Making Medical Assessments Available and Objective Using Smartphone Sensors

Alex Timothy Mariakakis

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Shwetak Patel, Chair

Jacob O. Wobbrock

James Fogarty

Program Authorized to Offer Degree:
Computer Science and Engineering

University of Washington

Abstract

Making Medical Assessments Available and Objective Using Smartphone Sensors

Alex Timothy Mariakakis

Chair of the Supervisory Committee:
Dr. Shwetak Patel
Computer Science and Engineering

Access to healthcare resources is a worldwide issue, but people do not always need access to such resources to discover a medical condition. Time and time again, people have been able to discover medical symptoms in themselves and others using their human senses—namely sight, touch, and hearing. However, observations with the senses are subjective, which can lead an untrained person to ignore their own symptoms and neglect treatment until their condition worsens. I propose that subjective health measures can be made objective with little additional burden using smartphone sensors. For my thesis, I provide three examples of how the smartphone camera can be used in place of visual inspection to automatically interpret diagnostic observations related to the eye; these projects cover medical conditions like glaucoma, pancreatic cancer, and traumatic brain injuries. My work in this space has lead me to uncover a number of challenges that impede progress in smartphone-based health-sensing. One of those challenges is ensuring that people make rational decisions when they are given health-screening tools despite not having formal training on diagnostic decision-making. I address this challenge by presenting a low-fidelity survey instrument that enables researchers to rapidly explore the effects of design decisions on the expected acceptability and effectiveness of a ubiquitous health-screening technology.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I'd like to thank my family. Our Sunday phone calls kept me connected with home, and you always offered to help me out whenever I needed it. Your constant love and support pushed me through my first long period of time away from our home state.

Second, I'd like to thank my advisors: Shwetak Patel and Jacob Wobbrock. Shwetak, you basically eliminated all the logistical hardships that people complain about when it comes to a graduate school experience. You gave me the freedom to work on whatever projects I wanted to work on, you gave me total freedom with my schedule, and I never had to worry about funding from quarter-to-quarter. I'll never take these luxuries for granted. Jake, you balanced my advising situation with precise and blunt feedback, which has helped me communicate my research to others. Every time I am about to write the word "while", I consider if I should use "although"; I can't get that out of my head. You also let me be a part of the MAD Lab, which exposed me to different opportunities and made me a more well-rounded researcher. Although he wasn't my advisor, I'd also like to thank Gaetano Borriello. If it weren't for you, I might not have applied to the University of Washington, and I may not have even been accepted here either. I wish I took better advantage of your mentorship while you were still around, but I look back fondly on the conversations we did have.

Third, I'd like to thank the UbiComp grad students who put up with me for up to six years. Sidhant Gupta, Gabe Cohn, Keyu Chen, Tien-jui Lee, and Tanvir Aumi, you all set a high bar for us to reach after I graduate; I hope I continue on that tradition. Mayank Goel, you were the second-in-command to Shwetak in the lab. You were my mentor on my first project in the lab and exemplified what it means to be a hard-working graduate student. Elliot Saba, let's be honest. At least half of the projects in the lab would not have been successful had it not been for your breadth of knowledge and wide skill set. You helped all of us with our annoying questions without a single complaint. Lilian de Greef, you are so kind and considerate. I know you'll be improving people's lives no matter

where you go in the future. Edward Wang, Ruth Ravichandran, Hanchuan Li, and Chen Zhao, most people can say that they joined a lab with one other person, yet rarely can someone say they joined with four others. Starting my graduate student career would have been a lot harder on my own. Edward, you were the Roger Federer to my Rafael Nadal. We worked together, struggled together, complained together, and grew together. You were my pacesetter throughout graduate school. Hanchuan, I really enjoyed the banter that went on between us while you were in the lab. We alternated between serious, thought-provoking conversations and arguments about penguins versus pigeons, but it was always a good discussion. Ruth, you truly exemplify what it means to be a well-rounded researcher. I marvel at your ability to be so technical and yet so mindful of the implications of your work. Chen, it's unfortunate that you didn't get to finish out your PhD here. You seemed to have it all figured out on day one, and I'm sure you've got it figured out wherever you are now. Eric Whitmire, where to begin. You lived in the same apartment with me for a couple of years. You worked with me on projects, both research- and school-related. Most importantly, we completed and failed multiple puzzle hunts together. Without question, you're the smartest and most tenacious person I know. Josh Fromm, we didn't see you around all that often, but it was always fun when you showed up to lab. Farshid Salemi Parizi and Morelle Arian, you both brought energy and fun to the lab right when things were about to get stale; keep doing what you're doing. Manuja Sharma, Xin Liu, and Matthew Whitehill, you all are the future of the lab. You seem to always ask the right questions at the right times. You're much quicker learners than I was when I started, and I know you'll do great. Chunjong Park, thanks for being my colleague and advising guinea pig at the same time. I hope you take the mantle of "computer vision + smartphone" from me and be twice as successful. Ravi Karkar, you're close enough to the lab that I'll mention you here. Our survey paper would not have been possible without you. Underneath all of the memes and jokes, you're a very

well-polished academic researcher with a lot of knowledge about the field that I hope to get someday.

Fourth, I'd like to thank my friends outside of the lab. Vincent Lee and Shumo Chu, we all started our Seattle experience together as housemates. It has been nice having people who appreciate good food, good trips, and good gossip. Nacho Cano, Daniel Miller, and Shrainik Jain, there was nothing more relaxing during graduate school than sitting at a bar with you all and complaining about whatever was on our minds. I hope we all stay in touch with the occasional international trip that Vincent organizes.

Fifth, I'd like to thank all of the clinical collaborators who handed me a research direction on a silver platter: Joanne Wen, Jim Taylor, Anthony Law, and Lynn McGrath. You all provided me guidance as I navigated an emerging field of research. Each of you has trained me so that I someday can be considered an expert in the mobile health space. I'd also like to thank my internship mentors: Souvik Sen, Vijay Srinivasan, Kiran Rachuri, Evan Welbourne, Daniel Avrahami, Gonzalo Ramos, and Asta Roseway. That's a long list, but each of you gave me a different research and advising experience. Each internship was a breath of fresh air during my career as a graduate student.

Chapter 1

INTRODUCTION

1.1 Problem Statement

Healthcare systems around the world are heavily strained. The World Health Organization reported that there were 1.53 physicians for every 1000 people worldwide as of the beginning of 2017 [263]. Although this statistic surpasses the organization's desired standard of 1 physician for every 1000 people, there are a number of caveats that must be considered. First and foremost is the fact that the distribution of physicians is not uniform. The global statistic is skewed in favor of developed nations. A deeper examination reveals that 44% of the WHO's member states fall below the WHO's standard [263]. Even within developed nations, physicians are far more likely to live in urban areas, forcing those in more rural areas to travel to receive medical attention. Not only are the numbers of physicians in favor of developed nations in urban areas, but also the expertise of those physicians and the resources available to them. Experienced and well-equipped physicians are more likely to live in urban, developed areas and are also likely to be more proficient in diagnosing diseases and prescribing treatment. These statistics also only consider physicians: medical professionals who are trained to make diagnoses and prescribe medications to treat symptoms of various illnesses. When a physician is in doubt, they will often refer their patient to a doctor who specializes in a particular area of medicine, such as orthopedics or dermatology. The number of specialized doctors worldwide is fewer than physicians, especially considering all of the different specialties they may have. No matter the distribution of physicians or doctors

in the world, it is impractical for people to visit clinics everyday to be tested for every possible condition daily.

The lack of clinical resources begs for a balance between centralized in-clinic testing and distributed testing in homes and communities. One way to support that balance would be to encourage the development of less expensive medical devices. Although Moore's Law would leave one to believe that new hardware could be the answer, such a vision is still impractical. People would need to purchase an automated blood pressure cuff, a "smart stethoscope", and tens of other devices; even if those devices are only $10 each, those costs may be prohibitive for people with lower incomes. Even if the devices were inexpensive, people would need to train themselves on how to use each of the devices, maintain the devices over time, and replace the devices once they have become obsolete. This vision is also flawed because it takes time for companies to develop and manufacture these tools, and problems related to health are serious enough that their solution cannot be delayed by years.

I believe that a way to avoid these barriers is by taking advantage of a device that is already ubiquitous throughout the world: the smartphone. Smartphone penetration reached 66% in 2018, accounting for 73% of internet consumption [243]. Perhaps more importantly, smartphones come with an array of sensors that can be used to understand physiological and biological information. Cameras can process visual information, microphones can process audio information, and even the accelerometer and gyroscope can process information related to motion and proprioception.

In fact, these sensors mirror human senses—sight, hearing, and touch—that we sometimes use to discover medical symptoms within ourselves or each other. We can see when a wound is not healing properly, listen for wheezing, and feel when we are shivering. The reasons for using sensors instead of our human senses include increased robustness, consistency, precision, and accuracy.

The fever serves as an interesting example. Fevers are characterized by an internal body temperature roughly greater than 37.5°C (99.5°F). If you ever had a fever as a kid,

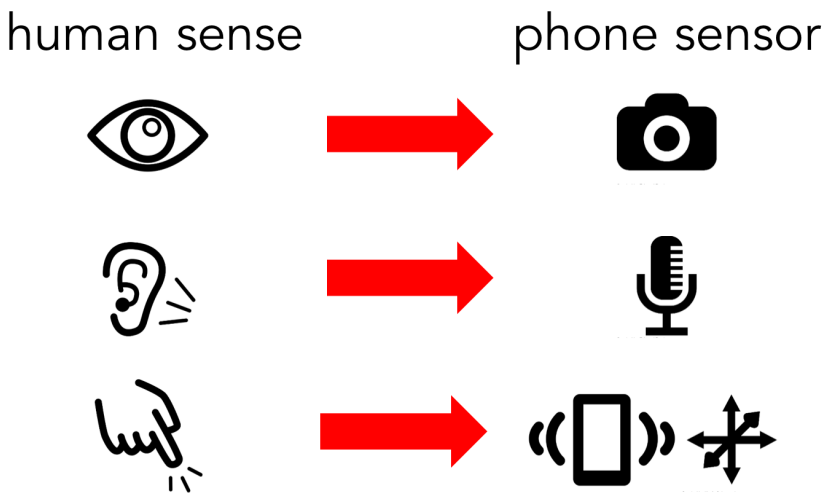

Figure 1.1: Smartphone sensors can process the same information that we can with sight, hearing, and touch.

you probably remember your mother or father placing the back of their hand against your forehead to feel your temperature. The body's surface temperature is directly correlated with its internal core temperature, but the surface temperature is usually cooler. Did your parents account for this difference? Did they account for the fact that their own hand may have been cold when they touched you? Most importantly, what did they have to feel in order to decide that you had a fever? At best, measuring temperature through touch leads to subjective descriptions like "warm" or "cold". Even if your parents knew exactly what they needed to feel, their judgment could have been clouded, for better or for worse. You may have felt only slightly warm, but they were so worried that they decided you were sick anyway. In the opposite situation, they may have been believed that you were faking sickness to get out of class and convinced themselves to send you to school anyway, even if you felt very warm.

Banco and Veltri conducted a study in 1984 [12] to test how accurate mothers were at subjectively assessing the presence of a fever in their children. They found mothers were

only correct 52.3% of the time when they said their children had a fever. Such studies motivate the need for medical devices that are available to provide quantitative measurements at home. In fact, devices already exist for measuring body temperature: in-ear and oral home thermometers. However, not everyone has a thermometer at home either because of cost or the mere inconvenience of purchasing one. Furthermore, thermometers are only good for measuring temperature, limiting their efficacy to a single task. Different symptoms warrant different tests, which in turn warrant different medical devices.

Smartphone-based health-screening apps can serve two purposes. First, a smartphone app can serve as an accessible screening tool that decides whether a particular measure of a person's health is normal or abnormal. Some may download such an app because they know that their family has a history of a particular condition, while others may download an app out of genuine curiosity. When the app determines that a person has an abnormal result, the app can recommend that they seek a medical professional who can perform a more direct clinical test (e.g., blood draw, MRI). Figuratively, the app could replace a physician's referral as the user's admission ticket to see a specialized doctor.

The second purpose that a smartphone-based health-screening app can serve is disease management. For example, a doctor might want to ensure that the beta blockers they prescribed for a patient's hypertension are working. To support or repudiate this decision, the patient would ideally check their blood pressure on a daily or even hourly basis. Getting such data would either require the patient to make frequent visits to the clinic, which would be inconvenient, or to purchase their own in-home blood pressure monitor, which constrains how often the patient can leave their home. A smartphone-based solution would be a more convenient alternative to those two options.

1.2 Thesis Statement

Throughout this dissertation, I provide evidence to support the following thesis statement:

> *Technological and scalability barriers to some medical assessments can be addressed through smartphone-based sensing tools; moreover, the acceptability of these tools can be addressed through surveys that reveal how these tools and their results are likely to be regarded by potential users.*

Clinical testing and expensive medical hardware introduce **technological and scalability barriers** to at-home health screening. I will demonstrate how we can take inspiration from subjective observations that people make with their senses and make them more accurate, precise, repeatable, and pervasive using smartphone sensors. In particular, my work focuses on the use of smartphone cameras for automating visual observations related to the eyes. Based on my experience from these projects, I have come to realize that we are far from seeing apps dedicated to specific symptoms in today's medical and technological infrastructure. I describe four specific challenges in this regard and propose possible opportunities for future work. The last part of my dissertation focuses on one of those challenges involving **acceptability barriers** that may impede the adoption of smartphone-based health-screening technologies. I present a survey instrument that helps researchers investigate the perceived utility and efficacy of hypothetical technologies without requiring a physical prototype.

1.3 Outline of This Document

Chapter 2 provides an overview of relevant prior work. It begins with a discussion of how researchers first began to use mobile devices for health applications in ways that did not involve sensing. That overview is followed by examples of sensor-based health apps. Since the projects I use to support my thesis relate to symptoms of the eye, I dedicate

a separate section to that research space. The chapter concludes with an overview of health-related decision-making support and people's understanding of sensing systems.

The next three chapters detail smartphone-based health-screening apps that address medical symptoms that manifest in the eyes. Chapter 3 describes iPressure, a smartphone app and inexpensive hardware attachment for measuring intraocular pressure. Eye pressure is an important risk factor for glaucoma, a progressive optic neuropathy that can lead to blindness. Besides the large and complicated setups found in ophthalmologist clinics, current methods for measuring eye pressure either require a dedicated battery-powered device or expertise in reading narrow analog scales with constantly fluctuating values. The iPressure system allows untrained, yet responsible individuals to perform fixed-force applanation tonometry, a clinically validated technique for measuring intraocular pressure. The main feature of the attachment is a clear, free-hanging mass that the examiner places on top of the patient's eye. When the mass is allowed to rest on the eye, an area of the eye is flattened out. The mass is clear, so the smartphone's camera can see the applanated surface. By measuring the surface's area, the counterpressure exerted by the eye to support the mass can be calculated, which in turn leads maps to a measurement of the pressure within the eye. To mitigate errors associated with manual observation, the iPressure app uses image processing to precisely measure the correct diameter of the applanation surface and aggregate those diameters in a clinically relevant manner.

Chapter 4 describes BiliScreen, a smartphone app that quantifies the extend of jaundice in an individual. Jaundice, or the yellow discoloration of the skin and eyes due to the buildup of a compound called bilirubin in the blood, is one of the few externally visible indicators of an issue in the pancreas. This is potentially critical for the diagnosis of pancreatic cancer, which has one of the worst survival rates amongst all forms of cancer because diagnoses are often made after the disease has already progressed. Ruiz et al. [214] found that jaundice is only perceptible to the naked eye at levels beyond the clinical threshold for concern, exposing a gap in coverage. BiliScreen aims to fill that gap

by quantifying the extent of jaundice in a patient's sclerae (i.e., the white part of the eyes). Computer vision is applied to pictures of the patient's eyes to isolate the sclerae from the rest of the face. After the color of the sclerae is summarized, machine learning is used to produce an estimate of the patient's bilirubin level. The ambient lighting of the room and the color response of the smartphone's camera sensor both have an effect on the appearance of the sclerae. I have investigated two approaches to mitigate these effects. The first involves a low-fidelity box (similar to a head-mounted virtual reality display) with the back-facing camera and flash facing towards the patient to provide complete control over the lighting environment. The second uses paper glasses with colored squares printed around the rims. Instead of standardizing the lighting conditions for the photo, the glasses allow for the images to be calibrated against known references.

Chapter 5 describes PupilScreen, a system that tests how a person's pupils react to a light stimulus. Measuring the pupillary light reflex is one of the clinical screening tests used by athletic team doctors and EMTs to judge the extent of a person's potential traumatic head injury (TBI). The most ubiquitous test used by those groups is the penlight test. In short, the penlight entails an examiner shining a miniature flashlight towards a person's eyes and describing how their pupils shrink in size (e.g., "quickly", "by a large amount"). Research has shown that the penlight test can lead to incorrect decisions or disagreement between clinicians, motivating the need for a more quantitative technique [165, 265]. To use PupilScreen, an examiner uses the box described earlier for BiliScreen. The box blocks out the ambient lighting while allowing the flash to stimulate the eyes, thereby standardizing the amount of light that reaches the patient's eyes. A fully-convolutional neural network architecture is used to measure the pupils' diameters within each frame to generate a graph of pupil dilation over time. That graph can be summarized with clinically relevant measures like constriction amplitude and velocity. After presenting results on PupilScreen, I describe some of the steps I have taken to make PupilScreen more usable and eventually support a large-scale study.

Chapter 6 describes four challenges I have identified in smartphone-based health-screening: hardware limitations, smartphone heterogeneity, quality control during data collection, and data interpretation for untrained users. These challenges have emerged from experiences by myself and my colleagues in the mobile health space. Beyond enumerating the challenges, I also describe current approaches that have been used to address them and offer future solutions that can be explored in the future.

Chapter 7 delves into a potential aid for researchers to address investigate the last challenge. I describe a survey instrument that allows researchers to rapidly explore the effects that design decisions have on a ubiquitous health-screening technology's acceptability and effectiveness earlier in the design process, without the need for a functional prototype The survey instrument is framed around the Health Belief Model [102, 109], an established psychological model that attempts to explain and predict short- and long-term health behaviors. The survey presents respondents with a hypothetical scenario regarding their health and then introduces a hypothetical health-screening technology that claims to screen for the medical condition in question. The respondent is probed about constructs within the Health Belief Model and other parameters that could explain their decision-making process. The results are analyzed using structural equation modeling (SEM) to determine the significance of hypothesized causal relationships. A deployment of the survey instrument revealed that in some cases, people were willing to take change their course of action in response to a test result regardless of the app's reported accuracy.

Chapter 2

RELATED WORK

My work draws inspiration from the vast literature regarding health-related smartphone apps, wearables, and sensing as a whole. In this chapter, I will describe related works from the mobile health and medical sensing communities. I will then cite works that relate specifically to the diagnosis of medical conditions within and through the eye since that is the focus my work has taken. I conclude this chapter by describing how

2.1 Health-Related Apps without Sensing

Before researchers began to rigorously explore the full capabilities of sensors on smartphones for health applications, people had already begun to take advantage of mobile devices' user interfaces and wireless capabilities for medical screening and diagnosis. Many such works come from the Information and Communication Technology for Development (ICTD) community, which focuses on how technology can improve the lives of underserved populations in low-income regions.

The most common use of mobile devices in mobile helath applications has been for one- or two-way communication between community health workers and affected populations. One-way communication can entail a server that sends bulk SMS messages to a large number of recipients. Kunutsor et al. [129] built such a system to limit missed clinic visits. One-way communication can also entail users having to submit messages in a pre-defined structure, such as the work of Asiimwe et al. [8] who used mobile devices to communicate diagnostic test results. Two-way communication provides feedback between community health workers and their patients. This feedback can be automated,

such as the work by Ngabo et al. [179] that seeks to improve maternal and child health. The feedback can also be manual. Lester et al. [138] propose a system called Weltel in which community health workers and patients exchanged one-word messages about HIV medication. Their system improved medication adherence, though almost 30% of their participants had to be called since they did not provide a response to the SMS message. Perrier et al. [194] provide an example that combines the strengths of both automated and manual messaging. Communication was initiated with bulk messaging, but nurses and health workers read free-form responses from those who needed attention.

There have also been examples of mobile health applications that simply provide information in a more convenient form factor. DeRenzi et al. [57], for example, designed and deployed an electronic version of the Integrated Management of Childhood Illness (IMCI). The IMCI is a protocol written by the World Health Organization to standardize the methods by which clinicians assess children for various issues (e.g., malnutrition, dehydration, malaria). To step through the procedures, clinicians are instructed to ask the child's parents questions and perform observational tasks like checking for sunken eyes or pinching the child's skin. The protocol is typically handed out as a printed booklet[1] that directs clinicians to different pages depending on the particular child's condition. The electronic IMCI by DeRenzi et al. automates the logical flow of instructions, but not the sensing of the observational data that guides the flow. Mitchell et al. [167] deployed a similar effort for standardizing the HIV screening methodology used in South African AIDS treatment centers.

