A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy

Emma Beede
Google Health
Palo Alto, CA
embeede@google.com

Elizabeth Baylor
Google Health
Palo Alto, CA
ebaylor@google.com

Fred Hersch
Google Health
Singapore
fredhersch@google.com

Anna Iurchenko
Google Health
Palo Alto, CA
annaiu@google.com

Lauren Wilcox
Google Health
Palo Alto, CA
lwilcox@google.com

Paisan Ruamviboonsuk
Rajavithi Hospital
Bangkok, Thailand
paisan.trs@gmail.com

Laura M. Vardoulakis
Google Health
Palo Alto, CA
lauravar@google.com

ABSTRACT

Deep learning algorithms promise to improve clinician workflows and patient outcomes. However, these gains have yet to be fully demonstrated in real world clinical settings. In this paper, we describe a human-centered study of a deep learning system used in clinics for the detection of diabetic eye disease. From interviews and observation across eleven clinics in Thailand, we characterize current eye-screening workflows, user expectations for an AI-assisted screening process, and post-deployment experiences. Our findings indicate that several socio-environmental factors impact model performance, nursing workflows, and the patient experience. We draw on these findings to reflect on the value of conducting human-centered evaluative research alongside prospective evaluations of model accuracy.

Author Keywords

human-centered AI, health, deep learning, diabetes

CCS Concepts

•**Human-centered computing** → *Empirical studies in HCI;*

CHI'20, April 25–30, 2020, Honolulu, HI, USA

ACM ISBN 978-1-4503-6708-0/20/04.
DOI: https://doi.org/10.1145/3313831.3376718

INTRODUCTION

Diabetes is a growing problem around the world, including in Southeast Asia [39]. As of 2016, 9.6% of Thailand's population was living with diabetes, comparable to 9.1% of the population in the United States [41, 40]. With diabetes comes complications, including diabetic retinopathy (DR), a condition caused by chronically high blood sugar that damages blood vessels in the retina, the thin layer at the back of the eye responsible for sensing light and sending signals to the brain. These blood vessels can leak or hemorrhage, causing vision distortion or loss. DR is one of the leading causes of vision impairment in the world [29], and causes 5% of cases of blindness worldwide, excluding refractive errors [38]. In Thailand, 34% of patients with diabetes have low vision or blindness in either eye [23].

In early stages of DR, a patient often has no symptoms, making it important for people living with diabetes to be screened regularly, as this is the stage in which damage can be reversed—progression of DR can be stopped or significantly reduced by blood sugar control. Early detection is key to initiate timely treatment and mitigate the risk of blindness.

Since 2013, the Ministry of Health in Thailand has set a goal to screen 60% of its diabetic population for diabetic retinopathy (DR). However, reaching this goal is a challenge due to a shortage of clinical specialists. In Thailand, there are 1500 ophthalmologists, including 200 retinal specialists, who provide ophthalmic care to approximately 4.5 million patients with diabetes [23]—a ratio of about 1:3000, about double of what it is in the United States [3, 11]. The shortage of doctors limits the ability to screen patients and also creates a treatment backlog for those found to have DR. As a result, nurses

conduct DR screenings when patients come in for diabetes check-ups, by taking photos of the retina and sending them to an ophthalmologist for review.

Our team has developed a deep learning algorithm that can provide an assessment of diabetic retinopathy, bypassing the need to wait weeks for an ophthalmologist to review the retinal images [20]. This algorithm has been shown to have specialist-level accuracy (>90% sensitivity and specificity) for the detection of referable cases of diabetic retinopathy. Through a large-scale, retrospective study comparing the algorithm to human graders, the deep learning algorithm shows significant reduction in the false negative rate (by 23%) at the cost of slightly higher false positive rates (2%) [33].

Currently, there are no requirements for AI systems to be evaluated through observational clinical studies, nor is it common practice. This is a problem because the success of a deep learning model does not rest solely on its accuracy, but also on its ability to improve patient care [35].

Prospective studies that involve evaluations of deep learning models within a clinical environment are beginning to emerge [1]. These studies are designed to provide additional evidence of model accuracy (sensitivity and specificity), but are not sufficient to evaluate true clinical effectiveness — that is, impact on patient care [25], nor do they explore socio-environmental factors that impact model performance in the wild. Furthermore, as Yang and colleagues note, when HCI researchers attempt to study AI systems in a hospital or clinic, they are often prevented from fully embedding into clinical workflows and from evaluating systems using authentic patient data [42].

This paper contributes the first human-centered observational study of a deep learning system deployed directly in clinical care with patients. Through field observations and interviews at eleven clinics across Thailand, we explored the expectations and realities that nurses encounter in bringing a deep learning model into their clinical practices. First, we outline typical eye-screening workflows and challenges that nurses experience when screening hundreds of patients. Then, we explore the expectations nurses have for an AI-assisted eye screening process. Next, we present a human-centered, observational study of the deep learning system used in clinical care, examining nurses' experiences with the system, and the socio-environmental factors that impacted system performance. Finally, we conclude with a discussion around applications of HCI methods to the evaluation of deep learning algorithms in clinical environments.

RELATED WORK

Design research in clinical settings

Researchers have shown that paper prototyping and wizard-of-oz methods can be used with real patient data to study contextual fit of prototypes in clinical settings [31, 36]. Additional human-centered evaluative research has occurred with more robust systems deployed within a hospital, and that operate on patient data during the deployment period for the purposes of conducting small-scale pilot studies [6, 37, 22]. However, the artifacts and systems used in these studies did not incorporate deep learning models, which add challenges to traditional formative design approaches.

Challenges evaluating AI in clinical settings

Past work on systems using artificial intelligence, such as computer-aided detection (CAD) [10] or Decision Support Tools (DSTs) [4] has highlighted several obstacles in going from research and development environments to hospital or clinical settings. These obstacles include frequent lack of utility to clinicians and logistical hurdles that slow or block deployment [15, 27]. Even systems with widespread adoption, such as CAD in Mammography, have been shown to require a radiologist to do more work, not less, [18] and generally do not improve a radiologist's diagnostic accuracy [12, 26].

Human-centered evaluation of interactive, deep learning systems—within clinical environments—is an open area of research [30]. Cai and colleagues created interactive techniques that can lead to increased diagnostic utility and increased user trust in the predictions of a deep learning system, used by pathologists in a lab setting [8]. In subsequent work, Cai et al. examined what information pathologists (in a lab setting) found to be important when being introduced to AI assistants, before integrating these assistants into routine prostate cancer screening practice [9]. Clinicians in their study emphasized a need to gain an initial "global" impression of a model (e.g. its limitations, medical point-of-view, idiosyncrasies, and overall objective). While these studies bring us closer to understanding the needs of clinicians as they interact with deep-learning-based systems, they do not account for the highly situated nature of activities in clinical environments [16].

