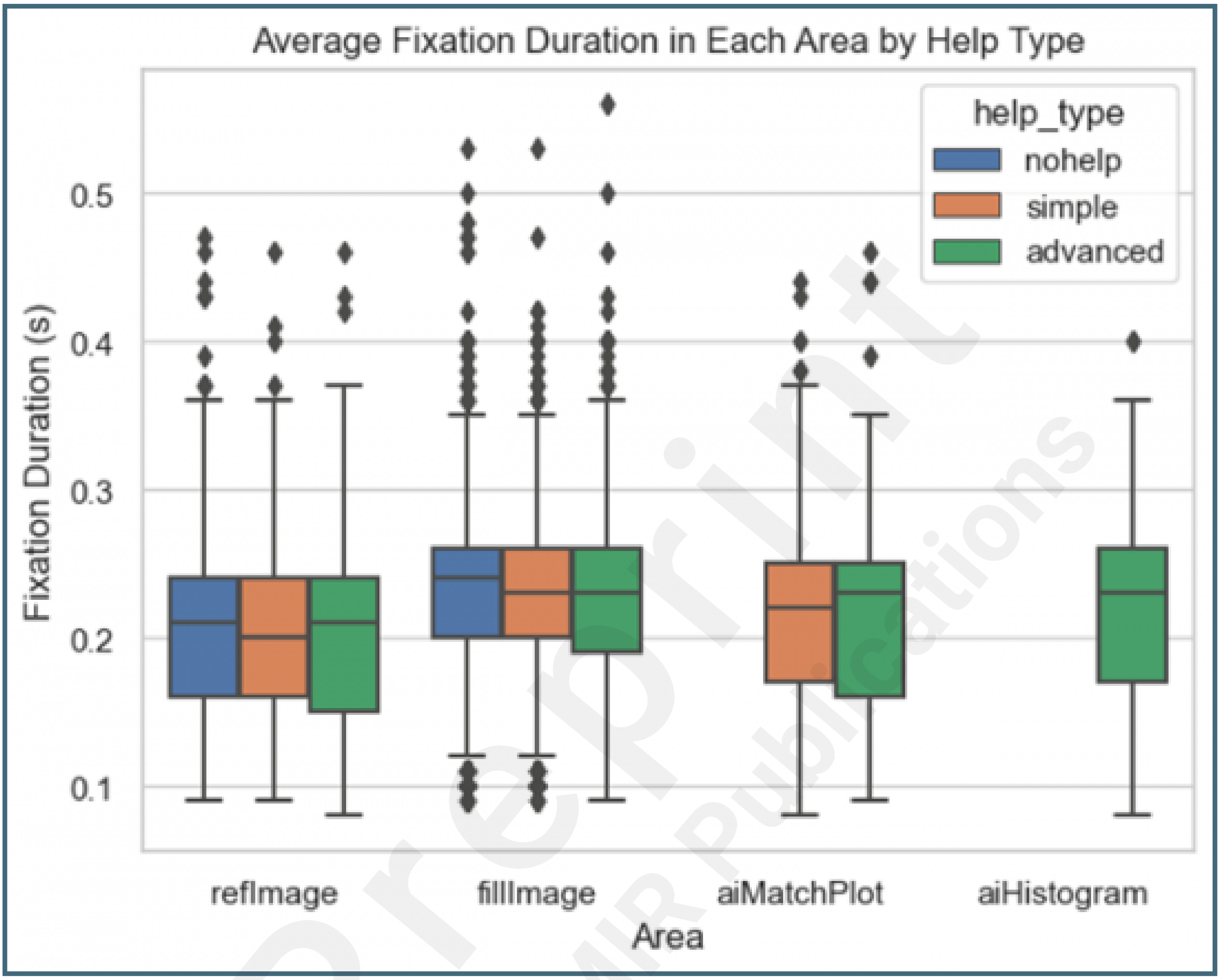
Percentage of trials by region of interest and AI help type. Left: first fixation, Right: last fixation.