In a step towards sensing through mobile devices, researchers have leveraged location information from call records (i.e., nearest cell tower) to track human mobility, which in turn improves predictions on the spread of parasites. Wesolowski et al. [254], Vazquez-Prokopec et al. [244], and others have are just some examples of how population-level

[1] http://www.who.int/maternal_child_adolescent/documents/IMCI_chartbooklet/en/

travel patterns can be combined with detailed geographical models of disease risk to do such analysis.

2.2 Smartphone-Based Health Sensing Apps

Smartphones come with a number of sensors that can be used to process physiological information in-the-wild. Below, I categorize projects in this space according to the sensor they used.

2.2.1 Camera

My work has examined how computer vision and machine learning can be used to objectify observations that are normally subjective using the smartphone camera. Smartphone cameras have been used by a number of other researchers for diagnosis. Wadhawan et al. [248], for example, use image processing and pattern recognition to categorize skin lesions as malignant melanoma or benign moles. Their system is intended to replace the "ABCDE rule" (asymmetric, irregular border, varied color, wide diameter, and elevated) that dermatologists teach their patients to screen themselves for skin cancer [175]. BiliCam, by de Greef et al. [91], uses the smartphone camera to screen newborns for jaundice, a yellow discoloration of the skin. Neonatal clinic nurses see so many babies that they can train themselves to identify cases of jaundice irrespective of their skin tone. Once the baby leaves the clinic, though, that responsibility falls upon the newborn's parents, who likely does not have such a mental model to rely upon. Face2Gene[2], by Ferry et al. [67], is a smartphone application that diagnoses rare diseases by analyzing deformities in facial structure. The original Face2Gene application was deployed as a searchable image database that could be used as a reference for observation, but the application was re-launched with deep learning and computer vision analysis to automate the diagnosis procedure. Cho et al. [38] and Hashemi

[2] https://suite.face2gene.com/

et al. [98] have both developed tools that screen children for autism. Cho et al. monitor the child's gaze pattern, noting that children with autism tend to have a more scattered gaze trajectory. Hashemi et al. analyze children's emotional reactions to video clips.

Grimaldi et al. [94] use a smartphone camera and flash in combination to perform photoplethysmography, a technique that measures a person's pulse. As blood rushes in and out of the fingertip with the same frequency as the heart rate, the transparency of the finger changes slightly, which can be picked up by the camera. Wang et al. [252, 253] take photoplethysmography a step further with HemaApp, a smartphone app that estimate the hemoglobin concentration against total blood volume by analyzing the color channels. HemaApp uses machine learning to estimate the absorption coefficients of hemoglobin and plasma at the smartphone flash's broadband wavelengths.

Computer vision-based solutions are not only useful for diagnosing conditions on the body, but also for reading disposable, biochemical immunoassay papers that change their appearance according to the presence of a specific substance. Mondanyali et al. [174], Dell et al. [56], and the company Mobalysis[3] each propose their own smartphone-based platform that quantifies the papers' change in hue or brightness. Commercial pregnancy tests are an example an immunoassay that is usually easy to read, but there are others that are more difficult. For instance, interpretation of the CD4 rapid test for HIV treatment depends on the intensity of the assay's capture line [47]. Using computer vision removes the need for qualitative guesswork from untrained users.

2.2.2 Microphone

Before smartphones truly became "smart", all mobile phones were guaranteed to have one sensor on them to process speech: the microphone. Thinking of smart devices as portable voice recorders, researchers have devised algorithms to process speech for various reasons. Mauremi et al. [173] use emotional acoustic features mined from phone

[3] `https://mobalysis.com/rapid-diagnostic-testing/`

conversations to predict manic and depressive episodes of people suffering from bipolar disorder. Dubey et al. [59], Bot et al. [24], and others also extract features from speech, but do so tracking the progression of Parkinson's disease. The speech of those affected by Parkinson's disease can be characterized by monotony in pitch, reduced loudness, irregular speech rate, and imprecise consonants.

Microphones have been used to detect and describe sounds other than speech. SpiroSmart [131] assesses a person's lung function after they perform an explicit breathing maneuver towards a smartphone, turning the microphone into an uncalibrated flow sensor. Microphones measure sound, which is a pressure wave. According to Bernoulli's principle, the flow rate of a fluid like air is inversely related to pressure; therefore, a microphone measures an increase in flow rate as a decrease in pressure. To perform the test, a user holds their phone out at roughly arm's-length, inhales, and then exhales as if they were using a spirometer mouthpiece—mouth wide open while forcing as much air as they can for as long as possible. Like a flute, a person's trachea becomes quieter when it is obstructed and higher pitched when it is restricted.

Work by Larson et al. [130] and Sun et al. [228] provide examples of systems that classify sound-related respiratory symptoms like coughs and sneezes. When pressed against the neck, a microphone can detect sounds from within the body. This is the intuition behind BodyBeat [204], a wearable system designed to detect non-speech sounds like chewing, laughter, and coughing. The microphone is not the only sensor that can detect sound in this way. An accelerometer pressed against the neck can also pick up the vibration of the vocal chords during speech. Mehta et al. [166] designed such a system to diagnose voice disorders.

2.2.3 Accelerometers and Gyroscopes

Inertial measurement units (IMUs) allow smartphones to sense their own orientation to deliver content on the screen in the best manner possible. However, motion sensors have

been used for decades to understand the motion of limbs and bodies. Given the variety of motions that people perform over the course of a day, most clinical motion tracking relies on the detection of periodic motions or explicit gestures. Joundi et al. [111], Keijsers et al. [122, 121], and Tsipouras et al. [242] are just a selection of researchers who have analyzed time- and frequency-domain features to track the progression of tremors for people with Parkinson's disease, multiple sclerosis, and other conditions characterized by essential tremor. Joundi et al. asked participants to strap their smartphones to their arms, while the other projects asked participants to wear dedicated motion sensors on different parts of their body (e.g., ankles, wrists, sternum). The mPower app, developed by Sage Bionetworks [24], tests how quickly people with Parkinson's can tap a touchscreen as a longitudinal measure of their motor abilities.

2.3 Diagnosis within the Eyes

The eye is one of the most complex organs in the body, making it susceptible to a number of different complications. Pamplona et al. have developed several inexpensive attachments for smartphones to diagnose conditions related to the eye. Much like an eye chart, their hardware presents stimuli to the user. Rather than conversing with a clinician, the user interacts with their smartphone depending on what they see; this is an iterative procedure that goes on until a result is reached. In NETRA [190], refractive errors are identified by asking the user to align patterns projected through a microlens display and pinhole. In CATRA [189], cataracts are localized by scanning the eye with a beam of collimated light and asking the user for feedback about the spread of the beam. EyeMITRA [134], being a wearable camera, varies slightly from the other two projects. It is meant for mobile retinal imaging, so it does not perform diagnosis on its own. The user is placed within the loop of the system by being asked to focus on focal points shown in the other eye, which in turn focuses the camera on the opposite side.

Researchers have explored other ways of diagnosing ocular conditions. Abdolvahabi et al. [1] discuss the possibility for digital photography to catch the early onset of

retinoblastoma in newborns; they found that if the common "red-eye" effect in the pupils is replaced with a milky white color (known as leukocoria), it could indicate tumors in the back of the eye. D-Eye[4] is a smartphone adapter for performing fundoscopy. Bastawrous et al. [15] and Giardini et al. [85] propose a number of attachments for diagnosing visual acuity and glaucoma. The first project I present in this dissertation, iPressure, uses a passive hardware attachment to perform fixed-force applanation tonometry, a technique for measuring intraocular pressure and assessing a person's risk for glaucoma.

The eye is part of the body's nervous system, so it is susceptible to non-ocular conditions that manifest within the body as well. Hyperemia and conjunctivitis are two symptoms that appear in the sclera, affecting both the amount and contrast of the surface-level blood vessels [99]. Osteogenesis imperfecta, a genetic disorder that results in brittle bones, produces a blue tinge in the sclera [224]. Diabetes results in fewer capillaries, dilated macrovessels, and changes in the curvature in the sclera's covering [186, 187]. My project, BiliScreen, automatically quantifies the extend of jaundice in the sclera as a potential predictor for pancreatic cancer.

Ramlee and Ranji [205] propose a system that identifies arcus senilis, a condition that manifests as a cloudy ring at the corneal limbus between the iris and the sclera. Arcus senilis can be a sign of impaired lipid metabolism [14], which is a risk factor for conditions like hypercholesterolemia, or decreased blood flow to the unaffected eye due to carotid artery disease [226]. Ramlee and Ranji's paper describes an algorithm that involves iris segmentation and the identification of lipids in the eye, but they do not provide an evaluation of their system. Limbus sign (i.e., dystrophic calcification) also appears as a cloudy ring at the corneal limbus, but it indicates a buildup of calcium in the blood [257]. Kayser–Fleischer rings are dark brown rings that develop at the corneal limbus from copper deposition caused by Wilson's disease [163].

---

[4] https://www.d-eyecare.com/

Finally, the pupils are particularly useful for assessing neurological function since their appearance is supposed to change due to stimuli like light and stress. PupilScreen tracks the pupillary light reflex, or how a person's pupils respond to changes in light. Non-reactive pupils can indicate a traumatic brain injury or elevated intracranial pressure [30]. The pupils' shape and color can also be important to observe. Oval-shaped pupils can indicate cerebrovascular illnesses like hypertensive cerebral hemorrhaging [70] or neurosyphilis [71].

2.4 Supporting Health-Related Decision-Making for Non-Experts

To the best of my knowledge, there has not been prior commentary on evaluation methods for health-related decision-making technologies, but there has been such commentary in the related field of behavior change. Behavior change aims to change a person's habits to prevent disease, whereas decision-making support focuses on the similar goal of getting a person to take a single health-promoting action (e.g., going to the doctor, stopping drinking coffee). Klasnja et al. [123] provide a thorough meta-analysis on different evaluation approaches for health behavior change, including interviews, field studies, and randomized control trials. They come to the conclusion that system evaluations should be tailored to their specific intervention strategies (e.g., self-monitoring, conditioning, tunneling [72]). Although Klasnja et al.'s commentary concentrates on evaluation strategies for after a technology is ready to be deployed to end-users, their call for additional evaluation strategies motivates my survey instrument for early-stage technologies.

Hekler et al. [100] urge HCI researchers to utilize and contribute to behavioral science theories. In particular, Hekler et al. call for the development of new strategies for investigating design recommendations that balance abstraction with contextual relevance. They note that many design guidelines for behavior change technologies are often tied to assumptions about the specific technology that was studied, leading to findings that are less generalizable than intended.

One way to provide abstraction is through vignettes: brief, carefully written situations that include a subset of key features to simulate a real-world scenario [3, 9]. My survey instrument uses hypothetical scenarios and technology descriptions to probe people's decision-making; however, I am not the first to do so. Evans et al. [65] and Bachmann et al. [10] both provide systematic reviews on this field of research. Two of the prominent vignette-based methods they describe are conjoint analysis [92] and judgment analysis [95, 44]. In conjoint analysis, participants are asked to rank or select among different versions of an object with slight variations across a feature set. As more of these decisions are made, the influence of each feature on the participant choices can be elicited. As an example of health-related conjoint analysis, Ryan [216] used conjoint analysis to examine the values that are important to people pursuing in vitro fertilization. In judgment analysis, participants are asked to decide whether they would take action in a series of scenarios with different features. Participant decisions are compared to the optimal decisions according to an oracle, producing correlations between the weighting of the features in both cases. As an example of health-related judgment analysis, Kee et al. [120] used the method for evaluating prioritization decisions within a dialysis program.

My work diverges from existing vignette-based methods in several ways. First, my survey instrument not only elicits preferences between different feature combinations, but also examines how those features influence people's health-related decision-making. Second, I do not assume that an optimal decision exists for my hypothetical scenarios. The fact that a person may change their course of action at all is an interesting result that I believe should be studied further.

2.4.1 Evaluating the Perception of Sensor-Based Technologies

Prior work within UbiComp and HCI has looked at the role that transparency plays in the perception of sensor-based systems. Dzindolet et al. [60] argue that decision-making

systems capable of generating explanations for their behavior (i.e., intelligible systems) yield increased trust and acceptance. Lim and Dey [144] investigated the issue further through a survey instrument. By varying inference certainty and explanation thoroughness, they found that intelligibility was helpful for applications with high certainty yet harmful for applications with low certainty that still performed their tasks successfully. Kay et al. [118] recognized that telling users the accuracy of a sensor-based technology affects their perception of it. They deployed a survey instrument where respondents were asked to decide if they would be willing to use hypothetical technologies with various levels of precision and recall. In doing so, they created a tool for measuring a technology's acceptable level of accuracy. Kay et al.'s tool supported the hypothesis that when a technology had a high cost for a false positive (e.g., a burglary alarm that automatically calls the police), respondents prioritized precision over recall.

My survey instrument differs from the aforementioned studies in several ways. First, prior work has primarily focused on technology acceptability. My survey instrument goes a step further, investigating how a technology might affect a person's course of action (i.e., its effectiveness), which is an important outcome for ubiquitous health-screening technologies. I am able to do so by restricting my instrument to health-related decision making and by building my survey on top of the HBM. The HBM allows a researcher to tease apart the effect that their technology intervention might have on a person's decision-making from other factors like a person's educational background or a person's perception of a medical condition. Second, prior work has looked at the roles intelligibility [144] and accuracy [118] play in the perception of a technology, but our survey considers transparency more broadly. I allow researchers to reveal as much or as little information as they want about their technology (e.g., interface, price, sensing modality). Using structural equation modeling, researchers can evaluate how the information they reveal affects the technology's perceived acceptability and effectiveness.

Chapter 3

IPRESSURE

Intraocular pressure (IOP) is the innate fluid pressure within the eye. IOP is maintained by the trabecular meshwork, which manages the leakage of the aqueous humor in the anterior chamber of the eye. The typical IOP of humans ranges from 7-21 mmHg with a mean of approximately 16 mmHg. Elevated IOP is an important risk factor for glaucoma, a progressive optic neuropathy that can lead to visual field defects or eventual blindness. A study carried out by Quigley and Broman in 2006 [201] predicts that the global population affected by glaucoma will reach 80 million by 2020; it further postulates that half of the people living with glaucoma are unaware that they have the disease, which can largely be attributed to a lack of resources or incentive for IOP assessment. Glaucoma also imposes a significant burden on the US healthcare system, costing roughly $3 billion USD and over 10 million visits to physicians per year [206].

Tonometry is the diagnostic procedure used for measuring IOP. Although tonometry comes in many different forms, most require the experience of a trained eye care professional and access to dedicated medical devices. These constraints make tonometry difficult in low-resource environments. Smartphones, on the other hand, have seen a rapid uptake all over the world and contain a myriad of sensors that can be used for mobile health applications.

In this paper, I propose iPressure, a smartphone-based system that allows for minimally trained individuals to perform IOP assessments on other individuals. Rather than requiring precision from specialized hardware or a trained professional, the precision of this system is placed within the smartphone. The user attaches a low-cost smartphone adapter that I have developed to emulate a technique called fixed-force

applanation tonometry. While the patient lies supine, the cylinder inside the instrument is rested on the patient's eye, allowing the smartphone's camera to automatically detect and measure the applanation surface, from which the patient's IOP may be inferred.

I evaluated iPressure's ability to measure IOP in a lab study with two *ex vivo* porcine eyes. With such a controlled setup, I am able to vary the IOP within the same eye, avoiding possible confounds attributed to physiology or diurnal variation; however, applanation tonometry requires data from a clinically validated lookup table to convert diameter measurements to IOP values, and such a table does not exist for porcine eyes. I instead fit those measurements to the underlying physical model that governs the process of applanation. I find that my results obey those models with Pearson correlation coefficients of 0.89 and 0.88 for the two porcine eyes.

My contribution comes in three parts:

1. The design of the iPressure smartphone attachment for performing fixed-force applanation tonomtry,

2. The use of computer vision to automatically detect and measure the applanation surface, and

3. An evaluation of iPressure on two *ex vivo* porcine eyes.

3.1 Related Work

Before discussing how iPressure is able to convert a smartphone to an automated tonometer, it is important to discuss the methods of tonometry that are currently available and their underlying physical principles. I also mention prior work that has been published on the considerations that must that must be taken into account with tonometry.

3.1.1 Forms of Tonometry

There are four broad classes of tonometry: applanation, indentation, dynamic contour, and rebound. I detail each of those classes below:

Applanation Tonometry

The clinical gold standard for measuring IOP is Goldmann applanation tonometry [227]. Applanation tonometry in general relies on Goldmann's observation that "the pressure in a sphere filled with liquid and surrounded by an infinitely thin membrane is measured by the counterpressure which just flattens the membrane" [87], also known as the Imbert-Fick law. In Golmann applanation tonometry, a form of fixed-area tonometry, a topical anesthetic with a fluorescein dye is placed on the eye. When the dye mixes with the tears and the eye is fluoresced with a cobalt blue light, the dye appears as a brighter yellowish green. A split optical prism is then pressed against the eye, resulting in two semicircles. The ophthalmologist adjusts the force exerted by the prism until the semicircles align on opposite ends, indicating that the area of the applanation surface has reached a predetermined quantity. That force measurement is mapped to an IOP value using a clinically validated lookup table [196]. The complement to fixed-area tonometry is fixed-force tonometry. Instead of measuring the force required to make an applanation surface of known area, a cylinder of known mass is allowed to rest on the eye without any external forces and the area of the applanation surface is mapped to an IOP value. An example of a fixed-force tonometer is the Maklakov tonometer, which entails applying ink on the eye and the mass that applanates the eye to measure the area of ink that was transferred between the two [150]. The system I propose in this work also uses fixed-force tonometry.

Indentation Tonometry

The underlying physical principle behind indentation tonometry is similar to that of indentation tonometry: a fluid-filled object will indent to a greater degree when the internal pressure is low. The most well-known indentation tonometer is the Schiøtz tonometer, first developed in 1905 [218]. Most Schiøtz tonometers are analog in nature, moving a needle across a narrow scale that must be precisely read by the user.

The most common portable tonometer is the TonoPen [191], a handheld, battery-operated device that uses a combination of applanation and indentation tonometry. To help the user operate the TonoPen, the device produces an audible click after contact with the cornea. The TonoPen has a built-in microprocessor that processes readings and accounts for variability, removing the possibility of misinterpretation.

Dynamic Contour Tonometry

Dynamic contour tonometry, first demonstrated by Kanngiesser et al. [114], uses a flexible material that can conform to the curvature of the cornea. When the device match's the curvature, a piezo-resistive pressure sensor can accurately measure the IOP. By taking a continuous pressure measurement against the eye, the ocular blood flow corresponding to the heart pulse can be estimated; this is evidence that this can also be a predictor of irregular IOP [106, 63]. An example of a dynamic contour tonometer is the PASCAL, a slit-lamp mounted device similar to the Goldmann applanation tonometer [101].

Rebound Tonometry

Rebound tonometers measure how far a tiny probe bounces as it dropped onto the cornea. The iCARE tonometer [49] uses a magnetic field to hold the probe in place and then drive it towards the eye in a controlled manner. Once the probe bounces off the cornea, the magnetic field measures the deceleration, which is lower at high IOPs. Unlike most of the aforementioned tonometers that require the application of a topical anesthetic, rebound

tonometers can be used as is.

3.1.2 Studies on Tonometry

A number of clinical studies have been conducted to both compare different forms of tonometry and identify factors that can affect IOP measurements. Posner and Inglima [198] compared the measurements from fixed-force applanation tonometry (Maklakov tonometer), fixed-area applanation tonometry (Goldmann applanation tonometer), and indentation tonometry (Schiøtz tonometer). Strong correlations were found between all three techniques, but Posner and Inglima found that the fixed-force applanation and indentation tonometry techniques overestimate low IOP measurements and underestimate high IOP measurements when compared to Goldmann applanation tonometry. Posner has also published an article that provides a more qualitative comparison between those three techniques [197], including the fact that Goldmann applanation tonometry is not portable and is more difficult to ues with children.

Perhaps the most famous study in regards to confounding factors of tonometry is the Rotterdam Study [259]. Wolfs et al. measured the IOP and central corneal thickness of 395 subjects in Rotterdam, Netherlands. With a linear regression, they found an increase of 0.19 mmHg for each 10 micrometer increase in the thickness of the cornea from the average thickness of 537 micrometers. Mansouri et al. [152] and others found that IOP has a natural diurnal fluctuation of 3-6 mmHg. Qureshi et al. [202] noted a correlation between days of the menstrual cycle and variations in intraocular pressure, although the variations alone were not significant enough to affect diagnoses. Regarding controllable effects on IOP, Pardianto [192] found that alcohol and marijuana can lower IOP, while caffeine can increase IOP. Schmidtmann et al. [219] even found that the playing of musical instruments with high intraoral resistance (particularly woodwind and brass instruments) can lead to drastic increases in IOP.