Observational studies of the introduction of diagnostic and information systems

A large body of ethnographic and ethnomethodologically-informed studies have focused on the effects of workflow when new diagnostic and information systems are introduced into clinical environments. In the medical imaging domain, researchers have long known that clinical trials fail to grasp the situated social and collaborative dimensions of medical work [21]. Alberdi et al.'s ethnographic study of the introduction of a CAD tool in breast screening found a range of effects on experts' decisions, both beneficial and detrimental [2]. Yang, et al. conducted a design and field assessment of a DST for cardiologists, and found that the more an AI system is unobtrusively embedded into current workflows, the more likely a clinician will embrace such a system [42]. Unfortunately, Yang and colleagues faced challenges evaluating their predictive model using authentic patient data, and were only able to evaluate their tool using one synthetic patient case.

Jirotka et al. studied work practices in digital mammography screening, finding that system design affected both trust in the screening system and humans' trust in each other, when their work was mediated by the system [24]. Their study found that clinicians developed their own tolerance and workarounds for system policies, in order to trust the system results. Their study highlighted the importance of understanding differences in data acquisition and analysis practices that tended to emerge in local contexts (e.g. different characteristic appearances

of images if taken at one site or another)—differences that clinicians naturally acclimated to and incorporated into their human decision-making processes [24].

In a recent study on computerized medication order entry, DSTs improved pharmacists' workflow but increased communication load and impaired aspects of human decision-making [32]. Reddy and colleagues examined the effects that a new wireless notification system had on surgical ICU clinicians, finding a number of disruptions to existing work practices and information flows. In their study, the new system prevented residents and fellows from managing critical events before they were relayed to the attending physician, disrupting their ability to manage problems that did not require escalation [34].

Our research builds upon the long tradition of studies that examine environmental and contextual factors surrounding systems designed for a clinical environment. To our knowledge, this paper presents the first human-centered evaluation of a production-level deep learning model for diagnostic prediction, being used across several clinics for patient care.

DR SCREENING IN THAILAND

In 2013, the Thailand Ministry of Public Health set up a national screening program where patients can be screened for diabetic retinopathy when visiting their local clinic for a diabetes check-up on designated eye screening days (available year round in some clinics, or during a period of two months for clinics that share the equipment). The Ministry of Public Health set an initial target of screening 60% of diabetic patients in each region. However, even with ample opportunity to receive an eye screen, Ministry data indicates that less than 50% of the diabetic patients are screened annually since the inception of the program.

During the eye screening portion of a diabetes check-up, nurses are tasked with taking photos of a patient's retina, called a fundus photo (Figure 1). These images are sent, either by email or on a compact disc in postal mail, to an ophthalmologist and are often reviewed weeks to months after a patient's fundus photo was taken. However, nurses often perform initial grading themselves, checking for apparent abnormalities that clearly warrant a referral.

When received, the ophthalmologist evaluates the images for the presence and severity of DR. DR generally has 4 levels of severity: Mild, Moderate, Severe Non-Proliferative, and Proliferative [28]. In addition, vision can be threatened by the development of Diabetic Macular Edema (DME), an accumulation of fluid or lipid in the macula due to leaking blood vessels that can occur at any stage of diabetic retinopathy. Depending on the severity of DR, presence of DME, and the visual acuity of the patient when screened, patients may be referred to an ophthalmologist or told to come back for more frequent screening exams. After results are available - up to 10 weeks later - patients with no DR are advised to come back for screening at approximately one year; patients with Mild DR are advised to come back at approximately six months, and patients with moderate, severe, or proliferative DR, as well as DME, get referred to an ophthalmologist for a comprehensive exam and possible treatment.

Figure 1. A nurse operates the fundus camera, taking images of a patient's retina.

AI-ASSISTED EYE SCREENING

Using a system employing a deep learning algorithm to assess DR could reduce the requirement for nurses to perform grading in the moment, and eliminate the need for nurses to send fundus images to an ophthalmologist for review (Figure 2). Removing this bottleneck could provide patients with immediate results, give nurses the ability to make a referral recommendation in the moment, facilitate faster treatment of patients, and allow more cases to be screened and reviewed. Furthermore, it could allow nurses to spend additional time with patients on diabetes management and education, which is what ultimately leads to improved patient outcomes [14].

Context of use

In partnership with Rajavithi Hospital under the Ministry of Public Health, the deep learning system is being deployed in multiple clinics throughout Thailand, in a large-scale prospective study with 7600 patients. The purpose of the prospective study is to evaluate the feasibility and performance of the algorithm in a real-world clinical setting. The research presented in this paper was conducted in parallel with this prospective study, across two regions in Thailand: Pathum Thani and Chang Mai, in order to evaluate the socio-environmental factors that influence the algorithm's success, and how use of the system affects nursing workflows and the patient experience.

As part of the prospective study, the deep learning system was initially deployed in three clinics within the province of Pathum Thani: Klong Luang, Nongsue, and Lamlukka. The three initial sites screened patients from December 2018 to May 2019, at which point their eye screening season for the year concluded, to be resumed in October 2019. Each clinic offered screening between two and eight days a month, and on each screening day, would see between 30–200 patients.

Prospective Study Protocol

On a designated screening day, patients due for an eye exam were identified as they registered at the clinic and provided information about the study. Prior to their exam, all patients were called by a nurse or nursing assistant and assessed for study eligibility, based on age and comorbidities. If eligible,

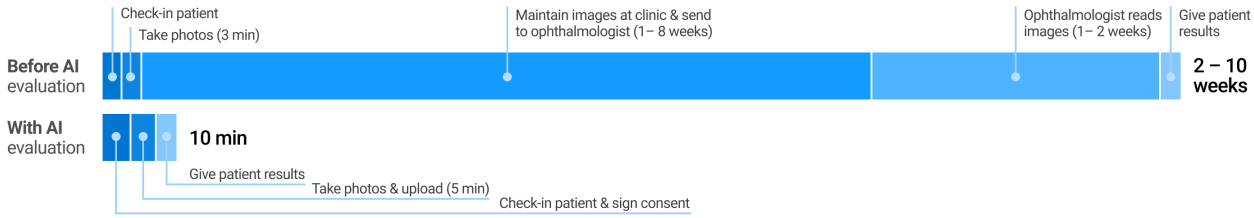

Figure 2. Eye screening process before and after deployment of the deep learning system.

the nurse invited them to participate in the study, explained how their data would be used, reviewed the written consent form, and answered any questions. If the patient consented to participate, a pre-generated case identifier was assigned. The patient's name, date of birth, and case identifier were entered into a paper register maintained by the local clinic staff, should any future follow-up be needed. Patients that did not sign the consent continued with their exam as per the typical local process, having their images sent to an ophthalmologist to be graded without use of the deep learning system.