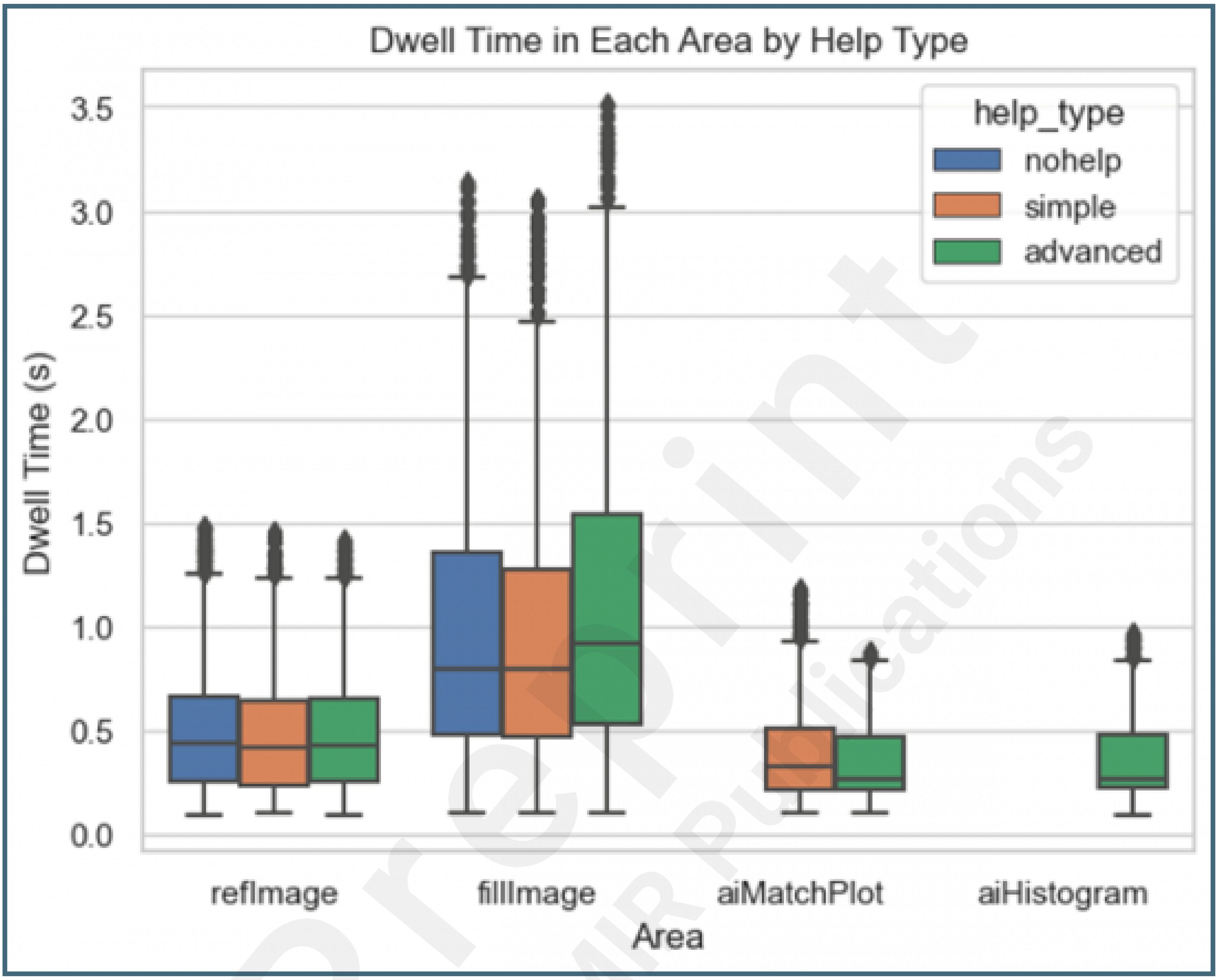
Boxplot representation of number of fixations in each area per trial, categorized by help type.



Boxplot representation of fixation duration in each area, categorized by help type.



Boxplot representation of dwell time in each area, categorized by help type.



Dwell time in each area, categorized by case type.