3.2 Data Collection

Figure 3.1: The proposed system emulates fixed-force applanation tonometry using the hardware adapter pictured above. The clear acrylic cylinder is allowed to move freely within the black casing so that its mass provides a constant force on the patient's eye.

The hardware adapter is shown in Figure 3.1 attached to an iPhone case. The most important part of the hardware is the clear acrylic cylinder inside the black casing. The acrylic cylinder is allowed to move freely within the casing, but has notches to ensure that it does not fall out of the adapter. The acrylic cylinder has a diameter of 8 mm and a height of 63 mm. The 8 mm diameter was chosen such that it would capture a fairly large circle from a low eye pressure without being too difficult to use on patients with small palpebral fissures. The height of 63 mm was chosen for two reasons: (1) If the applanation surface is placed too closely to the smartphone's camera, the resulting video becomes difficult to focus and the edges become blurry. (2) This combination of diameter, height, and material leads to a mass of 5.0 g, a mass for which the conversion from applanation surface diameter to IOP has already been clinically validated for human eyes [196, 262]. Although conversion tables for larger masses have been produced, studies have shown

that the weight of the tonometer induces an increased pressure due to the displacement of aqueous humor during applanation [262].

The black casing itself is designed such that the acrylic cylinder is optimally positioned in front of the smartphone's back-facing camera. Not only does this positioning include the alignment of the acrylic cylinder with the camera, but also the distance between the base of the acrylic cylinder and the camera. The black casing also blocks out ambient lighting to prevent any extraneous reflections from appearing in the acrylic cylinder.

To enhance the visibility of the acrylic cylinder in the camera, the edge of the cylinder's bottom surface is frosted. To emphasize the applanation surface in the camera, an external LED is mounted on the casing; in the future, the smartphone's flash could be redirected to the outside of the acrylic cylinder via a short fiber optic cable. When this lighting is reflected off of fluorescein dye, it shines as a bright yellowish green.

Before receiving the assessment, the patient assumes a supine position. The user conducting the test administers a topical anesthetic with fluorescein dye (Fluorescein sodium 0.25%/Proparacaine 0.5%) to the patient's eye. The user then holds the smartphone over the patient's eye such that only the weight of the acrylic cylinder is applied to it. This means that the smartphone should be as flat as possible (i.e.,parallel to the ground) and the user should not apply any extra force on the smartphone (i.e.,pressing down). The flatness of the smartphone is measured with the smartphone's accelerometer, operating as a sort of bubble level.

The weight of the acrylic cylinder creates an elliptical applanation surface with a yellowish green outline when the LED is shone on the patient's eye. The smartphone's camera records the applanation of the eye. The frames from the resulting video are then processed using computer vision to give a real-time estimate of the patient's IOP.

3.3 Algorithm

Figure 3.2 outlines the algorithm used to extract an IOP measurement from an RGB image. The overall goal of the algorithm is to detect two ellipses: the base of the clear

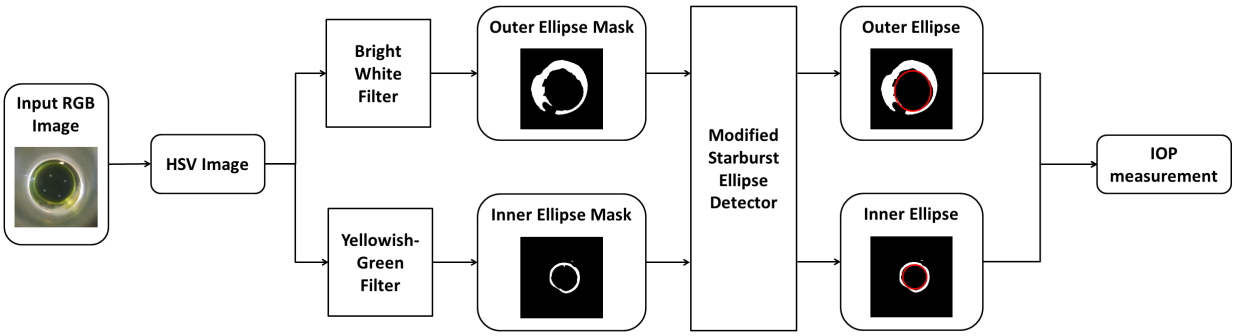

Figure 3.2: The steps taken to estimate intraocular pressure from an RGB image of the applanation procedure. After converting the image into the HSV space, masks are defined for the clear acrylic cylinder's base (outer ellipse) and the applanation surface (inner ellipse) using color and intensity features as filters. Ellipses are detected on the insides of those masks and then mapped to absolute measurements given the 8 mm diameter of the acrylic cylinder. The diameter of the applanation surface is then mapped to the patient's estimated IOP.

acrylic cylinder (outer ellipse) and the applanation surface (inner ellipse). Since the diameter of the acrylic cylinder is known, the applanation surface can be assigned an absolute measurement by using the cylinder as a reference. Both of the ellipses should be relatively circular; however, the acrylic cylinder may appear slightly elliptical if the hardware adapter is improperly mounted, and the applanation surface may be elliptical if the patient has significant astigmatism or corneal surface irregularities.

As shown on the far left of Figure 3.2, the edge of the acrylic cylinder appears bright and white, and the edge of the applanation surface appears as a dimmer yellowish green. By filtering the image according to intensity and color information, binary masks can be produced to select the outlines of the circles. This information is most intuitively recovered from the image after it is converted into the HSV space. The mask for the inner ellipse bounds the hue between 15-45%, the saturation between 35-100%, and the value 15-100%. Together, these thresholds encode the greenish yellow that appears due

to the fluorescein. The mask for the outer ellipse is simpler, thresholding the saturation between 0-20% and the value between 25-100%. Both masks are smoothed using morphological filtering operations to create contiguous contours.

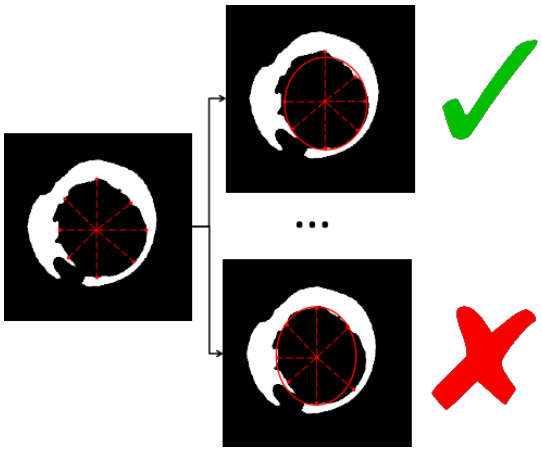

Figure 3.3: A variation of the Starburst technique by Li et al. [141] is used to estimate the innermost ellipse from a binary mask. After candidate points are selected from the inside, contiguous subsets of points are tested with least-squares ellipse fitting until the most circular is found.

Each of the masks will have some non-uniform thickness due to the application of the dye and extraneous reflections in the cylinder. The diameters of interest correspond to the innermost edges of these masks. Standard circle detection methods would either discover many overlapping circles or none at all, depending on the evenness of the masks. Even worse, only part of the applanation surface may be visible if it overlaps with the sclera, which makes it more difficult to see the fluorescein dye. For these reasons, I apply an adaptation of the pupil contour detection algorithm (Figure 3.3) used by Li et al. in their Starburst work [141].

The ellipse detection starts by assuming the center of the ellipses given that the position of the clear acrylic in the camera's view is known. The algorithm then steps radially at 20 evenly spaced angles until an edge is reached in the mask (illustrated with

fewer angles for clarity). This assumes that there are no contours that appear within the mask, which can happen for the applanation surface if the fluorescein pools in the patient's eye. Since extra blobs appear in the middle of the mask due to the distribution of the fluorescein, the radial steps start from a fixed distance just below the minimum expected radius to prevent them from stopping short.

Most of the detected edge points should belong to the desired ellipse, but some may still belong to artifacts along the edge of the contour. The original Starburst algorithm accounts for noisy ellipse points by fitting random subsets of points to ellipses and selecting the ellipse that minimizes the number of outliers. In the case of applanation, there is almost always a clean arc that appears in the image. Instead of randomized subsets of points, as used by Li et al., the proposed system fits contiguous subsets of points (three-quarters of the entire circumference) to ellipses. Although I noted earlier that the base of the acrylic cylinder and the applanation surface may appear elliptical, the ellipses should be relatively rounded. If the percent difference between an ellipse's major and minor axes is greater than 10%, it is automatically rejected. Amongst the rest of the ellipses produced by the different subsets of edge points, the ellipse that best fits the data according to Euclidean distance is selected.

The ellipses recovered from the two masks are then translated into circles with a radius equal to the average of the ellipses' axes. Given that the diameter of the clear acrylic cylinder is 8 mm, the absolute measurement of the applanation surface can be recovered by using the cylinder as a reference. Every time the user performs an applanation, a time series of diameter measurements is produced. The measurements of interest occur when the data is most stable since that is when the weight of the acrylic cylinder should be resting on the eye; therefore, the system combines diameter measurements by taking a mean over the measurements within a standard deviation of 0.25 mm over the course of 0.5 s. The final diameter measurement is mapped to an IOP value using a clinically validated lookup table, such as the one published by Adolph Posner [196].

3.4 Results

3.4.1 Data Collection

Given the invasive nature of contact applanation tonometry, a feasibility evaluation has been performed on two freshly enucleated *ex vivo* porcine eyes before deploying the system to living human patients. Although there is not a clinically validated table that maps applanation surface diameter to IOP for animal eyes, the Imbert-Fick law still applies. The porcine eyes were inserted into a clay mold such that the iris was horizontal to the ground, as if a patient were supine. The anterior chambers of the eyes were cannulated and the IOP was artificially varied by the height of a saline-filled reservoir. A topical anesthetic with a fluorescein dye (Fluorescein sodium 0.25%/Proparacaine 0.5%) was applied to the eyes to assist in imaging the applanation surface. Tonometry measurements were obtained three times at every 5 mmHg between 15 and 40 mmHg, leading to a total of 32 measurements. Below 15 mmHg, the applanation surface's edge begins to overlap with the edge of the acrylic cylinder, making it difficult to separate the two. The upper bound exceeds the limits of diagnostic significance for elevated IOP.

3.4.2 Comparison to the Imbert-Fick Law

Although other methods of tonometry are available as a point of comparison for validation, they all have two drawbacks for my validation: (1) they require manual inspection (e.g., Schiøtz or Maklakov tonometers) and/or (2) they are calibrated specifically for *in vivo* human eyes (e.g., Goldman applanation tonometers). Instead of a comparison to a possibly inaccurate ground truth, I validate my findings against what is expected from the Imbert-Fick law:

$$P = \frac{F}{A} = \frac{mg}{\frac{1}{4}\pi d^2} \tag{3.1}$$

where P is the intraocular pressure, m is the mass applied to the eye, g is the

acceleration due to gravity, and d is the diameter of the applanation surface. Although corrections have been proposed by ophthalmologists to account for properties like the coefficient of ocular rigidity and corneal curvature (e.g., [262]), these models break down for *ex vivo* eyes. The measurements from each of the eyes were independently fit to the Imbert-Fick law using non-linear fitting. The mass parameter was used as the unknown parameter; if the data were to follow the Imbert-Fick law without deviation, the parameter resulting in the best fit would be 5 g, the mass of the acrylic cylinder.

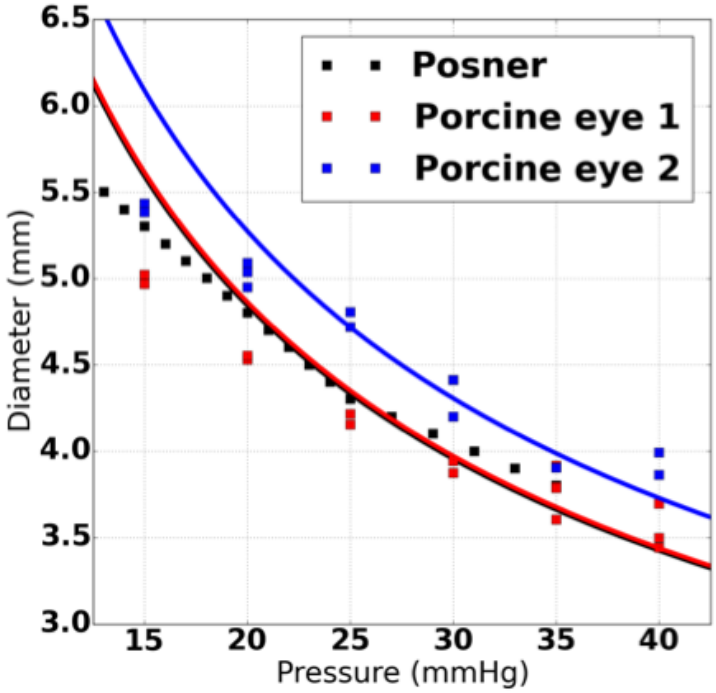

Figure 3.4: The data recorded from the smartphone system and fit to the physical model expected from the Imbert-Fick law. The two curves lead to coefficients of determination of 0.89 and 0.88.

Figure 3.4 shows the models that were fit to the two datasets. The clinical measurements validated by Adolph Posner [196] are also included as a point of reference. Even though Posner only dealt with human eyes, the shape of the data should

be similar to that of the ex vivo porcine eyes. When compared to the Imbert-Fick law, Posner's data shares a coefficient of determination (R^2) of 0.95 and the estimated mass according to the optimal fit is 5.01 g, showing that it follows the model very closely. The fit for the first porcine eye leads to a coefficient of determination of 0.89 and an estimated mass of 5.04 g. The fit for the second eye does not obey the Imbert-Fick law as well; it results in a lower coefficient of determination of 0.88 and an estimated mass of 5.94 g. The regressions overestimate low pressures and underestimates high pressures in all cases, a fact that has been observed by clinicians for other forms of tonometry as well [198]. The most promising observation is that there is a statistically significant separation between the diameters for 20 mmHg and 30 mmHg, the boundary that clinicians consider for the diagnosis of elevated IOP.

Chapter 4

BILISCREEN

Among all forms of cancer, pancreatic cancer has one of the worst survival rates [5]. Many attribute this statistic to the fact that the symptoms associated with pancreatic cancer often go unnoticed until the cancer is in a later stage; 80-85% of patients present themselves with tumors so advanced that they cannot be removed completely through surgery [23, 246]. One of the earliest symptoms to appear is jaundice, a yellow discoloration of the skin and eyes. In the case of pancreatic cancer, jaundice occurs because a cancerous growth obstructs the common bile duct, causing a buildup of bilirubin in the blood [55]. Being able to detect the very first signs of jaundice when levels of bilirubin are minimally elevated could enable an entirely new screening program for at-risk individuals. Jaundice also manifests as a symptom for a variety of other conditions, such as hepatitis and Gilbert's syndrome, but I am primarily motivated by the link between jaundice and pancreatic cancer for the purpose of this paper.

The clinical gold standard for measuring bilirubin is through a blood draw called a total serum bilirubin (TSB). TSBs are invasive, require access to a healthcare professional, and are inconvenient if done routinely for screening. Medical device manufacturers have investigated non-contact alternatives to a TSB for bilirubin. One such device is the transcutaneous bilirubinometer (TcB). A TcB shines a wavelength of light that is specifically reflected by bilirubin onto the skin and measures the intensity that is reflected back to the device. The computations underlying TcBs are designed for newborns; their results simply do not translate correctly for adults. Part of the reason for this is that normal concentrations of bilirubin are much lower in adults compared to newborns ($<$1.3 mg/dl vs. $<$15.0 mg/dl [20]). As it so happens, the sclerae are more

sensitive than the skin to changes in bilirubin because their elastin has a high affinity for bilirubin [148]. This presents an opportunity for early, non-invasive screening that has been previously unexplored. My contribution to this space is BiliScreen, a system that estimates the extent of jaundice in a person's eyes through pictures taken from the smartphone and produces an estimate of their bilirubin level.

To be effective, BiliScreen should be sensitive enough to measure the range of bilirubin levels exhibited by adults. Ruiz et al. [214] found that jaundice is not apparent to the trained naked eye until roughly 3.0 mg/dl; however, bilirubin levels greater than 1.3 mg/dl warrant clinical concern. There exists a detection gap between 1.3 and 3.0 mg/dl that is missed by clinicians unless a TSB is requested, which is rarely done without due cause. I hypothesize that diagnoses can be made much earlier and lead to better outcomes with a system that is precise enough to distinguish between bilirubin levels within and outside of those bounds.

Oftentimes, the trend of a person's bilirubin level over time is far more informative than just a single point measurement. If a person's bilirubin exceeds normal levels for one measurement but then returns to normal levels, it could be attributed to normal variation. If, however, a person's bilirubin shows an upward trend after it exceeds normal levels, it is more likely that a pathologic issue is worsening their condition, such as a cancerous obstruction around the common bile duct. Trends are not only important for diagnosis, but also for determining the effectiveness of treatment. One course of action for those affected by pancreatic cancer is the insertion of a stent in the common bile duct. The stent opens the duct so that compounds like bilirubin can be broken down again; a person's bilirubin level should decrease thereafter. If their bilirubin continues to rise, then there are either issues with the stent or the treatment is ineffective. Trends in bilirubin levels are difficult to capture because repeated blood draws can be uncomfortable and inconvenient for many people, especially those in an outpatient setting. BiliScreen takes advantage of the ubiquity of smartphones, dramatically reducing the effort required to perform these measurements.

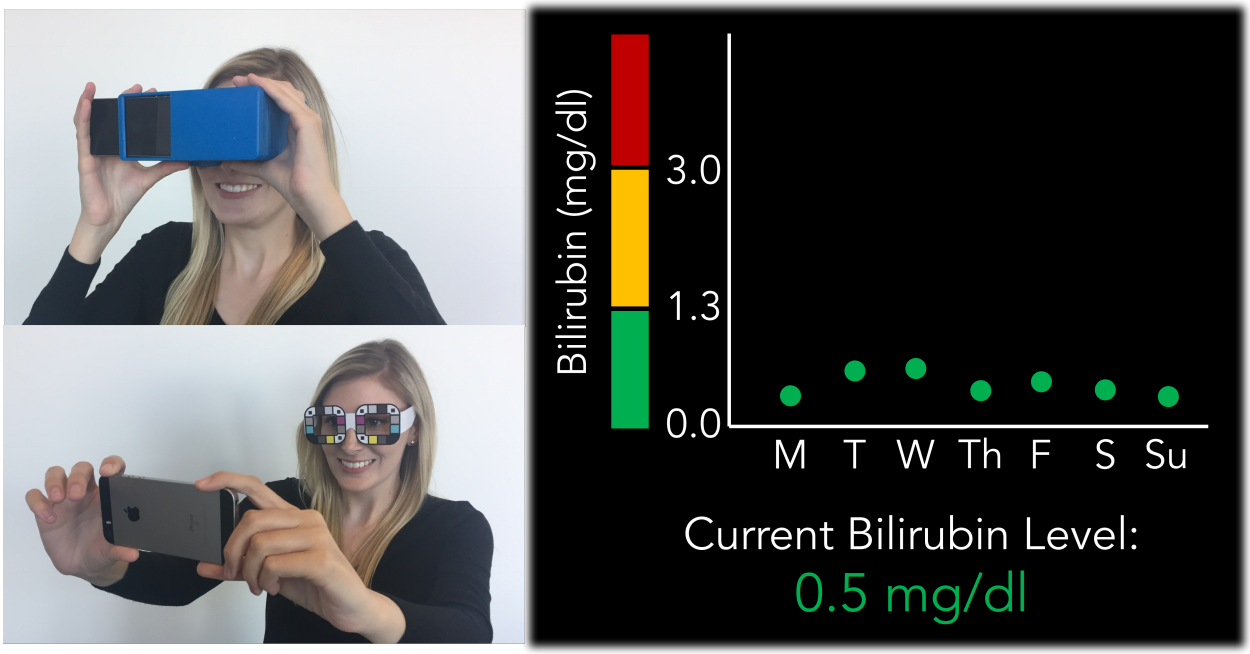

Figure 4.1: BiliScreen is a system that measures a person's bilirubin level using the smartphone's camera. I examine two methods for color normalization: **(top-left)** a box similar to a head-mounted VR display that controls the amount of light that reaches the eyes, and **(bottom-left)** paper glasses that provide colored squares for calibration.