Once a patient was called for their eye exam, the camera operator verified consent, took photos of each eye using the clinic's current fundus camera, and uploaded them to the deep learning system, using the case number as the sole identifier. The images were sent to the algorithm in the cloud, and an assessment of the presence and severity of DR, as well as the presence or absence of Diabetic Macular Edema (DME) was returned in real-time, including a recommendation for whether or not the patient should be referred to an ophthalmologist (Figure 3).

If the system recommended referral (due to either DR severity or an ungradable image), the nurse (often the camera operator) gave the patient two referral notes. The first was the standard referral letter given by the doctor and required for the appointment at the referral center. The second was provided to study participants to allow the referral centers to easily track referral adherence. The patient was to take both notes to the specialist.

Each week, images and corresponding predictions were reviewed by a member of the study staff, an ophthalmologist, referred to as an *overreader*, to ensure the system had not missed any patients that needed to be referred. If the overreader determined that a patient missed by the system should be seen by a specialist, a nurse followed up with the patient and issued a referral. If the system referred the patient but it should not have, no action was taken.

At the completion of the prospective study, all collected images will go through an adjudication panel to determine the deep learning system's accuracy.

HUMAN-CENTERED CLINICAL FIELD STUDY

Our work complements the prospective study, by adding an additional focus on the human aspects to deploying and using a deep learning system within a clinical environment. Through a multi-round study, both before and after deploying the deep learning system, we explore nurses' expectations of the system, clarity of the patient consent process, experiences of the operation of the technology, and shifts in workflow to accommodate the new system. This human-centered evaluation is the primary focus of this paper.

Figure 3. Web application displaying the deep learning model's predictions for diabetic retinopathy and diabetic macular edema, along with the fundus photos.

We conducted fieldwork at 11 clinics within the provinces of Pathum Thani and Chiang Mai during November 2018, April 2019, and August 2019. We timed our visits to observe and understand the eye screening process across multiple sites (five clinics within Pathum Thani and six within Chiang Mai) prior to the deployment of the deep learning system. After the system was deployed at three clinics in Pathum Thani, and used for patient screenings for at least one month, we returned to the deployment sites and conducted observations and interviews with all primary users of the system. We also held weekly feedback phone calls with the local study coordinator, who was in close contact with the nurse users on a day-to-day basis.

This study was approved by the Ethical Review Committee for Research in Human Subjects of Rajavithi Hospital and the Ethical Committees of hospitals or health centers from which retinal images of patients with diabetes were used. Patients gave informed consent allowing their retinal images to be used in the prospective study evaluating the accuracy of the deep learning system: registered in the Thai Clinical Trials Reg-

istry, Registration Number TCTR20190902002. Clinicians gave informed consent for participating in the interviews and observations reported in this paper.

Data Collection

Pre-Deployment Fieldwork

Before any system deployments, we conducted 26 hours of observation (adopting a non-member role [13]) across 11 clinics. We observed the eye-screening process of more than 100 patients, along with general clinical activity. Additionally, we conducted 15 hours of interviews with the 13 nurses who lead eye screening at eight clinics.

Post-Deployment Fieldwork

After deployment of the deep learning system at three clinics in Pathum Thani, we spent 12 hours observing the eye screening process of about 50 patients, and conducted seven hours of interviews among five nurses and one camera technician (all primary users of the system).

In all interviews (pre- and post-deployment), one researcher led the discussion with the aid of a Thai interpreter who translated the researcher's questions and the participant's responses. An additional researcher acted as a note-taker during the interviews and wrote down the interpreter's translation of the participant's responses. Interviews were approximately 90 minutes long and took place in the nurse's office.

Data Analysis

Our analysis, for both pre- and post-deployment studies, included observational notes, interview recordings, and system logs, including model predictions and usage data from 1838 image uploads. To analyze our observations and interviews, we created affinity notes at the end of each day, which included 1-2 clinic visits. At the conclusion of our clinic visits for a given trip (each lasting about a week), the site researchers gathered and performed an affinity mapping of the interview and observation data [5]. Themes that emerged included: gradability, referrals, design implications, satisfaction, consent considerations, eye-screening workflows, and patient education.

Upon returning from each trip, recordings of participants' responses (interpreter quotes) were machine-transcribed and used as a backup to written verbatim notes.

Participants

Several roles make up the clinician team in charge of eye screening at the clinic. During our visits, we focused on primary users of the system: non-communicable disease (NCD) nurses, who lead patient care for general diabetes cases and may also perform eye screening; ophthalmic nurses, who have completed a four-month specialty course to learn how to capture and assess fundus photos and discuss eye diseases with patients, including DR; and camera technicians, who manage the process of taking fundus photos before passing them to a nurse for initial assessment and patient consultation.

All nurses were female, and ranged in nursing experience from 2 to 30 years (*mean*=17). Eight of the nurses had completed a four-month course in ophthalmic nursing, which they considered to be their specialty, with an average of five years experience in the specialty. The remaining nurses considered themselves non-communicable disease nurses and had taken abbreviated 2–10 day courses in screening patients for DR. The camera technician was also female and had a tenure of three months taking fundus photos (Table 1).

Province	Clinic	Role	Nursing (years)	Screening (years)
PT	1	Oph nurse	20	2
PT	1	Oph nurse	0.25	0.25
PT	**2**	**NCD nurse**	**26**	**5–6**
PT	**2**	**NCD nurse**	**10**	**0.25**
PT	3	Oph nurse	19	1
PT	**4**	**Oph nurse**	**2**	**2**
PT	**4**	**Oph nurse**	**2.5**	**2**
PT	**5**	**NCD nurse**	**19**	**6**
PT	**5**	**Technician**	-	**0.25**
CM	6	Oph nurse	23	15
CM	7	Oph nurse	35	10
CM	8	NCD nurse	28	7
CM	8	Oph nurse	22	2
CM	9	-	-	-
CM	10	-	-	-
CM	11	-	-	-

Table 1. Clinic and Interview participant details. Provinces: Pathum Thani (PT) and Chiang Mai (CM). Rows in bold represent post-deployment participation. Clinics 9-11 involved pre-deployment observations only. Clinics 1-4 were visited, pre-deployment, in November 2018. In April 2019, clinics 2, 4, and 5 were visited post-deployment, and clinics 9-11 were visited pre-deployment. In August 2019, clinics 6-8 were visited pre-deployment. Participant #'s are not included to protect their anonymity.

FINDINGS

We begin by providing an overview of the eye screening workflows and processes that we observed, prior to deployment of the deep learning system. We describe the different workflows and processes for conducting eye screenings among clinical sites, and how nurses often needed to make initial DR referral determinations themselves, given the high-volume of patients being screened each day. We review nurses' expectations of an AI-tool, and their desire to have it improve their ability to read images.

Next, we describe results from our post-deployment research, evaluating the deep-learning system from a human-centered perspective. We found that several contextual factors affected model performance in the clinical setting, which in turn affected nurses' attitudes towards the system and their willingness to consent patients into the prospective study of the system.

Pre-deployment Findings

As a typical eye-screening workflow, diabetic patients fasted overnight and presented to the clinic in the morning for blood tests. Next, they went to a general waiting area to measure their blood pressure and weight (often self-measured with machines in the waiting area). If the patient was due for an eye screening, they waited (anywhere from five minutes to three hours, depending on their queue number) to have their fundus photo taken. The nurse would take an image of the patient's right and left eyes, sometimes re-taking an image if it was determined to be too dark—typically caused by a source of light in the room or from an age-related cataract, making the pupil too small or presenting an ocular media too cloudy for the camera to image the retina. This process was repeated for all patients waiting in the queue, and occured each screening day at the clinic, usually once or twice a week.

Observational findings

We observed a high degree of variation in the eye-screening process across the 11 clinics in our study. The processes of capturing and grading images were consistent across clinics, but nurses had a large degree of autonomy on how they organized the screening workflow, and different resources were available at each clinic. For instance, one of the clinics in Chiang Mai was part of a larger hospital and had an ophthalmologist on staff, while the others relied exclusively on on-site nurses to read the image. Some clinics in Chiang Mai had the ability to use dilation drops at their discretion, but none in Pathum Thani did. At a clinic in Pathum Thani, eye screening was organized in an assembly-line fashion, with a camera technician taking the fundus photos and a nurse leading the consultation with the patient. In another Chiang Mai clinic, the nurse utilized part of the screening day for patient education, gathering all the patients together at the beginning of the day, to watch a YouTube video on a weight loss success story. She also used her time with patients to coach them, and would pull patients with high blood sugar aside to "interview" them. P11 told us, *"I ask patients to reflect on their lifestyle, it's better than me telling them what to do, they don't listen to that."*

Clinic Screening Conditions

The setting and locations where eye screenings took place were also highly varied across clinics. Only two clinics had a dedicated screening room that could be darkened to ensure patients' pupils were large enough to take a high-quality fundus photo. In other clinics, eye screening took place in the nurses' offices, or where additional patients received a foot sensitivity screening or nutritional counseling. As a result, the lights were rarely turned off in these settings while capturing a fundus photo, even when a fluorescent light was situated directly above the camera. We were interested to see how these real-world conditions would affect our model performance.

Volume

Through our observations and interviews, we saw, first-hand, the challenges required to meet the national screening goal target. Some clinics attempted to screen two hundred patients within five hours, allotting only ninety seconds to screen each patient, which often resulted in extended clinic hours or patients being rescheduled for another day (Figure 4).

Figure 4. Busy screening site at a clinic in Pathum Thani.

> [The screening process], it's okay, but that's if there aren't 300 people putting pressure on you. With 300, it's not okay, there's not enough time in that crowded hospital setting." -P2

> Patients come in, around 100 at a time, and it's not just eye screening that needs to happen...There are many things to do. After the medical history, the patient goes to the waiting area, and if there are a lot of people I have to pitch in and help out everywhere in the clinic. That's a lot of work... Things would be better if I could just be focused." -P7

Referral Determinations

All images were initially assessed by a nurse then sent to an ophthalmologist for review. The ability to assess fundus photos for DR varied from nurse to nurse. While most nurses told us they felt comfortable assessing for the presence of DR, they didn't know how to determine the severity if present. P4 told us, *"I know if it's not normal, but I don't know what to call it."* To make the ultimate decision of whether a patient needs to be referred to an ophthalmologist for an exam and potentially for treatment, the nurse turned to the ophthalmologist or retinal specialist, who are most often remote.

Images that appeared normal to the nurses were typically sent via email to the ophthalmologist in batches that include several weeks worth of images. The ophthalmologist then determined whether or not the patient needs to be referred for an exam, and typically sent the results back to the nurses within 1–2 weeks. For images that seemed abnormal, some nurses sent the image to the ophthalmologist via instant messaging in hopes of getting a recommendation for the patient quickly. These recommendations usually took days, but were sometimes returned within hours.

Expectations for AI-assisted screening

During our pre-deployment interviews with nurses, we sought feedback on potential designs for the visual presentation of the DR and DME prediction. We found that fundus images were a deep part of the nurses' practice, and it became clear that the images need to be *prominently* displayed alongside the DR prediction. Displaying the fundus images would not only provide confidence to the nurse that the correct image

was being used for the assessment, it also provided nurses with information they could use to convince patients to seek treatment. If a patient needed urgent treatment, but wasn't experiencing any symptoms, nurses wanted to be able to show the fundus images directly to the patient, and point out the area of concern. They were excited about a combination of images *with* the prediction as a way to aid in those conversations.

Potential benefits

Nurses foresaw two potential benefits of having an AI-assisted screening process. The first was using the system as a learning opportunity—improving their ability to make accurate DR assessments themselves. In their typical practice, nurses enjoyed fundus photo reading as an educational experience and an opportunity to apply their training.

> I like to study and learn things—and this is where you learn, the way you become more knowledgeable. -P4

> In the past, I got some things wrong. I made referrals when I shouldn't have. The doctor told me it was something other than DR, and I learned from this. Now my readings are better. -P6

> I went to training to learn how to read images. I'm very interested in it. I asked for images from the ophthalmologist—he sent them to me, I read them, and then I asked him to check my work. I got 850/1000 right... right now no one can replace me. - P11

The second benefit was the potential to use the deep learning system's results to *prove* their own readings to on-site doctors. Several nurses expressed frustration with their assessments being undervalued or dismissed by physicians, and they were excited about the potential to demonstrate their own expertise to more senior clinicians. As P7 explained, *"They don't believe us."* P11 stated, *"It could confirm what we already know."*

Potential challenges

Nurses also anticipated challenges using the deep learning system within clinical care. With patient volume already a burden, nurses were concerned that following the study protocol (including uploading images) would add to their workload and ability to screen all patients arriving each day. Nurses were also concerned about the consequences for patients if the algorithm produced a false positive, including the additional travel burden to follow up on a referral, the cost of missing work associated with travel, and the emotional strain a positive result could place on them.