BiliScreen uses the smartphone's built-in camera to collect pictures of a person's eyes. The sclera, or white part of the eyes, are extracted from the image using computer vision. Features describing the color of the sclera are then produced and analyzed by a regression model to return a bilirubin estimate. Since different lighting conditions can change the colors of the same scene, I evaluate two accessories that account for the ambient lighting conditions. The first accessory is a head-worn box (Figure 4.1, top-left), similar to a head-mounted VR display, that simultaneously blocks out ambient lighting and provides controlled internal lighting through the camera's flash. The second accessory is a pair of paper glasses printed with colored squares that facilitate calibration (Figure 4.1, bottom-left). The latter accessory is reminiscent of a previous project called BiliCam [91] by some of the co-authors of this work, which uses a color-calibration card

to account for ambient lighting conditions in pictures of newborns that are processed to detect neonatal jaundice. Beyond their intent of assessing bilirubin levels by detecting jaundice through the smartphone camera, the two projects are quite different. BiliCam is intended for newborns, who exhibit a far wider range of normal bilirubin levels than adults. Because the sclera does not have a predefined shape, BiliScreen also requires an additional step of segmentation. Although BiliScreen has tighter precision requirements, it benefits from the fact that the typical sclera is race-agnostic; the same cannot be said for skin, which varies across different ethnicities.

I evaluated BiliScreen in a 70-person preliminary study including individuals with normal, borderline, and elevated bilirubin levels. I found that BiliScreen with the box accessory, which leads to better results than the glasses, estimates an individual's bilirubin level with a Pearson correlation coefficient of 0.89 and a mean error of -0.09 $\pm$ 2.76 mg/dl when compared to a TSB. BiliScreen with the glasses accessory leads to a Pearson correlation coefficient of 0.78 and a mean error of 0.15 $\pm$ 3.55 mg/dl.

My contribution comes in four parts:

1. An implementation of the BiliScreen system for convenient bilirubin testing with two different methods for color calibration,

2. A novel sclera segmentation algorithm that is robust for individuals with jaundice,

3. Models that relate the color of the sclera to a measure of bilirubin in the blood, and

4. An evaluation of BiliScreen on 70 participants.

4.1 Related Work

The BiliScreen algorithm has two fundamental components: automatic segmentation of the sclera and models that map sclera color to bilirubin level. I summarize the literature related to both components below.

4.1.1 Sclera Imaging

To my knowledge, BiliScreen is the first application that automatically segments the sclera for medical purposes. There is, however, a body of literature that has proposed various methods of segmenting the sclera for biometric verification and gaze estimation. For biometrics, individuals are recognized through the uniqueness of the blood vessel patterns in their sclera. For gaze estimation, researchers have relied on the fact that the exposed area of sclera changes as a person makes significant changes in gaze.

The most common method for sclera segmentation relies strictly on color information, noting that the sclera is normally white. Zhou et al. [267] use dynamic thresholds in the RGB and HSV color spaces to create binary masks that correspond to non-skin- and sclera-colored pixels, respectively. After taking the intersection of those masks, the iris and pupil are removed by using a visible glint within the iris as a seed for an iterative method that moves radially until it reaches the iris-sclera border. Marcon et al. [153] train a linear discriminant analysis classifier on pixel color values to distinguish between sclera and non-sclera pixels. Morphological operations and watershed flooding are applied to form fuller candidate regions for the sclera, after which a classifier trained on shape information is used to select the regions that most resemble the sclera. Das et al. [50] propose a method that involves fuzzy k-means clustering on the pixel color and location to form three clusters: the skin, iris, and sclera. These strictly color-based methods rely on the assumption that the sclera is bright and white, which is not the case for people with jaundice. As the sclera becomes more yellow, its color can be confused with the color of lighter skin tones, making it difficult to train a global classifier. Even if the person's skin tone is known beforehand, there is the chance that its color is too similar to the person's sclera for it to be removed without spatial information.

In a different paper, Das et al. [51] demonstrate a method of sclera segmentation that uses active contour-based segmentation. In active contour-based segmentation, a snake

(i.e.,deformable spline) is initialized roughly around an object of interest. An energy function is defined based on the presence of lines, edges, and corners in the image, and the position of the snake is iteratively adjusted until that energy function is minimized. For sclera segmentation, Das et al. initialize snakes to the left and right of the automatically detected pupil. This technique is suitable for BiliScreen in that does not depend on the color of the sclera, but the initialization of the snakes can be difficult when the geometry of the eye is not completely constrained. The location of the sclera relative to the pupil depends on both the geometry of the eye and the user's gaze direction. For instance, depending on the narrowness of the eye and how far the user looks up, the sclera may or may not appear directly under the iris. If the initial snakes are too far out from the sclera, they may stop short at glare spots or wrinkles near the eyelids as they constrict. More onus could be placed on the person whose picture is being taken to adjust themselves until their pose satisfies specific constraints, but such a procedure could lead to frustration. Instead of relying on the location of the pupil, eye detection algorithms [142, 247] could be used to standardize a region of interest around the eye; however, such techniques fail when nearby facial features are obstructed, as is the case with the BiliScreen accessories.

One more approach that has been explored for sclera segmentation is the use of dedicated hardware. Crihalmeanu and Ross [46] utilize near-infrared (NIR) lighting to make sclera segmentation straightforward. They observe that the skin has higher NIR reflectance than the sclera since the skin has less water, which makes the separation between the sclera from pale skin more apparent in NIR than in RGB. The use of dedicated hardware in BiliScreen beyond my box or glasses accessory is undesirable for cost and accessibility purposes.

Overall, these issues motivate the need for a more automated solution. The sclera segmentation approach I propose for BiliScreen uses two iterations of the GrabCut method [212]. The first iteration learns the color characteristics of the skin and removes the skin to isolate the eye. The second iteration isolates the sclerae by assuming that they

are the brightest regions within the eyes (not necessarily white).

4.1.2 Jaundice Assessment

The standard for measuring bilirubin in the blood is through a blood draw called a total serum bilirubin (TSB). The more convenient alternative used in neonatal clinics is a transcutaneous bilirubinometer (TcB). Beyond these two methods, there are several researchers who have investigated bilirubin measurement via the digital photography of areas susceptible to jaundice: the skin and eyes.

Leartveravat [136] proposes a completely manual system for assessing jaundice in a newborn's skin. Photographs of the skin with a color calibration card are captured using a digital camera. Once the photo is uploaded to image editing software (e.g.,Adobe Photoshop), the image is color-calibrated and converted to the CYMK color space. A technician then manually selects a pixel representative of the newborn's skin, subtracts its yellow component from its magenta component, and inputs that value into a linear regression to get a bilirubin estimate. The BiliCam system by de Greef et al. [91] also analyzes pictures of a newborn's skin with a color calibration card to estimate their bilirubin level. It differs from the work of Leartveravat in that BiliCam entails more complicated models that account for skin tone.

Leung et al. [139] compare the performance that a system could achieve by analyzing both the skin and the sclera for newborns. Similar to de Greef et al. and Leartveravat, the authors manually selected regions corresponding to the skin, sclera, and a color calibration card for their analyses. With a fairly modest linear regression model, the authors achieve far better Pearson and Spearman correlations using the sclera (0.75 and 0.72) than using the skin (0.56 and 0.54).

To the best of my knowledge, BiliScreen is the first non-invasive system to quantify an adult's bilirubin level. BiliScreen analyzes the sclera because, as Leung et al. confirmed, the sclera is more sensitive to changes in bilirubin than the skin. This is important because

higher precision is needed for adults. Bilirubin levels in healthy newborns may peak as high as 15.0 mg/dl [20], whereas bilirubin levels for healthy adults are normally less than 1.3 mg/dl. BiliScreen is also completely automated, from the segmentation of the glasses and sclera to the feature extraction and machine learning. Finally, BiliScreen benefits from the fact that healthy sclera colors are independent of ethnicity, so less training data should be needed in the long-term.

4.2 Data Collection

I collected images using the BiliScreen app with both the box and glasses accessories to train BiliScreen's models and evaluate their efficacy. Volunteers with normal bilirubin levels were recruited from the University of Washington. Volunteers with varying bilirubin levels (ranging from normal to elevated) were recruited from the University of Washington Medical Center. Below, I elaborate on the diversity of the participant pool. I then describe my data collection procedure, including the design of the BiliScreen accessories and my procedure for ground truth measurements. All facets of my study were approved by the University of Washington's Institutional Review Board.

4.2.1 Enrollment

Table 4.1: Participant demographics (N = 70)

BILIRUBIN CLASSIFICATIONS - N (mean $\pm$ std)	
Normal (<1.3 mg/dl)	31 (0.6 $\pm$ 0.2 mg/dl)
Borderline (1.3-3.0 mg/dl)	14 (2.1 $\pm$ 0.5 mg/dl)
Elevated (>3.0 mg/dl)	25 (9.7 $\pm$ 5.9 mg/dl)

My study included 70 volunteers. From the university, 18 were male and 13 were female. From the medical center, 13 were male and 26 were female. Table 4.1 shows the

distribution of the total serum bilirubin tests split across the two different populations. Note that the precision of the TSB is 0.1 mg/dl.

Thresholds classifying the concern warranted by a single bilirubin measurement can vary between clinics. For the purposes of BiliScreen, three classes are defined: normal (<1.3 mg/dl), borderline (1.3-3.0 mg/dl), and elevated (>3.0 mg/dl). The 1.3 mg/dl threshold is used by the University of Washington Medical Center as their upper limit for a normal TSB measurement, while the 3.0 mg/dl threshold is based on the findings of Ruiz et al. [214] concerning when jaundice is most apparent to clinicians. According to these thresholds, 31 participants had a normal bilirubin level, 14 had a borderline bilirubin level, and 25 had an elevated bilirubin level. Unsurprisingly, most of the university population had a normal bilirubin level. The lack of variation within that population was expected. Although the clinical upper threshold for normal bilirubin levels is 1.3 mg/dl, values near 0.6 mg/dl are the norm. The medical center population provided a much wider spread of bilirubin levels, ranging from normal to elevated.

4.2.2 Data Collection Procedure

The data collection procedure for the BiliScreen app was the same for both populations, but the methods of recruitment and collection of ground truth measurements were different. The participants from the university were volunteers recruited from emails on public mailing lists. After a research staff member collected data with the BiliScreen app (described in the next section), they were asked to undergo a TSB within 24 hours. Bilirubin can change over long periods of time but remains stable within a day barring any serious conditions.

The participants from the medical center were inpatients suffering from liver disease. A research staff member selected candidate participants on two criteria. The first criterion was a recorded TSB blood test within 24 hours. Again, this is to ensure that the patient's recorded bilirubin level matches closely with their level at the time of data collection. The

second criterion relies on the Model for End-stage Liver Disease (MELD) [256], a scoring system for assessing the severity of chronic liver disease. The MELD score is a summary metric that combines three measures of a patient's liver condition - TSB, serum creatinine, and the international normalized ratio for prothrombin time (INR) - with a higher score indicating a higher three-month mortality rate. There is no guarantee that a patient with a high MELD score has an elevated bilirubin level since TSB is only one component of the MELD score; however, a high MELD usually includes an elevated TSB. The original recruitment criteria was a minimum MELD score of 14. This threshold was later lowered to 6 in order to recruit more patients with borderline levels (1.3-3.0 mg/dl) of bilirubin. If a patient satisfied the two recruitment criteria, they were approached by a research staff member and told about the study. Patients were enrolled in the study if and only if they understood the study and gave consent.

4.2.3 BiliScreen Accessories

Physics-based models for color information typically consider an object's visible color to be the combination of two components: a body reflection component, which describes the object's color, and a surface reflection component, which describes the incident illuminant [126]. When using digital photography, color information that gets stored in image files is also impacted by the camera sensor's response to different wavelengths. For my study, I examine the efficacy of two different accessories to isolate the sclera's body reflection component in different ways (Figure 4.2).

The first accessory is a 3D-printed box reminiscent of a Google Cardboard headset[1]. There is no electrical connection between the phone and the box; the phone is simply slid into the box via a rectangular channel along the back. The channel at the back of the box also fixes the placement of the phone relative to the participant's face by centering the phone's camera and keeping it at a fixed distance. The box blocks out ambient lighting

[1] https://vr.google.com/cardboard/

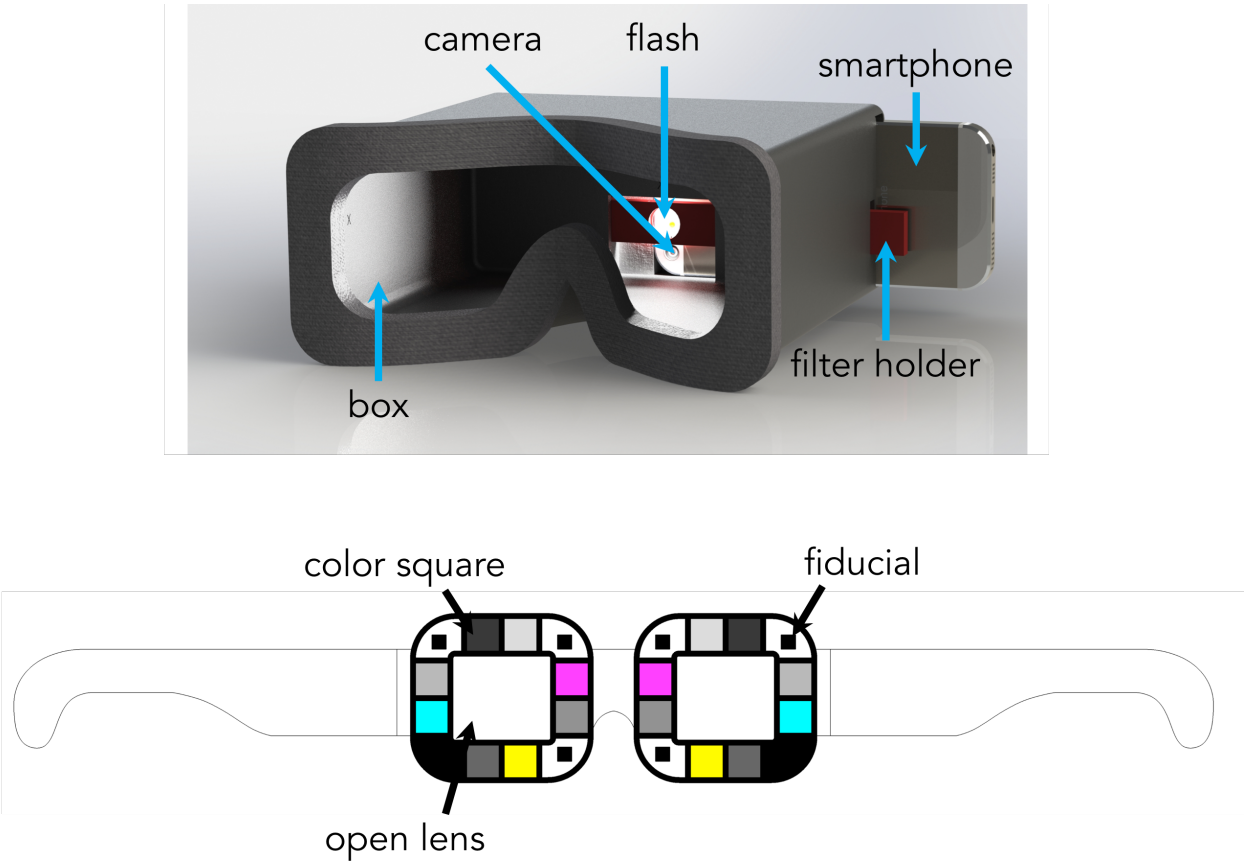

Figure 4.2: **(top)** A 3D rendering of the BiliScreen box. The smartphone's flash lies in the horizontal center of the box. The flash is covered with a neutral density filter and a diffuser to make the light more comfortable. **(bottom)** A rendering of the BiliScreen glasses.

while allowing the phone's flash to provide the only illumination onto the eyes. From pilot studies, some participants found the flash to be overwhelmingly bright. A neutral density filter and a diffuser were placed in front of the flash using a filter holder to soften the light slightly. The box used in my study was 3D-printed, but it could be made with an even cheaper material like cardboard (provided that it is sturdy enough to support the weight of the phone). By using the flash as the only illumination source on the sclera, the surface reflection component is kept constant for all images. This leaves the body

reflection component and the camera sensor's response as the only two components that affect the sclera's appearance. For the sake of this study, all images were captured using the same device, holding the camera sensor's response constant and leaving the body reflection component as the only variable left.

The second accessory (Figure 4.2, bottom) is a custom pair of paper glasses, reminiscent of the 3D glasses found at movie theaters. The glasses have no lenses inside their frames. Along the rims of the glasses are various colored regions. The corners near the temples and the nose have smaller black squares surrounded by the glasses' white background. These squares act as fiducials, similar to those seen in QR codes. The rest of the regions along the rims are the following colors (in no particular order): cyan, magenta, yellow, 17% gray, 33% gray, 50% gray, 67% gray, 83% gray, and black. The use of the colored squares is inspired by color calibration target cards like the Macbeth ColorChecker [193]. Rather than keeping the surface reflection component and the camera sensor's response constant, the colored squares allow for all images to be normalized to the same references. The colors along the rims of the glasses are known *a priori*. This means that their body reflection component is known and any deviation between their appearance and their true color is due to the surface reflection component and the camera sensor's response. Section 4.3.3 explains the calibration procedure that is used to define a calibration matrix that best simulates the effects of the latter two color information components, which can later be applied to the sclerae themselves to reveal their true body reflection components.

From a usability perspective, the glasses are more convenient for the user and cheaper to manufacture. However, the colors along the rims of the glasses must always be consistent, both across time and different pairs. If the colors were to fade over time, the colors would become a changing reference that could lead to inaccurate results. Although the box is bulkier, its requirements are far looser. The box's main purpose is to block out ambient lighting; control over the precise placement of the smartphone is convenient for aspects of the automatic segmentation, but the box's dimensions do not

require as strict precision as the glasses' colors.

From a technical perspective, the color calibration procedure for the glasses can incur its own inaccuracies. In BiliScreen's current state, though, the algorithm for the box accessory does not account for the camera sensor's response. If users were to use a phone with a camera different from that of the iPhone SE, no guarantee can be made that colors will appear the same between the two. Even though the color calibration procedure for the glasses may introduce noise, it allows for any device to be used without issue. The calibration procedure captures the effects of both the surface reflection component and the camera sensor's response.

4.2.4 BiliScreen Application

All data was collected by a research staff member through a custom app on an iPhone SE. The images collected by the app were at a resolution of 1920×1080. The research staff member ensured that participants complied with the procedure and noted any difficulties that participants had with the app and its accessories.

The BiliScreen app developed for my study was designed to collect data for both accessories in a similar manner. Before the use of either accessory, the smartphone's flash was turned on. When using the box, the flash is necessary since it is the only way to make the eyes visible within it. Keeping the flash constantly on rather than bursting it at the time of the pictures was a consideration for participant comfort since the stark change in lighting can be unpleasant. When using the glasses, the flash was left on in case there was insufficient lighting in the room or the glasses created a shadow on the participant's face.

After the flash was turned on, the research staff member placed the smartphone in the BiliScreen box. A hole in the back of the box provided access to the screen for starting and stopping data collection. The app prompted the participant to look in four different directions—up, left, right, and straight ahead—one at a time while taking a picture after

each. Having the participant look in different directions exposed different parts of the sclera, some of which may have exhibited more jaundice than others. The participant was not asked to look downward since doing so covers their eyes with their eyelids. Once the pictures were taken inside the box, the research staff member removed the smartphone and held it approximately 0.5 m away from the participant's face to take pictures with the glasses. This distance is roughly how far away I would expect participants to hold their smartphones if they were taking a selfie. The participant looked at each direction for two trials per accessory, yielding 2 BiliScreen accessories $\times$ 2 trials per accessory $\times$ 4 gaze directions per trial = 16 images per participant.

4.3 Algorithm

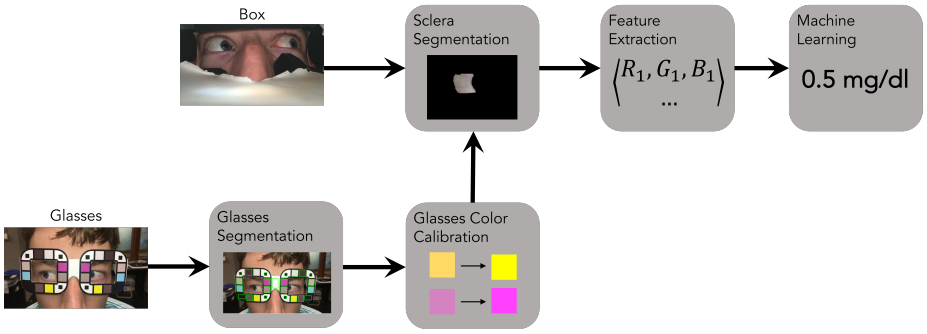

Figure 4.3: The algorithm pipeline for both BiliScreen accessories. Images from both the box and the glasses go through the same sclera segmentation, feature extraction, and machine learning steps (with their own respective models and small parameter changes). Images gathered with the glasses must go through the extra steps of glasses segmentation and color calibration.

Figure 4.3 outlines the high-level algorithm pipeline that transforms a BiliScreen image to a bilirubin estimate. I will provide further detail in this section on each of these steps, starting with the segmentation of various regions of interest, the transformation of those regions into feature vectors, and finally the machine learning itself.

4.3.1 *Sclera Segmentation*

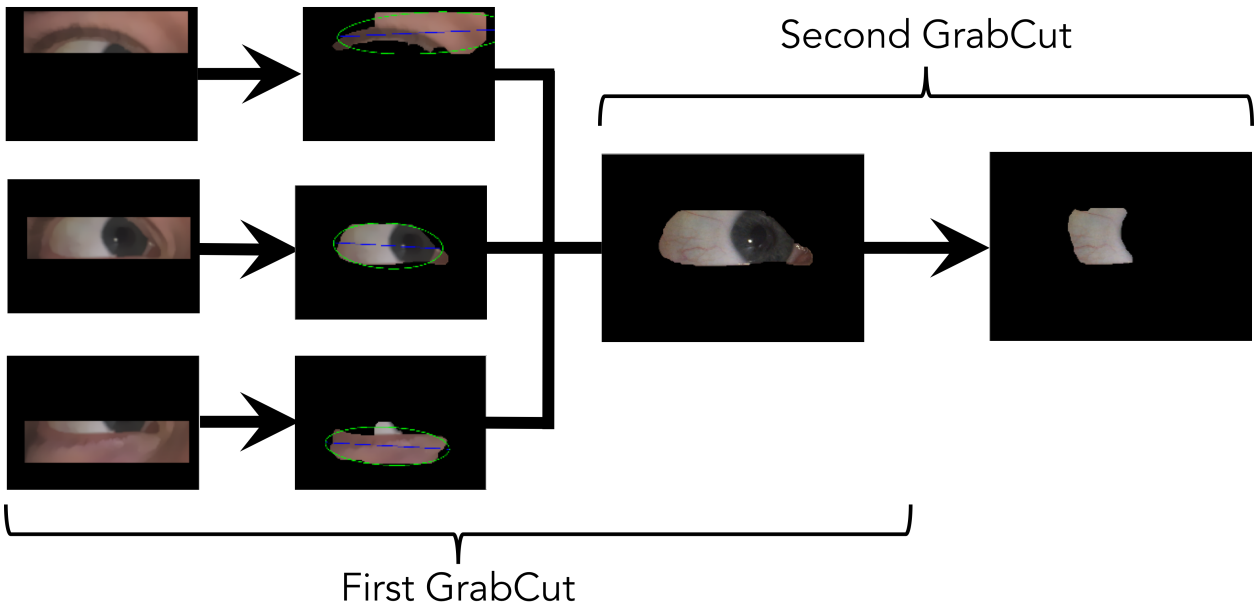

Figure 4.4: The procedure for sclera segmentation. The first iteration of GrabCut is initialized with several translated rectangles in parallel. The one that leaves a region that most resembles the eye is used as the region of interest for the second iteration of GrabCut. The second iteration of GrabCut uses adaptive thresholds to select the brightest regions within the eye.

The first step to segmenting the sclera from BiliScreen images is to define regions of interest where the sclera should be located. One way to logically identify these regions would be to use Haar feature-based cascade classifiers [142, 247] that are used in many applications that require eye detection. However, off-the-shelf eye detectors sometimes failed because features around the eyes (e.g., eyebrows) were obstructed by the BiliScreen box and glasses. To maintain consistency across images, regions of interest are defined through other methods depending on the BiliScreen accessory in use. Within the BiliScreen box, the regions of interest are defined as rectangular bounding boxes located on the left and right half side of the box using predetermined pixel offsets within the

image. This is possible because the placement of the camera within the box is always the same. The offsets were defined such that the regions of interest would cover various face placements and inter-pupillary distances. For the BiliScreen glasses, the regions of interest are more precisely defined as the regions surrounded by the colored squares (refer to Section 4.3.2 for how those squares are identified).

My approach to sclera segmentation relies on an algorithm called GrabCut [212], a technique for separating a foreground object from its background; in the case of BiliScreen, the sclera is the foreground, and everything else (e.g., skin, iris, pupil, hair) is the background; the terms "foreground" and "background" do not necessarily refer to the perceivable foreground and background of the image, but rather a region of interest versus everything else in the image. GrabCut treats the pixels of an image as nodes in a graph. The nodes are connected by edges that are weighted according to the pixels' spatial and chromatic similarity. Nodes in the graph are assigned one of four labels: definitely foreground, definitely background, possibly foreground, and possibly background. After initialization, graph cuts [25, 93] are applied to re-assign node labels such that the energy of the graph is minimized. Normally, GrabCut is an interactive technique that is typically initialized with a bounding rectangle and then followed with user-drawn strokes that further clarify the object of interest. BiliScreen uses GrabCut with a similar procedure, but without human intervention.

Before segmentation, bilateral filtering is applied to smooth local noise while maintaining strong edges. For the first iteration of segmentation, the eye is extracted using GrabCut with rectangles for initialization (Figure 4.4, left). This not only limits the search space for the sclera, but also removes most of the skin around the eye, reducing any effects those pixels could have on color histograms or adaptive thresholds later in the algorithm. The location of the eyes within the image can vary, so rectangular initializations at different locations are tested. To determine which output is most likely to only contain the eye, the segmented regions from each initialization are described using the calculations listed in Table 4.2. As the second column indicates, some of the

metrics are meant to be minimized, while others are meant to be maximized. Those that are meant to be minimized are negated so that higher values always imply that the region is more eye-like. The metrics are combined using the Mahalanobis distance relative to all of the other segmented regions. Overall, this calculation results in high distances for segmented regions that are small, elliptical, flat, and diverse in color, as well as rectangular initializations that likely do not crop out the eye. The segmented region with the highest distance wins out and is passed along to the second part of the sclera segmentation algorithm.

Table 4.2: Metrics used to rate a result of GrabCut as an eye

Name	**Min/Max**	**Description**
Area fraction	Min	The fraction of the region's area over the total area of the region of interest
Ellipse area fraction	Max	The fraction of the region's area over the area of the ellipse that best fits the region
Incline	Min	The incline of the ellipse that best fits the region
Color variation	Max	The standard deviation of the color across the region
Variation over borders	Min	The standard deviation of the brightness values across the top and bottom borders of the rectangle used to initialize GrabCut

After the first iteration of GrabCut is applied, the pixels that are assigned to the foreground are considered to be part of the eye, regardless of whether they are labeled as "definitely foreground" or "possibly foreground". A second iteration of GrabCut is then used to extract the sclera from the eye (Figure 4.4, right). The second iteration of GrabCut normally requires user interaction. In BiliScreen, however, the GrabCut

initialization can be bootstrapped automatically using adaptive and pre-defined thresholds. After converting the image to the HSL color space, the four possible pixel assignments are initialized as follows:

- Definitely foreground: Top 90^{th}-percentile of L channel values
- Definitely background: Bottom 50^{th}-percentile of L channel values
- Possibly foreground: Otsu threshold [185] on L channel values
- Possibly background: Inverse Otsu threhsold on L channel values

In cases when a pixel satisfies multiple assignments, the strongest assertion is prioritized (i.e., definitely foreground over possibly foreground). These assignments are based on the assumption that the brightest region in the eye should be the sclera. This assumption fails when glare appears within the eye, which is always the case with the BiliScreen box and sometimes the case with the BiliScreen glasses. Glare corresponds to high values in the lightness channel of the HSL image ($L > 230$). Pixels with glare are replaced using inpainting, a reconstruction process that re-evaluates those pixels' values via the interpolation of nearby pixels. Once GrabCut is run for the second time, the pixels that belong to the "definitely foreground" and "possibly foreground" labels are selected. The resulting mask is then cleaned by a morphological close operation to remove any tiny regions.

The distance between the smartphone and the person's face changes depending on which BiliScreen accessory is in use while the picture is being taken. The smartphone is at a fixed distance of 13.5 cm from the person's face when the BiliScreen box is in use and at a variable, farther distance when the BiliScreen glasses are in use. Changes in distance can have a modest effect on the lighting because the flash imparts more light on the eye when it is closer to the face. However, this effect is constant with the BiliScreen box and is canceled out by the color calibration procedure for the BiliScreen glasses. The distance

does, however, have a greater effect on the parameters for segmentation. As the distance between the smartphone and the person's face increases, the effective size of the eye in the image shrinks. The size of the rectangle used to initialize the first iteration of GrabCut has fixed dimensions for the BiliScreen box (∼600 × 200 px) and dynamic dimensions according to the size of the frames for the BiliScreen glasses (∼90% of width × 60% of height).

4.3.2 Glasses Segmentation

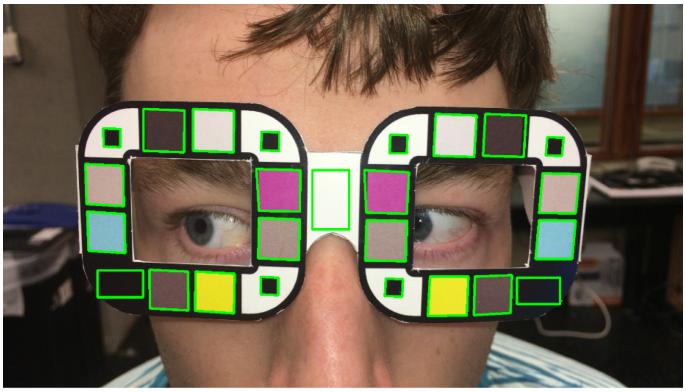

Figure 4.5: An example of correct segmentation for the glasses. The region over the bridge of the nose is used as a white reference for both sides.

The goal of the glasses segmentation is to identify the borders of the colored squares around the rims of the glasses and the white portion at the bridge of the nose so that their colors can be used for calibration. An example of correct segmentation is provided in Figure 4.5. The process starts with identifying the fiducials at the corners of the glasses. The fiducials are designed to be square-shaped, but unless they are viewed straight on, they can appear more as quadrilaterals. Black quadrilaterals are found by converting the image to grayscale and filtering it so that only the contours with four corners and a brightness value less than 60 are kept. The small quadrilaterals correspond

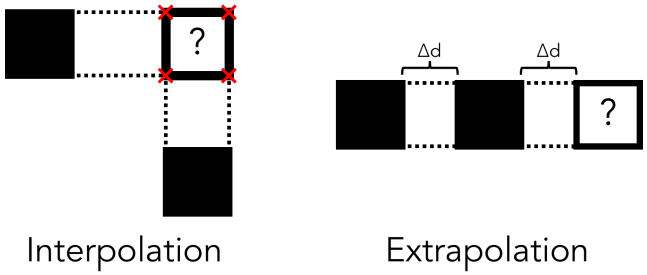

Figure 4.6: Illustrations showing how the positions of known fiducials or colored squares can be used to **(left)** interpolate or **(right)** extrapolate the positions of missing ones.

to the fiducials, while the others correspond to the outlines of the colored squares around the rims. The fiducials are roughly one-fourth the size of the colored squares. Therefore, quadrilaterals that are less than half of the average quadrilateral area are classified as fiducials; the other quadrilaterals are classified as colored squares. To confirm that the fiducials belong to the glasses and not something in the background, the algorithm checks that the pixels immediately outside of their borders are white. If any fiducials are not found because of glare or some other error, their locations are interpolated or extrapolated based on the locations of the discovered fiducials and the known geometry of the glasses. The left side of Figure 4.6 shows an example of interpolation. When there are known fiducials that are along the same vertical and horizontal axes as where the missing fiducial should be, the corners of the missing fiducial can be estimated by using the intersections of those lines. The right side of Figure 4.6 shows an example of extrapolation. If there are not enough known fiducials to use interpolation, BiliScren relies on the known relative dimensions of the glasses to estimate where they would most likely lie.

The positions of the fiducials are then used to check the positions of the colored squares. The fiducials are connected with straight lines to provide guides on which the other squares should lie. Any quadrilaterals found outside of those bounds are

discarded as the background. The fiducials are then used to develop a one-to-one mapping between the names of the colored squares (e.g., left yellow, right 33% gray) and their locations in the image. In the end, there should be two colored squares along each side of the lenses and black patches at the far bottom corners. The locations of the larger black-bordered quadrilaterals are compared to the expected positions of the colored squares. If the distance between a detected quadrilateral and the expected position of a colored square is less than a quarter of the expected square's width, the quadrilateral is matched with the corresponding label. There may not be enough detected black-bordered quadrilaterals to assign a border to every square label. This can be attributed to, among other reasons, glare from the camera or ambient lighting that obscures black outlines. Like the missing fiducials, the missing colored squares can be found using a combination of interpolation and extrapolation. After the squares around the rims of the glasses are found, the white rectangle that rests on top of the bridge of the nose is selected using a specified offset from the rims to provide a white color reference.

Both interpolation and extrapolation in this algorithm assume that the squares are linearly arranged around the glasses. The glasses were designed to make interpolation and extrapolation straightforward, but there were cases when users had to bend them so that they would fit comfortably on their faces. In these cases, it can be difficult to find fiducials and colored squares when quadrilateral detection has already failed. That being said, the advantage of the BiliScreen glasses design is that there are squares with the same color on each side. It is preferable to detect the squares on the same side as the eye of interest since they better represent the lighting shone on that particular side, but if one of those squares cannot be found, the other side can be used as a contingency.

4.3.3 Glasses Color Calibration

By identifying the colored squares of the glasses, BiliScreen images can be normalized to a common reference. Doing so removes the effects of the ambient lighting and the camera

sensor's response, both of which can change the appearance of the sclera.

The calibration procedure involves identifying the calibration matrix C that best maps the colors of the glasses' squares observed in the image to their actual colors. More formally, define O as the matrix of observed colors and T as the matrix of target colors, where each row contains an RGB vector that corresponds to a colored square. The matrix C defines the linear transform such that:

$$\begin{bmatrix} T_{R1} & T_{G1} & T_{B1} \\ T_{R2} & T_{G2} & T_{B2} \\ \vdots & \vdots & \vdots \\ T_{Rk} & T_{Gk} & T_{Bk} \end{bmatrix} = \begin{bmatrix} O_{R1} & O_{G1} & O_{B1} \\ O_{R2} & O_{G2} & O_{B2} \\ \vdots & \vdots & \vdots \\ O_{Rk} & O_{Gk} & O_{Bk} \end{bmatrix} \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix} \tag{4.1}$$

Because image files are gamma-encoded to optimize the usage of bits, gamma correction must be applied to the observed colors from the image so that linear operations on them are also linear. This is done by raising the values in O by a constant (γ = 2.2 for standard RGB image files). After a calibration matrix is applied, the gamma correction can be reversed by raising the values of the matrix to $1/\gamma$.

The calibration matrix C is calculated using an iterative least-squares approach detailed by Wolf [258]. The calibration matrix is first initialized under the assumption that the individual color channels are uncorrelated and only require a gain adjustment that would scale the mean value of the observed channel values to their targets:

$$C = \begin{bmatrix} \text{mean}(T_{Ri})/\text{mean}(O_{Ri}) & 0 & 0 \\ 0 & \text{mean}(T_{Gi})/\text{mean}(O_{Gi}) & 0 \\ 0 & 0 & \text{mean}(T_{Bi})/\text{mean}(O_{Bi}) \end{bmatrix} \tag{4.2}$$

For each iteration, the current calibration matrix is applied to the observed colors to produce calibrated colors. The colors represented by the rows are converted to the CIELAB color space so that they can be compared to the targets in T using the CIEDE2000 color error [221], the current standard for quantifying color difference. A

new calibration matrix C is computed that reduces the sum of squared errors, and the process repeats until convergence.

For BiliScreen, the rows of the target color matrix T are defined as the expected RGB color vectors of the glasses' squares according to their specification. The rows of the observed color matrix O are computed by finding the median vector in the HSL color space of the pixels within the bounds of the squares found in Section 4.3.2 (excluding the fiducials) and converting the vector back to RGB. For a region R with N 3-dimensional colors, the median vector is defined as:

$$v_m = \underset{v_i \in R}{\operatorname{argmin}} \sum_{j=1}^{N} \|v_i - v_j\|_2 \tag{4.3}$$

The median vector is preferred over taking the mean or median across the channels independently because it guarantees that the result is a color that exists within the original image; by treating the channels independently, the combination of values in the three channels may not ever appear in the image. The difference between the two approaches is typically insignificant when the region is uniform (as is the case with the colored squares), but is a precaution taken nonetheless.

The calibration procedure is repeated for both eyes using the colored squares closest to them. This is done because the ambient lighting effects are not always uniform; there may be a shadow or beam of light that creates a gradient across the face, making one side look slightly different from the other. There can also be cases when a colored square is washed out by glare from the smartphone's flash. If the error between the colored square and the expected color is 5 units more than the error between the corresponding colored square on the opposite side and the expected color (color difference is unitless), the latter is used. I never encountered a case when squares of the same color on opposite sides of the face were simultaneously obstructed. If that were the case, however, that color could simply be thrown out of the calibration procedure.

4.3.4 Feature Extraction

Jaundice is characterized by yellow discoloration, so the features extracted from BiliScreen's images should summarize the color of pixels belonging to the sclera. The color of the sclera is described using the median vector over the pixels that belong to the sclera for the same reasons described at the end of Section 4.3.3. More often than not, the sclera contains other components like vessels or a gradient from the eye's curvature. In these cases, aggregating the color channels independently can lead to a color that is not present in the sclera. For example, if an otherwise pristine sclera contains many blood vessels, taking the mean of the color channels independently will represent the color of the sclera as a pinkish color; the median vector will represent it as white assuming there is more white area than there is red. The median vector is also useful for when sclera segmentation includes superfluous pixels outside of the sclera. Assuming most pixels belong to the sclera, those pixels do not factor in to the final sclera color.

Table 4.3: Variations for feature extraction

PIXEL SELECTION METHODS	
All pixels	All pixels
No glare	$L \leq 220$ in HSL space
No glare or vessels	$L \leq 220$ and $H \geq 15$ in HSL space
No glare or eyelashes	$5 \leq L \leq 220$ in HSL space
No glare, vessels, or eyelashes	$5 \leq L \leq 220$ and $H \geq 15$ in HSL space
COLOR SPACES	
RGB, HSL, HSV, L*a*b*, YCrCb	

There are two considerations that must be considered for feature extraction (Table 4.3). The first is which pixels are considered in the calculation. The most obvious answer is to use all the pixels that survived the sclera segmentation presented in

Section 4.3.1. As mentioned earlier, though, not all pixels within the boundaries of the sclera actually represent the color of the sclera. Blood vessels and eyelashes can add undesired complications to the data. The median vector is meant to alleviate their effects, but as an extra precaution, BiliScreen uses the 5 different pixel selection methods described in Table 4.3. The thresholds for the different methods were selected empirically by examining images with prominent cases of glare, vessels, and eyelashes. They are by no means intended to capture all cases of non-sclera pixels; in fact, they are kept conservative on purpose to ensure that enough pixels remain in the calculation.

The second consideration for feature extraction is which color space is used. Images are saved from the smartphone camera in the RGB color space. Converting the image to a different color space is simply a calculation across the three channels that expresses those numbers in a different way, something that various machine learning models and feature transformation techniques can learn on their own. Nevertheless, explicitly carrying out color conversions can rearrange the color data in such a way that fewer features are needed. BiliScreen computes features for the 5 different color spaces listed in Table 4.3. Beyond features from the various color spaces, BiliScreen also computes the pairwise-ratios of the three channels in RGB. The intuition behind these features is that a yellower color will have low blue-to-red and blue-to-green ratios.

BiliScreen computes color representations of the sclera using every combination of pixel selection method and color space. Each color has 3 channels, resulting in 5 pixel selection methods $\times$ (5 color spaces $\times$ 3 channels per color space + 6 RGB ratios) = 105 features per eye. Not all of the features are used in the final model. Some pixel selection methods across the same regions can result in the same pixels, and some channels across color spaces represent the same information in similar manners. Automatic feature selection is used to select the most explanatory features and eliminate redundant ones. The top 5% of the features that explain the data according to the mutual information scoring function are used in the final models. Mutual information measures the dependency between two random variables [127]. The features that best explain the data

come from looking at the ratio between the green and blue channels in the RGB color space. A healthy sclera should be white, which produces high values across all three color channels. Blue is the opposite of yellow, so as the blue value of a white color is reduced, it becomes more yellow. This means that a high green-to-blue ratio implies a more jaundiced sclera.

4.3.5 Machine Learning

Separate models were developed for the two BiliScreen accessories to determine which would yield the better accuracy. The models use random forest regression and are trained through 10-fold cross-validation across participants. Note from Table 4.1 that the distribution of bilirubin levels is not evenly distributed; the healthy participants recruited from the university generally had similarly low values within 0.1 mg/dl, while the patients from the medical center had a far wider spread. The thresholds used in BiliScreen split the participants such that the normal and elevated classes have roughly equal sizes (31 vs. 25). The borderline class is roughly half as large (14), which is to be expected given that it is hard to catch such cases. To ensure that the training sets are balanced during cross-validation, splits are assigned using stratified sampling across the three bilirubin level classes. To be more specific, the typical fold includes 3 participants with normal bilirubin levels, 1 participant with a borderline bilirubin level, and 3 participants with elevated bilirubin levels.