Post-deployment findings

With the addition of the deep learning system, patients generally followed the same journey through the clinic as normal, with the exception of now being able to receive an immediate determination of whether or not a referral is needed (Figure 2). In this section, we present findings that we were only able to uncover as a result of studying the deep learning system in the context of actual clinical care. We discuss several challenges embedded in the intended journey of AI-assisted eye screening: consenting patients, factors influencing system performance and the clinician/patient experience, and resulting system workarounds.

Consenting patients

Given that the deep learning system was deployed in an observational, prospective study, it was critical for nurses to obtain patient consent prior to using the system. The informed consent process was the first challenge we observed, and was made more complicated by the need to explain the deep learning system.

To participate, patients provided both written and verbal consent before having their eyes photographed. As the system provides an immediate referral recommendation, nurses knew that they may need to convince a patient to visit a specialist, depending upon the results. This was a large change from their previous workflow, where results may not be available for up to 10 weeks, long after a patient has left the clinic. As a result of the prospective study protocol design, and potentially needing to make on-the-spot plans to visit the referral hospital, we observed nurses at clinics 4 and 5 dissuading patients from participating in the prospective study, for fear that it would cause unnecessary hardship.

At one clinic, we observed nurses mentioning to patients that if they received a positive DR result they will be referred to Pathum Thani Hospital (an hour drive away). Not all patients have cars, and transportation for them is uncertain. While patients at all sites were given the same information via written consent forms, some nurses felt the need to "warn" patients that they would need to travel should a referral be given. Given the far distance and inconvenience of getting to Pathum Thani Hospital, 50% of patients at clinic 4 opted out of participating in the study, even though it was unlikely that they would be referred.

> [Patients] are not concerned with accuracy, but how the experience will be—will it waste my time if I have to go to the hospital? I assure them they don't have to go to the hospital. They ask, 'does it take more time?', 'Do I go somewhere else?' Some people aren't ready to go so won't join the research. 40-50% don't join because they think they have to go to the hospital. - P6

Through observation and interviews, we found a tension between the ability to know the results immediately and risk the need to travel, versus receiving a delayed referral notification and risk not receiving prompt treatment. Patients had to consider their means and desire to be potentially referred to a far-away hospital. Nurses had to consider their willingness to follow the study protocol, their trust in the deep learning system's results, and whether or not they felt the system's referral recommendations would unnecessarily burden the patient.

Clinical Factors Influencing System Performance

We suspected, even using a high-quality dataset to train the deep learning model, that environmental and contextual factors would impact the system's performance in a clinical setting. Through our field research, we indeed found this to be true and observed several factors that affected the system's performance, as well as the clinician and patient experience.

Gradability

For an image to be gradable by a human or an algorithm, it needs to capture the retinal field clearly. To achieve a clear

image, enough light from the camera needs to get through to the back of the eye. For that light to get through, a patient's pupil needs to be large, helped by either a dark environment, dilation drops, or both.

The deep learning system has stringent guidelines regarding the images it will assess. For patient safety reasons, it was designed to decrease the chance that it would make an incorrect assessment, and therefore only assesses the highest-quality images. If an image has a bit of blur or a dark area, for instance, the system will reject it, even if it could make a strong prediction. Because the deep learning system cannot guarantee that it hasn't missed something, these images are deemed ungradable.

After clinics 2, 4, and 5 all reported issues with gradability, we reviewed system logs to determine how many images were rejected by the algorithm. Out of 1838 images that were put through the system (in the first six months of usage), 393 (21%) didn't meet the system's high standards for grading. Through our observations and interviews, we found that low-quality images were caused by fundus photos being taken in a non-darkened environment, as observed in our pre-deployment findings, or from a camera that needed repair. In addition, clinics in Patham Thani were not using dilation drops on patients, which could have aided in capturing a quality image.

In the case of an ungradable image, the system notifies the nurse and recommends the patient be referred to an ophthalmologist, as part of the prospective study protocol. This immediate gradability feedback is something that the nurses did not have before, and turned out to be frustrating as images they felt were human-readable were rejected by the system. Because of this, nurses somewhat questioned the power of the deep learning system. P6 said, *"It gives guaranteed results, but it has some limitations. Some images are blurry, and I can still read it, but [the system] can't."* P3 shared the same sentiment, *"It's good but I think it's not as accurate. If [the eye] is a little obscured, it can't grade it."* The system's high standards for image quality is at odds with the consistency and quality of images that the nurses were routinely capturing under the constraints of the clinic, and this mismatch caused frustration and added work.

This in-the-moment feedback caused the nurses to take more photos, in an attempt to achieve an image the system will grade. In doing this, they noticed that they can create a semi-complete image from two ungradable photos. If in one photo she was able to see the top half of the image field, the nurse would take a second photo where she could capture the bottom half from the same eye (Figure 5). She would try to assess for DR using the top half of one image and the bottom half of another. They expected the deep learning system to perform this workaround as well; however, it couldn't, because it requires one high-quality image per eye, and cannot make assessments based on a composite from two images. *"I want to be able to upload a second image with the focus on the macula. I want to focus on one area and crop it so I don't keep getting an ungradable result,"* said P7.

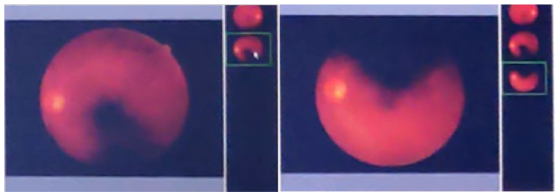

Figure 5. A nurse attempts to form a composite image of one eye by taking two images of the same eye, with varied lighting.

With poor lighting conditions commonly causing low-quality images, and many patients waiting in the queue, these ungradable images frustrated both patients and nurses. We observed nurses spending 2–4 minutes with each patient, retrying taking the photos a second time if the first ones were unable to be assessed, but never retrying for a third due to the discomfort the camera's bright flash causes to patients. P6 expressed this concern, *"I'll do two tries. The patients can't take more than that."* In the event that a nurse cannot capture a gradable image, the patient is to be referred to a specialist (potentially unnecessarily).

Once we understood the rate of images that could not be assessed by the deep learning system, we explored solutions to improve lighting conditions that will lead to higher-quality images, such as darkening the room as much as possible, using a cloth as a hood or veil to limit light further, and waiting 60 seconds between photographing the right and left eyes, allowing the pupils to re-adjust, as they constrict from the flash of the camera. However, these solutions are often difficult to implement in practice. For instance, turning a light off is difficult when the eye screening takes place in the same room where another patient is discussing their results with a nurse, receiving a foot sensitivity screening, or receiving nutritional counseling. Waiting 60 seconds before imaging the second eye is nearly impossible when a nurse has 150 patients in the queue waiting to be screened.

Internet speed and connectivity

One key difference in the eye screening workflow before and after the implementation of the deep learning system is that images are now uploaded to the cloud to get an assessment while the patient waits for results. On a strong internet connection, these results appear within a few seconds. However, the clinics in our study often experienced slower and less reliable connections. This causes some images to take 60-90 seconds to upload, slowing down the screening queue and limiting the number of patients that can be screened in a day. In one clinic, the internet went out for a period of two hours during eye screening, reducing the number of patients screened from 200 to only 100.

> Patients like the instant results but the internet is slow and patients complain. They've been waiting here since 6 a.m. and for the first two hours we could only screen 10 patients. -P8

Workarounds to the prospective study protocol

To the nurses, ensuring quality care for patients was just as important as not overly-inconveniencing them. In the case of

an ungradable image, the prospective study protocol originally stated that the patient should be referred to Pathum Thani Hospital, a larger provincial hospital with an in-house ophthalmologist that can examine the patient for DR and treat if needed. However, it is unlikely that the patient has severe or proliferative DR. Additionally, Pathum Thani Hospital is far away and inconvenient for many patients. Patients either need to take off work, or for older patients that cannot drive on their own, a child of theirs needs to take off work to drive them. We observed P8 urging a patient to go to the hospital when he was diagnosed with pterygium, another eye disease:

> P8: You should go to the hospital, there's pterygium in both eyes on the corneas. They will check for DR too.
>
> Patient: My child doesn't have time to take me.
>
> P8: The problem will get bigger. If it covers your eyes, you won't be able to see. If your child doesn't understand, have her call me.

This nurse was willing to go above and beyond to ensure patients got the treatment they need. But because of the large inconvenience it would cause patients to go to Pathum Thani Hospital, nurses were generally hesitant to refer patients there as a result of ungradable images alone, given the low likelihood that they have severe or proliferative DR.

Instead, we observed nurses working around the prospective study protocol; relying on previous workflows and criteria for determining whether or not to refer the patient to a specialist.

> I look at results and then determine the blood sugar result, the HbA1c. [Sometimes] people go to the hospital and it's a waste of time. You should always look at the blood result and if one eye is okay [and the other is ungradable], then I don't recommend it. But every time I ask if they want to go. -P6

We found that nurses took this approach across the three sites: in the case where an image was ungradable by the system, but the patient had no history of diabetic retinopathy and their blood sugar was well-controlled at the time of the visit, the nurse would make her own judgement call to send the patient home without a referral. P8 described what she does in the case of an ungradable image at her clinic, *"I look at their history. If it was bad last year, I refer. If it's okay, I send [the photo] to the doctor."*

DISCUSSION

Through our field research before and after deployment of the deep learning system, we discovered several factors that influenced model use and performance. Poor lighting conditions had always been a factor for nurses taking photos, but only through using the deep learning system did it present a real problem, leading to ungradable images and user frustration. Despite being designed to *reduce* the time needed for patients to receive care, deployment of the system occasionally caused unnecessary delays for patients. When network connectivity issues occurred, all patients present for a screening experienced delays or worse, rescheduled appointments. Finally, concerns for potential patient hardship (time, cost, and travel) as a result of on-the-spot referral recommendations from the system, led some nurses to discourage patient participation in the prospective study altogether.

Of these factors, *ungradability* had the largest impact on model performance and the patient experience, and this finding generalizes past system evaluation within a prospective study. The findings illuminate the tension between designing a threshold for the quality of images that the system will use to make an assessment, and the quality of images that arise from an imperfect, resource-constrained environment.

System thresholds that impact accuracy and safety (e.g., sensitivity and specificity) will always be an important consideration when deploying a deep learning system within a clinical environment. These thresholds are also something that need to be reevaluated as a system moves from prospective evaluation to widespread use. Appropriate thresholds during evaluation may be different from thresholds set once a system's accuracy is better understood. Researchers will always need to consider: what is the appropriate threshold to ensure accuracy and safety, and what is the risk of deferring a prediction based on imperfect data?

Through observations during a prospective study, we saw ungradable images increased wait times for patients at the clinic, and caused nurses to think twice before making referrals to ophthalmologists. As a result of conducting this research, we were able to inform an iteration of the study design: the study protocol now delays the patient referral for an ungradable image until an ophthalmologist, the overreader, has made a determination. This will presumably lower the rate of referrals, as the human overreader is willing to make a calculated judgment on the likelihood of disease, even if the system is not.

Our research highlights that end-users and their environment determine how a new system will be implemented; that implementation is of equal importance to the accuracy of the algorithm itself, and cannot always be controlled through careful planning. A recent longitudinal ethnography explored theoretical and empirical factors that lead to non-adoption of technology in healthcare [19]. Complexity across factors (e.g., medical conditions treated, organizational structure) increased the likelihood of non-adoption. When introducing new technologies, planners, policy makers, and technology designers did not account for the dynamic and emergent nature of issues arising in complex healthcare programs. The authors argue that attending to people—their motivations, values, professional identities, and the current norms and routines that shape their work—is vital when planning deployments [19].

Our findings suggest that even when a deep learning system performs a relatively straightforward task (e.g., focuses on retinal images and does not cross into multiple domains, organizational implementation challenges, or policy challenges), socio-environmental factors are likely to impact system performance. Our research also suggests that many environmental factors that negatively impact model performance in the real world have the potential to be significantly reduced or eliminated by tactical means (in our case, through lighting adjustments and camera repairs). However, such adjustments could

be costly and even infeasible in low-resource settings, making it even more important to deeply engage with contextual phenomena early on.

HCI Research on Deep Learning Systems

Thus far, most HCI studies of clinical systems have focused on interaction and presentation techniques for clinical data, or observations of the introduction of diagnostic and information systems that are not deep-learning based. Logistically, researchers have faced hurdles conducting studies within clinical environments [42], and often are limited to lab-based studies where clinicians use retrospective patient data [8, 9], or synthetic patient data in order to complete simulated tasks.

With machine learning, and in particular, deep learning, showing great promise to advance the field of medicine and improve care, there is a need for the HCI community to develop approaches for designing and evaluating machine learning systems in clinical settings.

In many ways, classic user-centered approaches still apply when conducting research in the age of deep learning. Formative work including clinical ethnography, observational studies, user interviews, and participatory design, have allowed researchers to build a foundational understanding of clinical problems and potential design solutions, prior to system development and evaluation. These methods are still critical to the success of deep learning-based systems, and can and should be conducted prior to, *and* in parallel with, system development. Similarly, paper- and digital-prototyping still hold as strong methods to help system designers understand user reactions to tools that make clinical predictions.