The data collection procedure resulted in 2 trials per accessory $\times$ 4 gaze directions per trial = 8 images per accessory. Note that each image contains 2 eyes, leading to 16 eye images per accessory. Each eye is summarized with a feature vector that leads to its own bilirubin level prediction. For the results that are presented in this paper, the estimates from the 8 images are averaged to produce a final bilirubin level estimate that is reported back to the user. In the future, I plan to examine methods for selecting the best subset of images and only using them in the calculations.

4.4 Results

My evaluation examines BiliScreen's two major components: the segmentation algorithms and the sclera color-to-bilirubin level regression. I first examine the performance of the glasses' and sclera segmentation. I then show how accurate BiliScreen *can be*, assuming near-perfect segmentation of the glasses and sclera, as well as how accurate BiliScreen *is* with the current segmentation algorithms. I conclude by framing the accuracy of BiliScreen as a classification problem, showing how likely BiliScreen is to make the correct diagnostic decision.

4.4.1 Segmentation

All of the images were hand-annotated by the same researcher for ground truth. For the images taken with the BiliScreen box, the sclerae for both eyes were annotated; for the images taken with the BiliScreen glasses, the sclerae of both eyes, the colored squares, and the lenses were annotated. The performance of BiliScreen's segmentation algorithms can be described using precision and recall. The ground truth pixels annotated by the researcher are treated as targets. Precision defines the fraction of selected pixels that were correct, while recall defines the fraction of correct pixels that were selected. A low precision with a high recall would imply that the algorithm selects most of the pixels that belong to the target, but also includes several pixels outside of the target. A high precision with a low recall would imply that the algorithm only selects a fraction of the necessary pixels, but they are mostly within the target.

Glasses Segmentation

Finding the general region of interest for the sclera when the box is in use is trivial; it is based on rough rectangles on either side of the box. For the BiliScreen glasses, however, the region of interest is defined by the region within the glasses' lenses. I found that the glasses segmentation algorithm was able to locate the region of interest for the sclera with

a mean precision of 94.0 $\pm$ 15.0% and a mean recall of 94.4 $\pm$ 15.1% across all images relative to the lens borders defined by the human annotator. Recall is more important than precision for this problem because, as a region of interest, it is okay for superfluous pixels to be included as long as those belonging to the lens are included. The first step of the sclera segmentation algorithm attempts to rule out pixels outside of the eye agnostic of whether they represent skin or something else.

The glasses segmentation algorithm is also important for locating the colored squares around the lenses for color calibration. On average, the algorithm found the squares with a mean precision of 83.5 $\pm$ 24.2% and a mean recall of 88.2 $\pm$ 24.1% across all images. Unlike the sclera region of interest, precision is more important than recall for the colored squares because superfluous pixels can add noise to the calculation that summarizes the pixel colors to a single color value. Nevertheless, that is the specific reason for why the median vector is used over other aggregation functions. BiliScreen can tolerate mediocre precision as long as most of the pixels belong to the colored squares, which is true even within a standard deviation of my results. BiliScreen also takes advantage of the fact that there is a copy of each colored square on both sides of the face. The expected colors of the squares are known beforehand, so if a square on one side appears significantly different from the other with respect to the expected color, BiliScreen prioritizes the one that is closer to expectations.

Many of the issues that arose for the glasses segmentation can be attributed to their deformability. The glasses were made from a thin cardstock that could bend if the glasses were not large enough to fit on the participant's head. If BiliScreen cannot find a square, the algorithm fits the squares it has found to linear rows and columns and uses their intersection to find the missing one. When the rows and columns are actually curves, lines do not properly infer the squares' locations. Higher-order polynomials could have been used to model the curvature, but most of the squares required extrapolation rather than interpolation. That is to say, the locations of the squares had to be inferred outside of the range of the available squares, so even higher-order functions would not always

properly locate squares. In the future, I plan on improving the design of the BiliScreen glasses with a stiffer material and adjustable stems to avoid bending in the future.

Sclera Segmentation

As was the case for the glasses, ground truth for the sclera segmentation came from manual annotations. Pixel perfect labels are impossible by hand because of artifacts like eyelashes and blood vessels that encroach into the region. Nevertheless, those artifacts are handled post-hoc during feature extraction, so neither the ground truth annotations nor the segmentation algorithm are required to handle them.

Table 4.4: Sclera segmentation results per eye

	Precision	**Recall**
Box	$74.8 \pm 34.1\%$	$56.9 \pm 28.6\%$
Glasses	$74.8 \pm 35.0\%$	$43.1 \pm 27.1\%$

For sclera segmentation, a high precision with a low recall is also preferred over a low precision and a high recall. During the feature extraction phase, the colors of the individual pixels are summarized into single color vectors that describe the entire region. Having a small but correct region is likely to result in a similar calculation outcome, but including pixels outside of the target region can contribute noise to the result. Table 4.4 shows the per-eye precision and recall for both BiliScreen accessories. The spread of these measures can be misleading since the performance of the algorithm is roughly binary; the segmentation algorithm either identifies a region that corresponds to the sclera and only the sclera, or it completely misses and identifies another region, though it is correct more often than it is not. Looking deeper into the results, I find that 57.1% of the images from the box were segmented with $\geq$90% precision and 56.5% of the images with the glasses were segmented with $\geq$90% precision. Failures were not evenly distributed amongst all

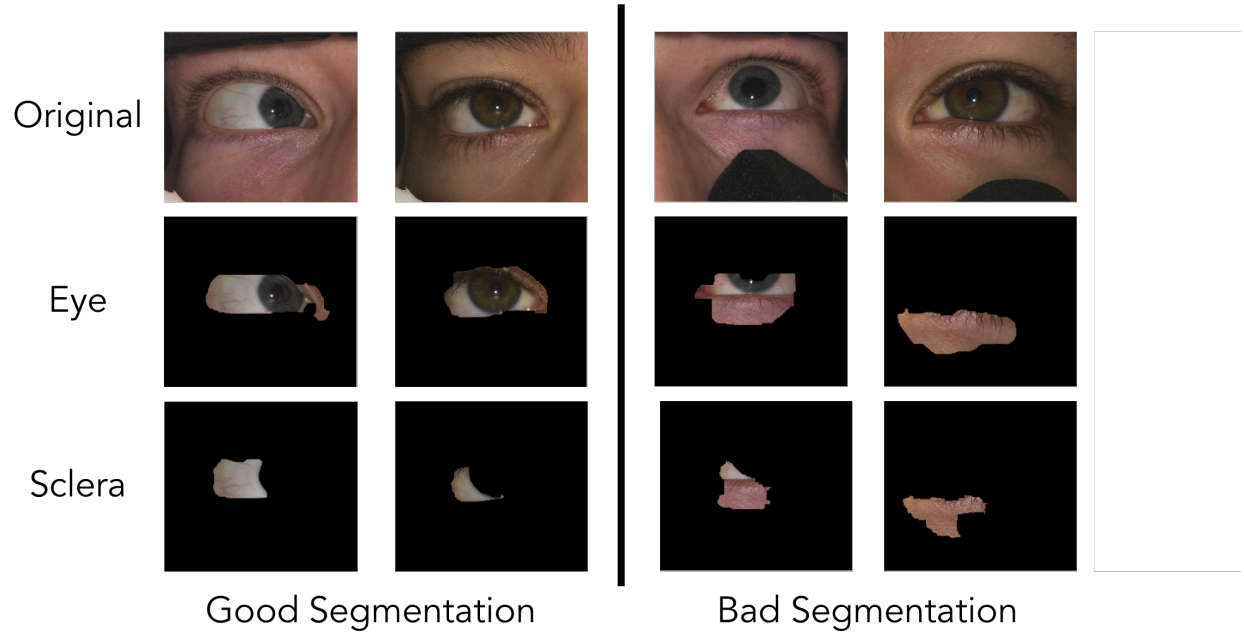

Figure 4.7: Example cases of BiliScreen's segmentation working **(left)** correctly and **(right)** incorrectly while the BiliScreen box was in use. These images come from individuals who were not recruited for the study in order to protect the privacy of those participants.

users.

Figure 4.7 shows successful and unsuccessful cases of sclera segmentation. For the most part, failures can be attributed to mistakes in the first half of the sclera segmentation algorithm, which uses GrabCut on the region of interest to locate the eye. In the first example of poor segmentation (third column of Figure 4.7), a faint shadow is cast onto the top-right part of the sclera since its curves away from the smartphone's flash. The sclera is assumed to be the brightest part of the image. Therefore, the algorithm prefers the rectangular initialization that includes the lower half of the sclera, which is bright, and the region just below the eye, where the flash reflects off of the skin and back to the camera. In the second example of poor segmentation (fourth column of Figure 4.7), the sclera has a naturally darker tint. Again, the flash produces a reflection

under the eye, so the algorithm completely fails to select any part of it. The dataset includes some users who squinted or blinked during the study. No attempts were made to manually curate images, and there was usually still enough exposed sclera so that a human observer could barely pick out the correct region. Nevertheless, I plan on implementing quality checks in a future version of the BiliScreen app to handle such cases. For the sclera segmentation with the glasses, errors can also be attributed to incorrect regions of interest from the segmentation of the glasses themselves. If BiliScreen could not the properly locate the glasses, then the algorithm makes its best guess, which can hinder later parts of the pipeline. This is another quality check that I believe will be necessary in the next version of the BiliScreen app.

4.4.2 BiliScreen as a Measurement Tool

Figure 4.8 shows the BiliScreen's optimal performance for estimating a person's bilirubin level when the exact boundaries of the sclera are known *a priori*. Of course, this claim assumes that the color-calibration procedure for the glasses and the feature extraction for both accessories properly capture the information needed to properly describe the color of the sclera. Although there are likely aspects of improvement in these regards, I suspect that automatic segmentation is the largest contributor of error since all calculations thereafter are dependent on its results.

The results are presented in two different arrangements. On the left, Figure 4.8 shows the correlation of BiliScreen's predictions with the ground truth measurements gathered from TSBs. The points are shown on a log-scale for clarity since the distribution is biased towards lower values. The dotted lines on the correlation plots indicate the 1.3 mg/dl and 3.0 mg/dl thresholds that separate the three groups of measurements. With the optimal segmentation, the Pearson correlation coefficient between BiliScreen's predictions and ground truth are 0.86 with the box and 0.83 with the glasses. On the right, Figure 4.8 shows the Bland-Altman plots of the same measurements. Again, the

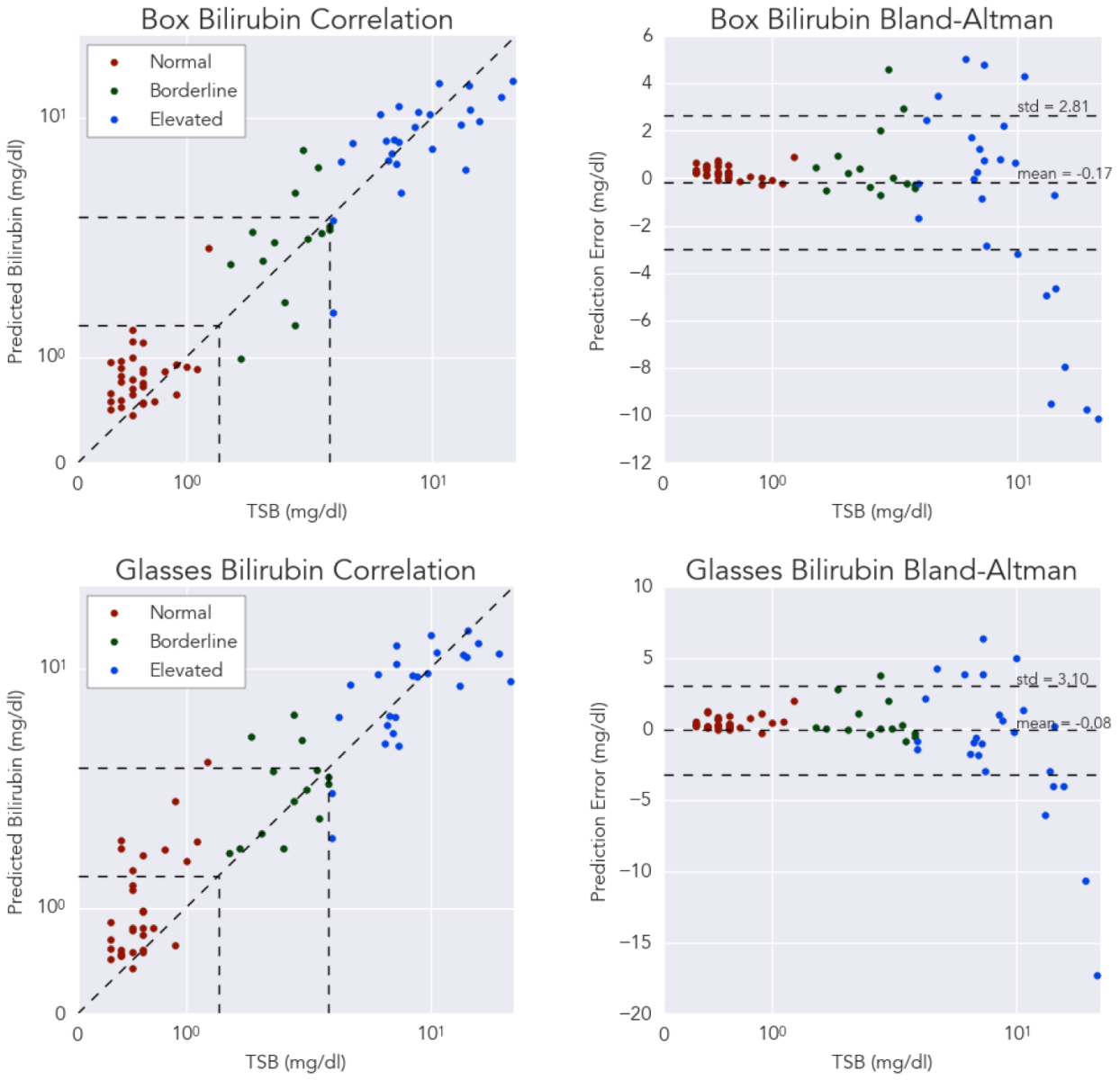

Figure 4.8: The **(left)** correlation and **(right)** Bland-Altman plots for BiliScreen's bilirubin measurements with the **(top)** box and **(bottom)** glasses using the optimal sclera and glasses segmentation. Note that the axes of the correlation plot are both in log-scale, as is the x-axis of the Bland-Altman plot.

x-axis shows the ground truth measurements using a log-scale for clarity. With the box, BiliScreen estimates the user's bilirubin level with a mean error of -0.17 $\pm$ 2.81 mg/dl. With the glasses, BiliScreen estimates the user's bilirubin level with a mean error of -0.08 $\pm$ 3.10 mg/dl.

The optimal models in their current state are more accurate for lower levels ($<$1.3 mg/dl) than they are for higher levels ($>$3.0 mg/dl). This can be attributed to the underlying distribution of bilirubin measurements for my participants'. Two participants returned a TSB value greater than 20 mg/dl, far beyond the threshold between borderline and elevated values. Since these participants were not thoroughly represented in my dataset, the optimal models underestimates their bilirubin level to fall more in line with the rest of the distribution. In general, higher TSB values lead to larger prediction errors for this very reason. Comparing the box and glasses accessories, the box yields better results. The box eliminates the effects of ambient lighting on the appearance of the sclera. The glasses require the extra step of color calibration, which introduces its own errors into the pipeline.

The results shown in Figure 4.9 are presented in the same manner as those in Figure 4.8, but were calculated using BiliScreen's automatic segmentation algorithms for the sclera and glasses. I anticipated that BiliScreen's overall performance would degrade with the use of imperfect segmentation. Regarding the sclera segmentation, extra pixels almost always belonged to the skin surrounding the eye. Skin often appears more yellow than the typical white of the sclera, so significant patches of skin can improperly lead to overestimation. The median color vector is used during feature extraction to counteract such behavior, but it is not sufficient for cases when the majority of the extracted region belongs to the skin. The prediction results using BiliScreen's automatic segmentation algorithms confirm my hypothesis, particularly for the glasses. The Pearson correlation coefficient for pictures taken with the glasses drops to 0.78, and the mean error of that model widens to 0.15 $\pm$ 3.55 mg/dl. To my surprise, the Pearson correlation coefficient for the box rises to 0.89, and the mean error improves to -0.09 $\pm$ 2.76 mg/dl. A careful

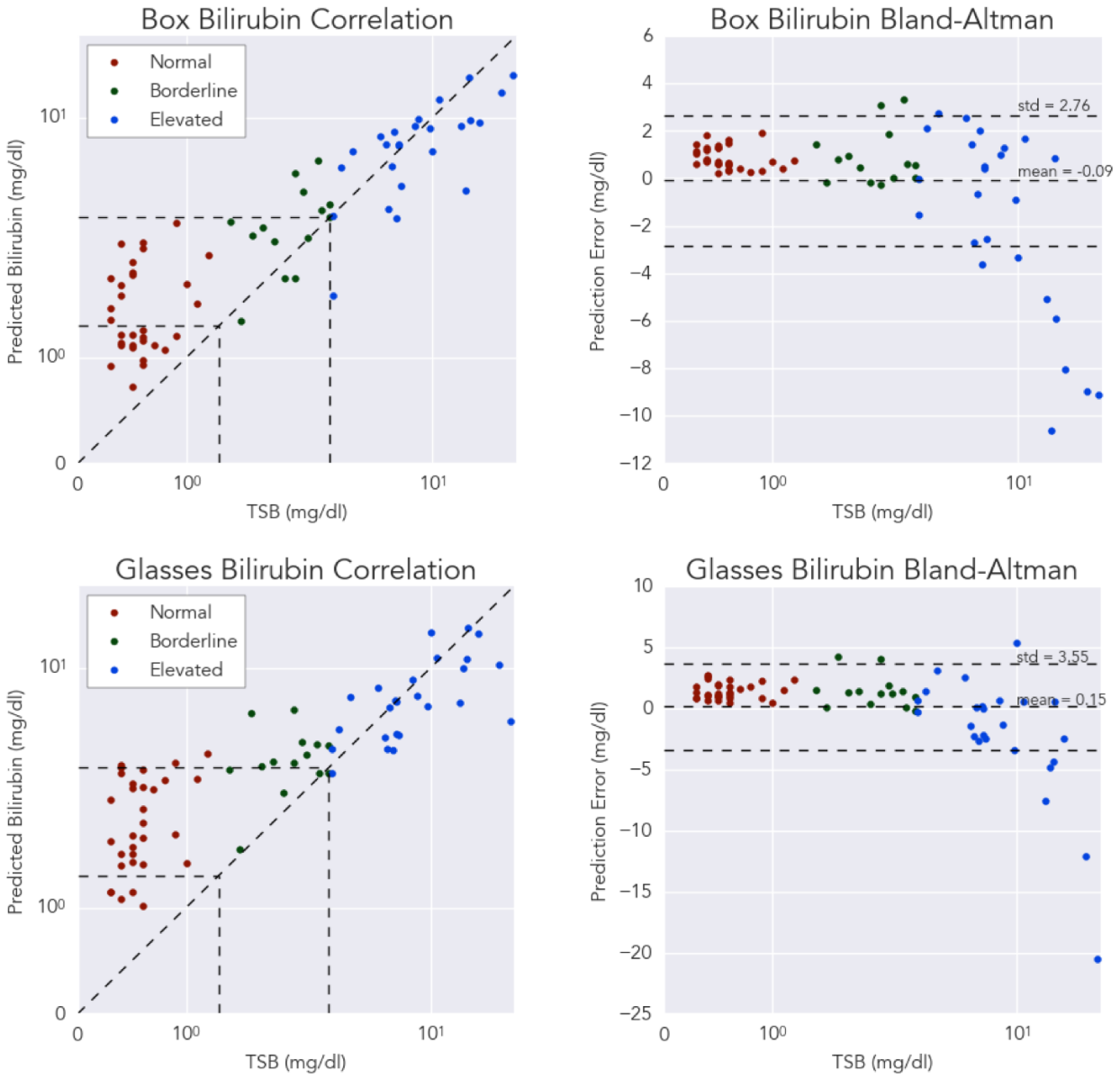

Figure 4.9: The **(left)** correlation and **(right)** Bland-Altman plots for BiliScreen's bilirubin measurements with the **(top)** box and **(bottom)** glasses using BiliScreen's sclera and glasses segmentation algorithms. Note that the axes of the correlation plot are both in log-scale, as is the x-axis of the Bland-Altman plot.

comparison of Figure 4.8 against Figure 4.9 reveals why this is the case. When the optimal segmentation is used to extract features, the model underestimates high TSB values. Because the addition of skin pixels can lead to overestimation, the underestimation is reverted and those predictions are improved. The model still overestimates all users, including those with normal and borderline bilirubin levels, but the improvement on the elevated levels outweighs the smaller errors that are incurred for those lower levels.