The problem occurs once researchers need to conduct studies using live clinical data. Once a deep learning-based system has been built, evaluations of model accuracy (sensitivity and specificity) have historically used only retrospective datasets, which may or may not be ecologically valid. With observational, *prospective* studies becoming an emerging approach to validate real-world clinical accuracy of deep learning models [1], we argue that conducting HCI research alongside prospective evaluations has many advantages. This approach enables researchers to evaluate using live data, with target clinicians, in a contextual environment. Furthermore, it provides opportunities to identify vital socio-environmental factors *ahead* of widespread deployment, such as issues that arise from system use or misuse, perceived utility, workflow integration, and experiential measures for both clinicians and patients.

As we observed, the design of the prospective study *protocol* itself needs careful consideration and is a ripe opportunity for participatory design [7] and service design [17]. The closer a protocol mirrors an ideal deployment post-evaluation, the more researchers will be able to understand the likelihood of future system use.

Practically speaking, prospective studies are a late-stage evaluation, meaning that the opportunity for multiple cycles of design iteration are significantly reduced. As such, teams developing deep learning models for clinical care should have a high level of confidence in a potential system and its deployment approach, prior to running a prospective study that evaluates model accuracy. For system developers, this means that formative research that provides a strong understanding of clinical users and their context is critically important to the success of such a system. By incorporating human-centered evaluations into deep learning model evaluations, and studying model performance on live data generated at the clinical site, we can reduce the risk that deep learning systems will fail in the wild, and increase the likelihood for meaningful improvements to patients and clinicians.

LIMITATIONS AND FUTURE WORK

This research is not without limitations. Although we evaluated a deep learning system in the wild, our study focused on primary users of the system: nurses and camera technicians. Further research is needed to understand how the system affects the patient experience; in particular, patients' trust of the result, and likelihood to act on the result. Additional research is also needed to understand how the system may alter the practices of ophthalmologists who evaluate patients that have received a prediction from the deep learning system. Lastly, as more systems are evaluated in clinical environments, an important area of future work includes the design of study protocols for conducting human-centered prospective studies and studies on end-to-end service design of AI-based clinical products.

Since this research, we have begun to hold participatory design workshops with nurses, potential camera operators, and retinal specialists (the doctor who would receive referred patients from the system) at future deployment sites. Clinicians are designing new workflows that involve the system and are proactively identifying potential barriers to implementation.

CONCLUSION

We are in a critical time in healthcare and technology, with deep learning algorithms poised to advance the field of medicine and improve patient outcomes. In this paper, we describe a sociotechnical study of a deep learning system for the detection of diabetic eye disease, used with patients in clinical care. Through observation and interviews with nurses and technicians in parallel with a prospective study to evaluate the deep learning system's accuracy, we discover several socio-environmental factors that impact model performance, nursing workflows, and patient experience. By conducting human-centered evaluative research prior to, and alongside, prospective evaluations of model accuracy, we were able to understand contextual needs of clinicians and patients prior to widespread deployment, and recommend system and environmental changes that would lead to an improved experience.

ACKNOWLEDGMENTS

We are especially thankful to the clinicians, staff, and patients at all participating clinics. Thanks to Kasumi Widner, Richa Tiwari, Nanlaphat Kuntee, and Variya Nganthavee for their research support. Thanks to Sara Gabriele, T Saensuksopa, Rebecca Shapley, and Brian Levinstein for their contributions to the system design. Thanks to Tayyeba Ali, Wing (Eric) Li, Ilana Traynis, Jonathan Krause, Rory Sayres, Elin Pederson, Greg Wolff, Lily Peng, and Greg Corrado for their helpful comments on this paper.

REFERENCES

[1] Michael D Abràmoff, Philip T Lavin, Michele Birch, Nilay Shah, and James C Folk. 2018. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med* 1 (Aug. 2018), 39.

[2] E Alberdi, A A Povyakalo, L Strigini, P Ayton, M Hartswood, R Procter, and R Slack. 2005. Use of computer-aided detection (CAD) tools in screening mammography: a multidisciplinary investigation. *Br. J. Radiol.* 78 Spec No 1 (2005), S31–40.

[3] American Academy of Ophthalmology. 2015. Eye Health Statistics. https://www.aao.org/newsroom/eye-health-statistics#_edn25. (2015). Accessed: 2019-9-7.

[4] Eta S Berner. 2007. *Clinical Decision Support Systems: Theory and Practice*. Springer Science & Business Media.

[5] Hugh Beyer and Karen Holtzblatt. 1997. *Contextual design: defining customer-centered systems*. Elsevier.

[6] Timothy W. Bickmore, Laura M. Pfeifer, and Brian W. Jack. 2009. Taking the Time to Care: Empowering Low Health Literacy Hospital Patients with Virtual Nurse Agents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 1265–1274. DOI: http://dx.doi.org/10.1145/1518701.1518891

[7] Keld Bødker, Finn Kensing, and Jesper Simonsen. 2009. *Participatory IT design: designing for business and workplace realities*. MIT press.

[8] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019a. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 4, 14 pages. DOI:http://dx.doi.org/10.1145/3290605.3300234

[9] Carrie Jun Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019b. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *CSCW Conf Comput Support Coop Work* 2019 (2019).

[10] Ronald A Castellino. 2005. Computer aided detection (CAD): an overview. *Cancer Imaging* 5, 1 (2005), 17.

[11] CDC. 2019. More than 100 million Americans have diabetes or prediabetes. Press Release. (February 2019). https://www.cdc.gov/media/releases/2017/p0718-diabetes-report.html.

[12] Elodia B Cole, Zheng Zhang, Helga S Marques, R Edward Hendrick, Martin J Yaffe, and Etta D Pisano. 2014. Impact of Computer-Aided Detection Systems on Radiologist Accuracy With Digital Mammography. *AJR Am. J. Roentgenol.* 203, 4 (Oct. 2014), 909.

[13] Kathleen Musante DeWalt and Billie R DeWalt. 2002. *Participant Observation: A Guide for Fieldworkers*. Rowman Altamira.

[14] Shelley E Ellis, Theodore Speroff, Robert S Dittus, Anne Brown, James W Pichert, and Tom A Elasy. 2004. Diabetes patient education: a meta-analysis and meta-regression. *Patient Educ. Couns.* 52, 1 (Jan. 2004), 97–105.