The results presented up until this point use all 8 images for each accessory, coming from the 4 gaze directions and the 2 trials. Asking the user to look in different directions provides different views of the sclera, some of which may exhibit more jaundice than others. Although these pictures take less than a minute to collect in total, I recognize that requesting users for 8 images can be burdensome. Using the optimal segmentation results, there is little disadvantage to using the images from a single gaze direction; the Pearson correlation coefficient for the box and glasses accessories varies by no more than 0.05 for any given direction.

Table 4.5 presents the Pearson correlation coefficient and error for BiliScreen's bilirubin measurements with the box and glasses accessories using the system's segmentation algorithm. Far more variation can be seen using the automatic segmentation, particularly when using the glasses and looking straight ahead. This could be because when the person looks straight ahead, the only parts of the sclera that are exposed are thin regions near the frames of the glasses. These regions are more likely to be covered in a shadow since they curve away from the camera and into the eye socket. The shadow not only affects segmentation, but also the color that is conveyed to the camera. Beyond this behavior, I do not believe there is any significant trend across different gaze directions. Incorporating more images into the final calculation allows BiliScreen to better tolerate an single image with incorrect segmentation. Sometimes, the results improved because incorrectly segmented images were removed from the final calculation. Other times, the results worsened because those same images were the only

Table 4.5: BiliScreen measurement results across different subsets of images

BOX - Pearson correlation coefficient, mean error $\pm$ std error	
All images	0.89, -0.09 $\pm$ 2.76 mg/dl
Looking up	0.84, -0.06 $\pm$ 3.03 mg/dl
Looking left	0.85, -0.15 $\pm$ 2.89 mg/dl
Looking right	0.82, -0.13 $\pm$ 3.21 mg/dl
Looking straight ahead	0.87, -0.05 $\pm$ 2.78 mg/dl
GLASSES - Pearson correlation coefficient, mean error $\pm$ std error	
All images	0.78, 0.15 $\pm$ 3.55 mg/dl
Looking up	0.72, 0.06 $\pm$ 3.18 mg/dl
Looking left	0.82, -0.06 $\pm$ 3.22 mg/dl
Looking right	0.83, -0.31 $\pm$ 3.09 mg/dl
Looking straight ahead	0.51, 0.28 $\pm$ 4.72 mg/dl

ones available for final calculation.

4.4.3 BiliScreen as a Classifier

The previous analyses have shown the accuracy with which BiliScreen can estimate a person's bilirubin level. Accuracy is always important, especially for capturing trends in the data. Nevertheless, the average user without a medical background is likely to be more concerned about how their estimated bilirubin level is classified rather than the value itself. In other words, if BiliScreen were to suggest that users with a borderline or elevated bilirubin level refer to a doctor for further tests, they would want assurances about BiliScreen's sensitivity (true positive rate) and specificity (true negative rate). From the perspective of the user, I group borderline and elevated bilirubin levels as positive cases when the user would be referred to a doctor and normal bilirubin levels as negative cases.

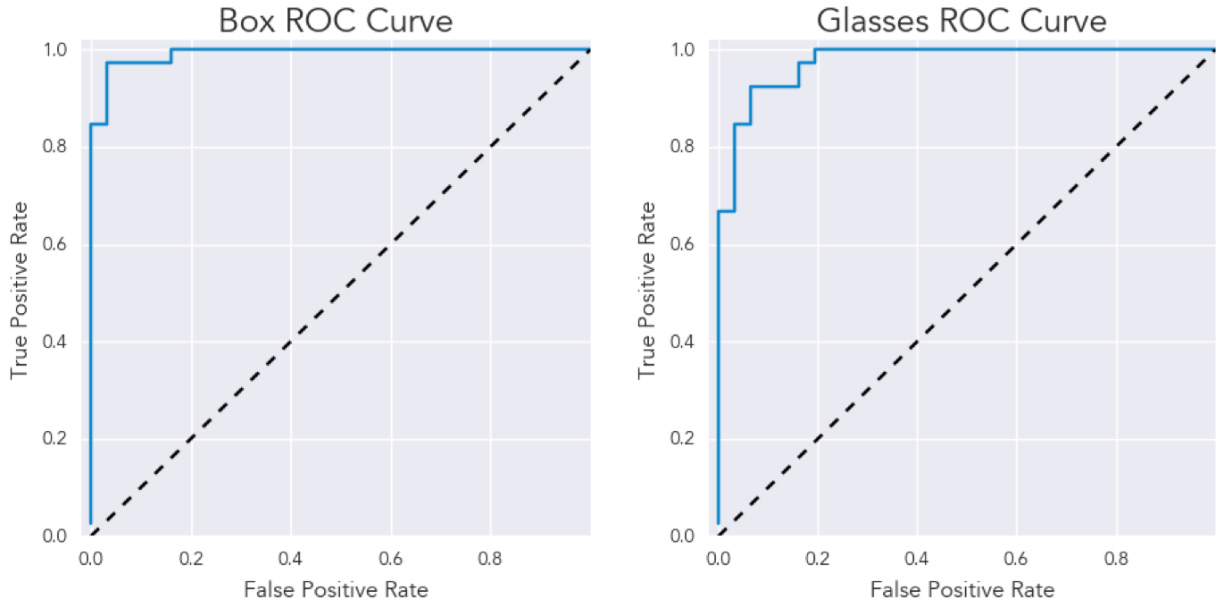

Figure 4.10: ROC curves showing BiliScreen's efficacy as a screening tool using the **(left)** box and **(right)** glasses using the optimal sclera and glasses segmentation. For the purposes of this analysis, normal bilirubin levels are considered negative cases, while borderline and elevated levels are considered positive cases.

Figure 4.10 shows the ROC curves for BiliScreen as a classifier using the optimal sclera and glasses segmentation. The area under the ROC curve (AUC) is 0.99 for the box and 0.98 for the glasses. Using the pre-determined threshold of 1.3 mg/dl used by the medical center, BiliScreen with the box achieves a sensitivity of 95.7% and a specificity of 97.4%. The threshold that maximizes the accuracy is only 0.1 mg/dl higher, increasing the sensitivity to 97.4% without a change to the specificity. Since the BiliScreen model with the glasses is more prone to overestimating lower bilirubin levels, it achieves a sensitivity of 100% and a specificity of only 71.4%. The threshold that optimizes accuracy leads to a sensitivity of 92.8% and a specificity of 94.3%.

Figure 4.11 shows the same curves for BiliScreen as a classifier using the system's segmentation algorithms. The AUC is 0.96 for the box and 0.95 for the glasses. Again,

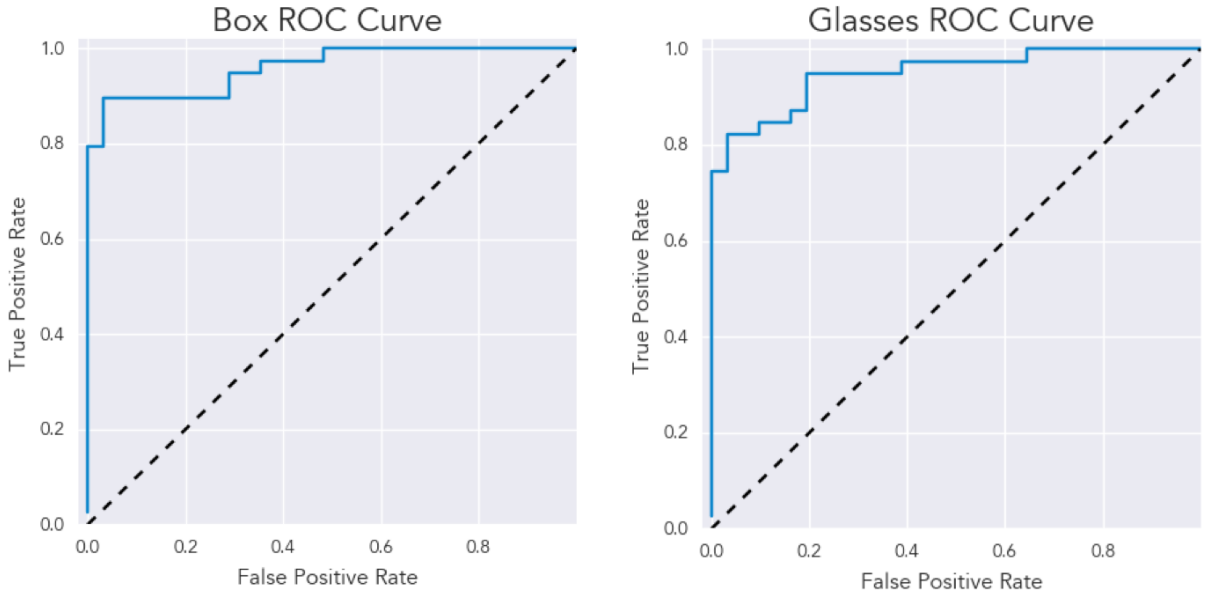

Figure 4.11: ROC curves showing BiliScreen's efficacy as a screening tool using the **(left)** box and **(right)** glasses using BiliScreen's sclera and glasses segmentation algorithm. For the purposes of this analysis, normal bilirubin levels are considered negative cases, while borderline and elevated levels are considered positive cases.

using the pre-determined threshold of 1.3 mg/dl leads to high sensitivity and low specificity since both models overestimate with the accidental incorporation of skin pixels. Using the optimal thresholds that maximize accuracy, BiliScreen with the box achieves a sensitivity of 89.7% and a specificity of 96.8%. With the glasses, BiliScreen has a sensitivity of 82.1% and a specificity of 96.1%.

4.5 Discussion

4.5.1 Hardware

The BiliScreen box was designed to block out ambient lighting, allowing the smartphone's flash to replace an otherwise varying surface reflection component with a constant one. However, the model associated with the box does not account for different

camera sensors. All of the data for this study was collected using the same iPhone SE device. Should another device be used, the BiliScreen model for the box would need to account for the camera sensor's response. This issue could be remedied in one of two ways. First, images could be gathered from the different cameras and separate models could be trained for each of them. This would clearly be a time-consuming endeavor, but lead to results like the ones presented in the paper. An alternative approach would be to perform a one-time calibration procedure as prescribed in Section 4.3.3 using a color calibration card within the box. The resulting calibration matrix would then be stored and applied on all images taken with the same device. This could be done offline by a researcher with a collection of devices, or the user could be asked to do it before using the BiliScreen app. The colored squares from the BiliScreen glasses could even be integrated into the BiliScreen box so that a separate color calibration card does not need to be purchased.

The latter approach assumes that a calibration matrix can perfectly correct an image's representation of color. Of course, this is the same assumption behind the BiliScreen glasses. If the assumption is not true, then the effects of the surface reflection component and the camera sensor's response cannot be fully eliminated. I believe that this assumption holds well enough that the lingering external effects on the sclera's color are negligible, but have yet to conduct a formal study on the matter.

4.5.2 Software

As mentioned in various parts of the paper, optimizations can be made throughout BiliScreen's pipeline. I use the mutual information scoring function [127] to automatically select the top 5% of the features that best explain the sclera color. In the future, I plan on manually examining the contributions of the features and determining if certain feature calculations are redundant. The final bilirubin estimate is also based on all 8 images captured per accessory. Taking so many images can be burdensome for the

user, but I also believe that getting the different views of the sclera ensures that any regions particularly affected by jaundice are captured. That being said, I have found that using all of the images only provides a small improvement to the final results. I plan on investigating this trade-off further.

4.5.3 Future Applications

BiliScreen does not directly assess a person's risk of pancreatic cancer; it examines the sclera for jaundice, one of pancreatic cancer's symptoms. Jaundice appears in other conditions, such as hepatitis and Gilbert's syndrome. Examining if there are differences between the visible symptoms of these diseases warrants further investigation.

The deployed implementation of BiliScreen depends on the target demographic for whom the app is designed. If BiliScreen were to be deployed as a screening application, I would prioritize notifying users about the possible risk of pancreatic cancer, even at the cost of extra false positives. This would be implemented by lowering the decision threshold for classifying a user's bilirubin level to increase sensitivity and decrease specificity; for example, lowering the decision threshold for BiliScreen with the box accessory improves its sensitivity from 89.7% to 95.2% while degrading the specificity from 96.8% to 71.2% (Figure 4.11, left). The downside to this change is that BiliScreen could induce a great deal of stress by falsely informing users that they may have a condition as serious as pancreatic cancer. To combat this issue, BiliScreen could require multiple consistent, high measurements before prompting the user to consult a clinician. If BiliScreen were to be deployed as a disease management tool, the trend of the data would be most important to clinicians.

Chapter 5

PUPILSCREEN

Traumatic brain injury (TBI) accounts for 30% of all injury-related deaths in the United States [66]. TBI can occur in a variety of situations, including car accidents, falls, and blunt force trauma. A concussion is a specific form of TBI caused by a swift blow to the head; these injuries tend not to be life-threatening, but can have serious and long-term effects on a person's memory, motor abilities, and overall cognition [171]. One area in which concussions have garnered national attention is sports, particularly contact sports such as boxing, hockey, and American football. The CDC estimates that there are roughly 3.8 million concussions per year in the US, and about half of them will go undiagnosed [96]. Patients suffering a concussion have a 600% increased risk of a future head injury and 15% increased risk of permanent cognitive deficits [96]. This is particularly more problematic for younger athletes who are not as well-educated on concussion prevention measures such as proper tackling technique. Roughly 250,000 young Americans (<20 years old) were treated for sports-related concussions in 2009 [32]. High school football players are 3 times more likely to suffer a catastrophic head injury than college football players [21]. Athletic departments with major funding can afford to have a team doctor with years of experience on-hand to diagnose concussions. For teams that are not as well-funded (e.g., pee-wee, middle school, high school), a school nurse, volunteer, or parent must put themselves in the same position as those doctors, but without the same tools or knowledge at their disposal. Identifying concussions immediately is essential because allowing a concussed athlete to return to play can lead to further significant injury [162]. There exists a need for accessible concussion screening that anyone can use at any moment. My proposed system,

PupilScreen, is meant to address this need by using a technology that most people have within arm's reach: a smartphone.

The methods that team doctors currently use to assess the probability of a concussion on the sidelines fall in one of two categories. The first category is task-based methods, which grade the performance of an athlete at a particular task using quantitative measures. For example, the King-Devick test [81] requires an athlete to read single digit numbers from left-to-right in different configurations. The second category includes survey-based methods, such as the Sport Concussion Assessment Tool (SCAT) [1]. Although a great deal of research supports the efficacy of these methods [80, 82], they capture indirect effects of concussions, require the athlete to be responsive, and take minutes to complete. These methods also require baseline measurements taken at the beginning of the season, which Broglio et al. [26] found were not repeatable for 118 healthy student volunteers. Furthermore, there is anecdotal evidence that athletes sometimes intentionally fail the baseline assessment so that there is little difference following an injury and they can remain in play [160].

A more quantitative method to assess a TBI is to check a person's pupillary light reflex (PLR), or the manner in which their pupils react to a light stimulus. The PLR of those who have suffered a TBI is typically either slower or not as pronounced [30]. The clinical gold standard for measuring the PLR uses a device called a pupillometer. Pupillometers are expensive (∼$4,500 USD) and are therefore mainly used in hospital intensive care units. Another method for assessing the PLR is through a penlight exam, in which a clinician directs a penlight towards each of the patient's eyes and observes the pupils' responses. This procedure is simple to perform, but has many drawbacks, including a lack of standardization, a need for deliberate training, and poor inter-observer reliability [182]. Those who provide first aid in emergency situations (e.g., EMTs and battlefield medics) will often conduct penlight exams despite these limitations because rapid assessment is

[1] http://www.sportconcussions.com/html/SCAT3.pdf

prioritized over precision.

PupilScreen combines the repeatability, accuracy, and precision of a pupillometer with the ubiquity and convenience of the penlight test for quantifying a person's PLR. The PupilScreen system consists of two ubiquitous components: a smartphone app and a box (Figure 5.1). Most people own a smartphone, and the box can be easily created since it does not require any wiring or expensive components. This means that PupilScreen can be available to almost anyone just hours before a sports event. The PupilScreen app records an 8-second video of a person's eyes as the pupils constrict in response to the smartphone's flash. The video is analyzed by convolutional neural networks (CNNs) in order to estimate the diameter of the pupils in each frame. I explored two different architectures. The first architecture uses two CNNs in sequence, where the first estimates the locations of the pupils and the second estimates their diameters given images cropped around their locations. The second architecture uses a fully convolutional network to perform pixel-wise segmentation. By examining how the pupil diameter changes over time, PupilScreen extracts metrics used by clinicians for diagnosis (e.g., constriction velocity, magnitude of diameter change). To standardize the results of the PupilScreen app, the smartphone is placed in a 3D-printed box. The box simultaneously eliminates ambient lighting conditions and controls the distance between the person's face and the flash.

Training CNNs requires a large quantity of diverse data, which is difficult to collect from patients with TBI. Therefore, I evaluated PupilScreen's ability to track the PLR on a dataset from 42 healthy adults. The range of pupil sizes encountered in non-reactive pupils is a subset of that encountered in reactive pupils; because the networks are trained on video frames in isolation, training PupilScreen on data from healthy individuals allows it to measure pupil diameter in individual video frames regardless of pupil reactivity. I found through my analysis that the PupilScreen was able to track pupil diameter with a median error of 0.30 mm with the fully convolutional network, the more accurate of the two approaches. Meeker et al. [165] found that manual pupil examination

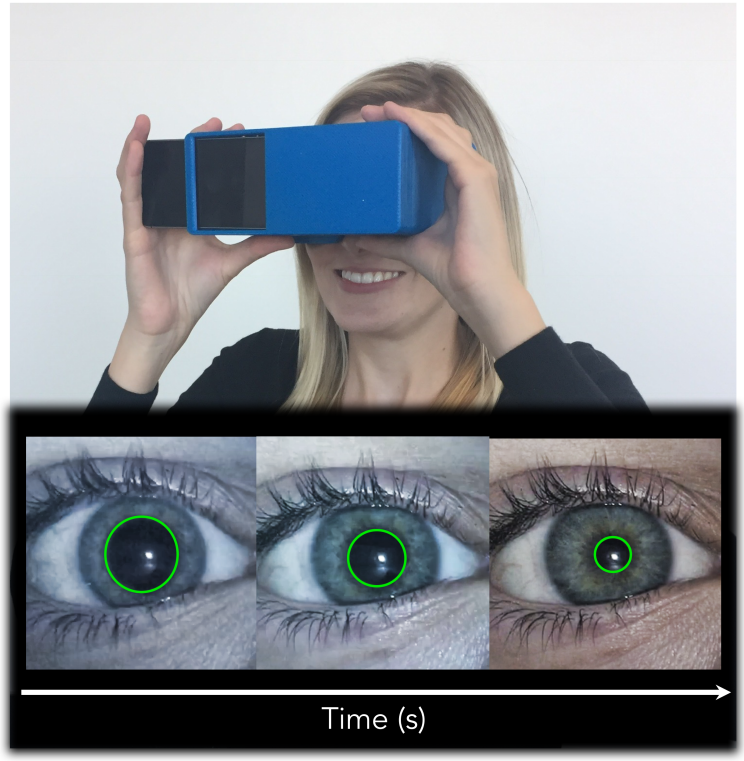

Figure 5.1: PupilScreen is a system that measures the pupillary light reflex to determine the severity of a traumatic brain injury. A smartphone app records a video of the patient's eyes as the camera's flash illuminates them. The VR headset-like box controls the position of the phone and the lighting that reaches the eyes.

has a median error of 0.5 mm, and a clinical pupillometer has a median error of 0.23 mm, which places the accuracy of PupilScreen between the two. PupilScreen was also able to track the pupil center with a median error of 0.20 mm. Using information about the pupil diameter over time, PupilScreen extracts three clinically relevant measurements: constriction amplitude, percentage, and velocity. I found that PupilScreen estimates constriction amplitude with a mean absolute error of 0.62 mm for a range of measured amplitudes that spanned 0.32-6.02 mm, constriction percentage with a mean absolute error of 6.43% for a range that spanned 6.21-62.00%, and max constriction velocity with a mean absolute error of 1.78 mm/s for a range that spanned 1.37-8.99 mm/s. To support

PupilScreen's efficacy as a diagnostic tool, I conducted a pilot clinical evaluation with six patients who had suffered a TBI. I found that clinicians were able to distinguish between normal and abnormal PLR curves produced by PupilScreen with almost perfect accuracy.

In designing a smartphone-based pupillometry system, the main challenges are:

1. Designing a controlled setup that is portable and inexpensive, and

2. Accurately identifying the pupils in video using only visible light.

My contribution comes in four parts:

1. The design and implementation of the PupilScreen system, which allows a smartphone to perform repeatable PLR tests at a fraction of the cost of a clinical device,

2. Two different CNN-based approaches for estimating the pupil diameter in videos,

3. An evaluation of PupilScreen's accuracy on 42 healthy participants, and

4. An evaluation of PupilScreen's ability to assist with diagnosis on 6 individuals who have suffered a TBI.

5.1 Background

Papers by Martinez-Ricarte et al. [157], Larson and Behrends [132], and Zafar and Suarez [264] provide thorough discussions on the mechanics of the pupil, the pathophysiology of the PLR, and the diagnostic power of the PLR. I summarize their content here for a broader audience, but refer the reader to their papers for a more detailed discussion of the PLR.