[15] Glyn Elwyn, Isabelle Scholl, Caroline Tietbohl, Mala Mann, Adrian GK Edwards, Catharine Clay, France Légaré, Trudy van der Weijden, Carmen L Lewis, Richard M Wexler, and others. 2013. "Many miles to go...": a systematic review of the implementation of patient decision support interventions into routine clinical practice. *BMC medical informatics and decision making* 13, 2 (Nov. 2013), 1–10.

[16] Geraldine Fitzpatrick and Gunnar Ellingsen. 2013. A review of 25 years of CSCW research in healthcare: contributions, challenges and future agendas. *Computer Supported Cooperative Work (CSCW)* 22, 4-6 (2013), 609–665.

[17] Jodi Forlizzi and John Zimmerman. Promoting service design as a core practice in interaction design.

[18] Maryellen L Giger, Heang-Ping Chan, and John Boone. 2008. Anniversary Paper: History and status of CAD and quantitative image analysis: The role of Medical Physics and AAPM. *Med. Phys.* 35, 12 (Dec. 2008), 5799.

[19] Trisha Greenhalgh, Joe Wherton, Chrysanthi Papoutsi, Jenni Lynch, Gemma Hughes, Sue Hinder, Rob Procter, Sara Shaw, and others. 2018. Analysing the role of complexity in explaining the fortunes of technology programmes: empirical application of the NASSS framework. *BMC medicine* 16, 1 (2018), 66.

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

[21] Mark Hartswood, Rob Procter, Mark Rouncefield, Roger Slack, James Soutter, and Alex Voss. 2003. 'Repairing' the Machine: A Case Study of the Evaluation of Computer-Aided Detection Tools in Breast Screening. In *ECSCW 2003*. Springer, 375–394.

[22] Matthew K Hong, Clayton Feustel, Meeshu Agnihotri, Max Silverman, Stephen F Simoneaux, and Lauren Wilcox. 2017. Supporting families in reviewing and communicating about radiology imaging studies. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 5245–5256.

[23] Jenchitr W, Hanutsaha P, Iamsirithaworn S, Parnrat U, Choosri P. 2007. The national survey of blindness low vision and visual impairment in thailand 2006–2007. *Thai J Pub Hlth Ophthalmol* 21, 1 (2007), 10.

[24] Marina Jirotka, Rob Procter, Mark Hartswood, Roger Slack, Andrew Simpson, Catelijne Coopmans, Chris Hinds, and Alex Voss. 2005. Collaboration and trust in healthcare innovation: The eDiaMoND case study. *Computer Supported Cooperative Work (CSCW)* 14, 4 (2005), 369–398.

[25] Pearse A Keane and Eric J Topol. 2018. With an eye to AI and autonomous diagnosis. *npj Digital Medicine* 1, 1 (Aug. 2018), 1–3.

[26] Ajay Kohli and Saurabh Jha. 2018. Why CAD failed in mammography. *Journal of the American College of Radiology* 15, 3 (2018), 535–537.

[27] Mark A Musen, Blackford Middleton, and Robert A Greenes. 2014. Clinical Decision-Support Systems. In *Biomedical Informatics*. Springer, London, 643–674.

[28] American Academy of Ophthalmology. 2002. International Clinical Diabetic Retinopathy Disease Severity Scale. http://www.icoph.org/dynamic/attachments/resources/diabetic-retinopathy-detail.pdf. (Oct. 2002). Accessed: 2019-12-17.

[29] World Health Organization. 2018. Vision impairment and blindness. https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment. (2018). Accessed: 2019-9-13.

[30] Sun Young Park, Pei-Yi Kuo, Andrea Barbarin, Elizabeth Kaziunas, Astrid Chow, Karandeep Singh, Lauren Wilcox, and Walter Lasecki. 2019. Identifying Challenges and Opportunities in Human–AI Collaboration in Healthcare. (2019).

[31] Laura Pfeifer Vardoulakis, Amy Karlson, Dan Morris, Greg Smith, Justin Gatewood, and Desney Tan. 2012. Using mobile phones to present medical information to hospital patients. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1411–1420.

[32] Sarah K Pontefract, Jamie J Coleman, Hannah K Vallance, Christine A Hirsch, Sonal Shah, John F Marriott, and Sabi Redwood. 2018. The impact of computerised physician order entry and clinical decision support on pharmacist-physician communication in the hospital setting: A qualitative study. *PloS one* 13, 11 (2018), e0207450.

[33] Paisan Raumviboonsuk, Jonathan Krause, Peranut Chotcomwongse, Rory Sayres, Rajiv Raman, Kasumi Widner, Bilson JL Campana, Sonia Phene, Kornwipa Hemarat, Mongkol Tadarati, and others. 2019. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *npj Digital Medicine* 2, 1 (2019), 25.

[34] Madhu C Reddy, David W McDonald, Wanda Pratt, and M Michael Shabot. 2005. Technology, work, and information flows: Lessons from the implementation of a wireless alert pager system. *Journal of biomedical informatics* 38, 3 (2005), 229–238.

[35] Nigam H. Shah, Arnold Milstein, and Steven C. Bagley, PhD. 2019. Making Machine Learning Models Clinically Useful. *JAMA* (08 2019). DOI: http://dx.doi.org/10.1001/jama.2019.10306

[36] Lauren Wilcox, Dan Morris, Desney Tan, and Justin Gatewood. 2010. Designing patient-centric information displays for hospitals. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2123–2132.

[37] Lauren Wilcox, Janet Woollen, Jennifer Prey, Susan Restaino, Suzanne Bakken, Steven Feiner, Alexander Sackeim, and David K Vawdrey. 2016. Interactive tools for inpatient medication tracking: a multi-phase study with cardiothoracic surgery patients. *J. Am. Med. Inform. Assoc.* 23, 1 (Jan. 2016), 144–158.

[38] World Health Organization. 2007. Global Initiative for the Elimination of Avoidable Blindness : action plan 2006-2011. https://www.who.int/blindness/Vision2020_report.pdf. (2007). Accessed: 2019-9-7.

[39] World Health Organization. 2014. WHO: Diabetes factsheet. https://www.who.int/news-room/fact-sheets/detail/diabetes. (2014). Accessed: 2019-9-13.

[40] World Health Organization. 2016a. Diabetes country profiles 2016 : Thailand. https://www.who.int/diabetes/country-profiles/tha_en.pdf?ua=1. (2016). Accessed: 2019-9-7.

[41] World Health Organization. 2016b. Diabetes country profiles 2016 : USA. https://www.who.int/diabetes/country-profiles/usa_en.pdf. (2016). Accessed: 2019-9-7.

[42] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 238, 11 pages. DOI: http://dx.doi.org/10.1145/3290605.3300468