5.1.1 *The Characteristics of the PLR*

A normal PLR is defined as symmetric constriction or dilation of both pupils in response to a light stimulus or its absence, respectively. The pupil size must change by a non-trivial amount within a specified time frame and should change in both eyes, regardless of which eye is stimulated. For example, when a person covers one eye while the other is exposed to bright light, the pupils of both the covered and exposed eyes should constrict, producing a phenomenon known as the consensual response.

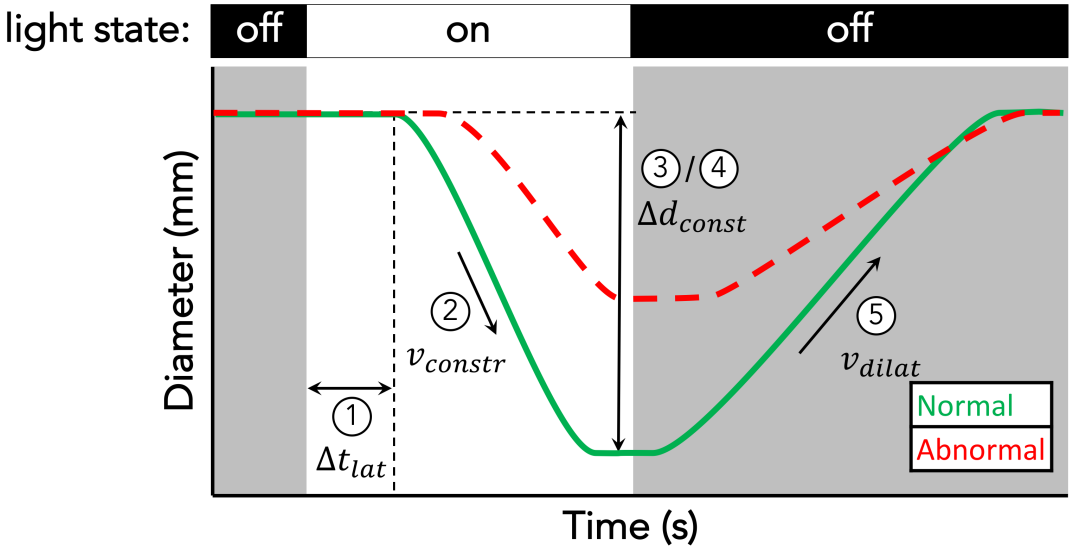

Figure 5.2: A PLR curve annotated with the five common descriptive measures: (1) latency, (2) constriction velocity, (3) constriction amplitude, (4) constriction percentage, and (5) dilation velocity. An abnormal PLR curve with increased latency, slower velocities, and diminished amplitude is also included for comparison.

When given pupil diameter as a function of time, clinicians focus on five simpler quantitative measures (Figure 5.2):

- **Latency (ms):** the time between the beginning of the light stimulus and the start of pupil constriction

- **Constriction velocity (mm/s):** the speed at which pupil constricts; reported as mean or max
- **Constriction amplitude (mm):** the difference between the maximum pupil diameter before light stimulation and minimum pupil diameter after light stimulation
- **Constriction percentage (%):** the constriction amplitude expressed as a percentage of the initial size
- **Dilation velocity (mm/s):** the speed at which the pupil dilates; reported as mean or max

5.1.2 Diagnostic Significance of the PLR

Because the neural pathways underlying the PLR include multiple brain regions and traverse many others, it is sensitive to a variety of injuries [250]. My motivating use case is traumatic brain injury. When the brain shifts inside the skull, it has the potential to injure both the cranial nerves carrying signals necessary for the production of the PLR or the brain regions that process these signals. A survey by Zafar et al. [264] in 2014 notes that the literature relating PLR to concussions is limited because it often includes a small number of patients ($\leq$10 patients with TBI) or individual case studies; however, researchers such as Ciuffreda et al. [234, 238, 239, 237, 240] have recently published the results of studies with larger datasets. In 2015, Thiagarajan et al. [234] quantitatively evaluated the PLRs of individuals with non-blast-induced, chronic, mild TBI (mTBI). That study included 15 healthy individuals and 17 patients with mTBI. Thiagarajan et al. found statistically significant differences between the two populations for most of the PLR metrics listed in Section 5.1.1. In a study published a year later, Truong et al. [237] carried out a larger study with 40 healthy individuals and 32 patients with mTBI. Beyond the larger study population, Truong et al. also studied how different light stimuli (e.g., pulses, step changes, different colors) could be used to better discriminate

certain PLR metrics. Populations of the same size were later examined to determine how pupillary asymmetry [239], photosensitivity [238], and refractive errors [240] affected the PLR. With more accessible pupillometry, such as that provided by PupilScreen, I believe that larger scale studies will be easier than ever before, particularly for examining the immediate effects on the PLR following a crisis.

Changes in the PLR are much better described by the literature in the context of severe TBI since those patients are often hospitalized and the changes are more obvious as a result of the severe cerebral dysfunction. Taylor et al. [232], for example, found that elevated intracranial pressure (ICP) for >15 minutes in patients with midline shift was associated with a decrease in pupillary constriction velocity. The PLR has also been examined as an indicator of the outcomes for patients following cardiac arrest. In a case study with 30 patients, Behrends et al. [18] found that the presence of a reactive pupil during the first five minutes of CPR was associated with increased survival and good neurologic outcome.

5.1.3 Techniques for Measuring the PLR

There are two methods used by clinicians to measure the PLR. The clinical gold standard method uses a device called a pupillometer. Infrared-based pupillometry takes advantage of the fact that there is a better demarcated boundary between the pupil and the iris when infrared imaging is used. While pupil diameter is tracked using infrared light, a ring of white LEDs stimulates the eye, causing the pupillary constriction. The components needed to make a pupillometer can be inexpensive, but the total product costs ∼$4,500 USD because, among other reasons, it is a self-contained system with proprietary algorithms and strict hardware requirements. Nevertheless, pupillometers provide two main benefits: precision and consistency. A study conducted by Meeker et al. [165] revealed that, for a modest participant pool, a pupillometer can track the pupil diameter with a median error of 0.23 mm. Couret et al. [45] asked multiple

clinicians to perform PLR measurements on 200 healthy volunteers in a variety of ambient lighting conditions. They found high intra-class correlation for maximum resting pupil size (0.95) and minimum pupil size after light stimulation (0.87) regardless of ambient lighting or device operator.

Figure 5.3: A penlight test being performed by a clinician.

A low-cost alternative for measuring the PLR involves using a penlight - a pen-sized flashlight (Figure 5.3). A penlight test is performed by directing the penlight toward and away from the patient's eye. Because the PLR is manually observed by a clinician, penlight-based pupil measurements are more likely to be inaccurate and imprecise. Meeker et al. [165] found that manual measurement of pupil diameter resulted in a median error of 0.5 mm, more than twice that of a pupillometer. Couret et al. [45] found a poor Spearman's rank correlation coefficient (0.75) between manual pupil size measurements and pupillometer readings. Only 64% of the cases when volunteers had pupils smaller than 2 mm were properly identified, and only half of the cases of anisocoria (i.e.,unequal pupil sizes) were caught. Larson et al. [133] note the inability of

clinicians to detect small, but clinically significant responses. Characteristics such as constriction velocity and amplitude also cannot be measured in absolute terms when using a penlight; instead of reporting a constriction velocity as 3.8 mm/s, observers can only describe the PLR as "normal", "sluggish", or "fixed". Penlight exams lack standardization as well. Clinicians purchase penlights from different companies, each with their own brightness specifications. Even if two health care providers use the same penlight, the patient may not experience the same light stimulus because of how the clinicians hold their penlights (i.e., distance and angle) or due to differences in ambient lighting conditions. Prior work has also discussed how penlight tests can lead to poor inter-observer reliability in PLR characteristics. Olson et al. [182] performed a single-blinded observational study where two practitioners were asked provide subjective scores for pupil reactivity. Across 2,329 paired assessments, Cohen's kappa coefficient was only moderate for pupil size (κ = 0.54), shape (κ = 0.62), and reactivity (κ = 0.40). In fact, only 33.3% of the pupils that were judged to be non-reactive by the practitioners were scored as non-reactive by pupillometry.

My prototype of PupilScreen is the first step towards combining the advantages of a pupillometer (repeatability, accuracy, precision) with the advantages of a penlight test (ubiquity, convenience). Before discussing how PupilScreen works, I will first provide an overview of pertinent related work.

5.2 Related Work

In this section, I summarize previous work concerning concussion diagnostics, gaze tracking, and pupil measurement.

5.2.1 Concussion Diagnostic Applications

Regarding concussions, metrics other than the PLR have been examined for diagnosis. Maruta et al. [158, 159] measured visual tracking performance in terms of gaze positional

error relative to a target and found that the performance variability increased for those with a TBI. Joiiv Lindsay [145] is one of the many researchers who have noted that involuntary eye movements are more prevalent in those with a TBI. Such work has been conducted in a clinical setting with dedicated devices; along with measuring the PLR, I look forward to investigating these metrics with PupilScreen in the future.

Lee et al. [137] provide a thorough survey of publicly available smartphone and tablet apps that are intended for assessing sports-related concussions. I refer the reader to their survey for a complete list of the smartphone apps that were examined, which includes both apps that are intended for non-medical personnel (e.g., coaches or parents) and medical personnel (e.g., team doctors). Lee et al. compared the purpose of each app to the SCAT2 and found that all of them exhibited partial or imperfect compliance to it. Furthermore, they found that the apps serve as a means of presenting, managing, and documenting various aspects of the SCAT2 rather than automating them.

5.2.2 Gaze Tracking

My work proposes a novel method for measuring pupil diameter. Although gaze tracking is a different problem - one that cares about the position of the pupil relative to the eye - the techniques used in both problems share many similarities.

The easiest way to track gaze involves the use of infrared light to emphasize the pupils. Infrared light is invisible to the naked eye and reflects off of the cornea, a fact which is leveraged in one of two ways. In bright pupil tracking, the light source is aligned with the camera so that the reflection can be tracked; in dark pupil tracking, the light source is off-angle so the pupil remains darker than the rest of the eye. There are a variety of commercial products by companies such as Tobii and LC Technologies that leverage this phenomenon for pupil detection. These products are primarily intended for controlled, desktop situations, but researchers have proposed form factors meant for on-site and outdoor scenarios. Fischer and van den Heever [68], Świrski et al. [229], and

Kassner et al. [117] are just three examples of techniques that take advantage of custom-designed headsets with an infrared camera pointed directly at the eyes for gaze tracking. All three of those systems are intended for gaze tracking and are evaluated as such, but their algorithms calculate both the pupil center and diameter as a means to that end. It should also be noted that Fisher and van den Heever's device uses gaze tracking alongside visual tasks like the King-Devick test with the intent of diagnosing sports-related concussions on the sideline, although there is no formal study on how that data improves the power of those tests.

There is a variety of methods for tracking the pupil without the help of infrared light. Qualcomm's SnapDragon SDK[2] provides facial features like gaze direction using a smartphone's front-facing camera, but their algorithm is proprietary. Timm and Barth [235] propose a mean of gradients approach for identifying the pupil center; essentially, the center is found using an optimization technique that identifies the pixel where a vector field of image intensity gradients is most likely to converge. For smartphones and tablets, EyeTab [261] relies on the observation that the pupil and the iris are normally concentric, so the center of the ellipse that best fits the edge between the iris and the sclera also corresponds to the center of the pupil.

Fuhl et al. have proposed a number of methods for detecting the pupil center. ExCuSe [77] utilizes two different techniques depending on whether the image contains a reflection or not. If there is a reflection, curved edges are found using dynamic thresholding and morphological operations; if there is no reflection, the coarse center is estimated using histograms oriented at various angles and then refined using an iterative ellipse-fitting technique [141]. ElSe [76] defines the pupil as the location where an image of the eye responds to two pre-determined convolutional filters: a circular mean filter and a surface difference filter. Finally, PupilNet [78] uses two CNNs for gaze tracking; the first CNN returns a coarse pupil center estimate, which is used to select a

[2] `https://developer.qualcomm.com/software/snapdragon-sdk-android`

region of interest that is fed into a second CNN to refine the prediction.

In this work, I explored two different network architectures. The first architecture is similar to PupilNet in that it involves two CNNs in sequence. However, instead of using the second network to provide a more precise estimate of the pupil center, I use the second network to estimate the pupil diameter. Although the first network only provides a coarse estimate of the pupil center, I demonstrate that it is sufficiently accurate for my purposes. The second architecture is an implementation of FCN-8, a fully-convolutional neural network proposed by Long et al. [147] for achieving pixel-wise segmentation.

5.2.3 Pupil Measurement

Researchers have extended existing techniques for identifying the pupil center to measure the contour of the pupil. Starburst [141] initializes an estimate of the pupil center using the mean of gradients approach. The algorithm then increments a marker in different directions from that seed until the first strong edge (defined by the gradient along this path crossing some threshold, which is expected to occur between the iris and pupil) is reached. An ellipse is fit to those edge points and its center is used as the seed for subsequent iterations of the same procedure until convergence.

A subset of the work in this area is particularly motivated by the use of pupil dilation as a proxy for assessing cognitive load. PupilWare [203] proposes improvements on the Starburst technique for use with a desktop web camera. These improvements include avoiding directions that could contain eyelash shadows and adding randomness to seed selection. Klingner, Kuman, and Hanrahan [125] do not discuss their pupil measurement algorithm in great detail, but provide a deeper analysis on task-evoked pupillary responses.

Many of the non-infrared-based techniques anecdotally cite issues for people with dark irises, even going as far as removing users with extremely dark irises from their studies. They primarily rely on the presence of an edge between the iris and the pupil.

PupilScreen uses a completely model-based approach that can learn features beyond edges (e.g., gradients and contiguous black pixels) for tracking the pupil.

5.3 Data Collection

I collected video recordings using the PupilScreen app and box to train its CNNs and evaluate its ability to track pupil diameter. Since my approach to segmenting pupils relies on CNNs, I require a large number of training examples from individuals with various pupil sizes and iris colors. This is difficult to attain through a patient population with TBI. Cases of TBI are limited, and the pupils of those with TBI usually stay a fixed size. Because of this, my networks are trained on data from healthy volunteers at the University of Washington and Harborview Medical Center. Below, I elaborate on the diversity of the participant pool. I then describe my data collection procedure, including the design of the PupilScreen box and my methods for gathering ground truth measurements. In Section 5.5.4, wIe present a preliminary evaluation conducted on six individuals with TBI to examine PupilScreen's clinical efficacy. All facets of my study were approved by the University of Washington's Institutional Review Board.

5.3.1 Enrollment

My training dataset comes from 42 volunteers: 16 males and 26 females. Typical non-infrared computer vision-based systems are reliant on determining the border between the iris and the pupil, which is more obvious for those with light blue eyes than those with dark brown eyes. For this reason, it was important to recruit participants with various iris colors. My study includes a balanced mix of iris colors: 17 blue, 20 brown, and 5 with a noticeable gradient between different colors. In most cases, the irises that were classified as mixed were light brown near the pupil but primarily blue.

Ideally, ethnicity should have no effect on PupilScreen's ability to measure the pupil diameter since the two are uncorrelated. I crop the images beforehand to reduce the

Table 5.1: Participant demographics (N = 42)

SEX - N (%)	
Male	16 (38.1%)
Female	26 (61.9%)
IRIS COLOR - N (%)	
Blue	17 (40.5%)
Brown	20 (47.6%)
Mixed	5 (11.9%)

number of skin-related pixels that are utilized by the CNNs; however, since my model-based approach for tracking the pupil is agnostic to the eye's structure, no guarantees can be made that the CNNs will not learn to estimate the pupil center or diameter from skin tone features. Although I did not specifically ask participants for ethnicity information, I note that one-sixth of the participants had a darker skin complexion.

5.3.2 Data Collection Application

All of the data was collected by the researchers using a custom app on an iPhone SE. The phone was placed into a slot in the back of the PupilScreen box (Figure 5.4). The design of the box is the same as the one used in BiliScreen [155], a project by a subset of this work's authors that aims to estimate the color of a person's sclera to detect cases of jaundice. The box-phone combination serves three purposes: (1) the box controls the position of the phone relative to the person's face, including the distance to and alignment with the face, (2) the box eliminates the effects of ambient lighting conditions, and (3) the phone provides its own lighting using the flash. The dimensions of the box are loosely modeled after the Google Cardboard. Besides the fact that the camera is

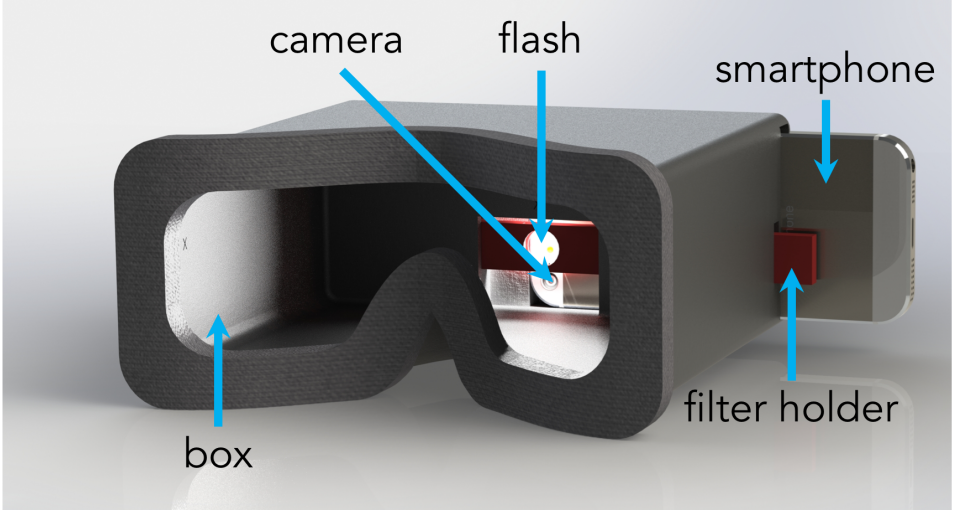

Figure 5.4: A 3D rendering of the PupilScreen box. The smartphone's flash lies in the horizontal center of the box. The box has a hole on the side so that a neutral density filter and a diffuser can be aligned with the flash using a sliding stick.

centered for the PupilScreen box, rather than the screen as in the Google Cardboard, the main difference between the two is the fact that the PupilScreen box is deeper. Having the camera close to the participant's face increases the effective resolution of their eyes, which allows PupilScreen to detect smaller changes in pupil diameter and measure the PLR with increased precision. On the other hand, moving the phone further away allows the camera to see both eyes at once and reduces the discomfort caused by the intense flash.

Although the box used in this study was 3D-printed for durability, I believe that it could be made with an even cheaper material like cardboard (provided that it is sturdy enough to support the weight of the phone). Also note that there is no electronic connection between the phone and the box, simplifying its manufacturing requirements. Apple iOS 9 does not provide complete dimming control over the brightness of the flash LED. At close distances, participants from a pilot study found the intensity of the light to be uncomfortable. To make the light more manageable, a neutral density filter and

diffuser were placed directly in front of the flash using a sliding stick. These components were chosen because they had precise specifications available online, but they could be replaced with a cheaper alternative like a sheet of white computer paper in the future.

Prior to putting the box up to their face, participants were asked to take off glasses if they wore them. Once the phone was placed in the box and the participant held it up to their face, the flash was turned on briefly and autofocus was enabled. The resulting camera focus was fixed for the remainder of the study to avoid blurriness as the lighting in the box changed. The flash was then turned off and after a brief pause to allow the pupils to recover, data collection commenced. The video was recorded at 30 fps with 1920×1080 resolution. After an audible 3-second countdown from the phone's speakers, the flash illuminated the participant's eyes. The stark change in lighting maximized the degree to which the pupil constricted, akin to the difference experienced when using a pupillometer. The recording stayed on for another five seconds, resulting in an 8-second long recording. The five second period after the introduction of the light stimulus was far longer than what was needed to capture the PLR, but provided extra video frames for evaluation. For each study participant, the PLR was recorded three times. Between recordings, a one-minute break was added to allow the participant to rest their eyes.

5.3.3 Ground Truth Measurements

Videos were manually annotated to generate ground truth labels. Using custom software, two researchers labeled frames by selecting points along the edges of the pupils and letting OpenCV's ellipse fitting algorithm generate a corresponding outline (Figure 5.5). The researchers could see and adjust the outlines to better fit the images. If the pupil was difficult to distinguish from the iris, the researchers could adjust the contrast to make it more visible. If the pupil was still too difficult to see after that, either because of poor focus or lighting, the frame was skipped; this only happened for 1.8% of the total frames encountered. The points were fit to an ellipse because not all pupils are

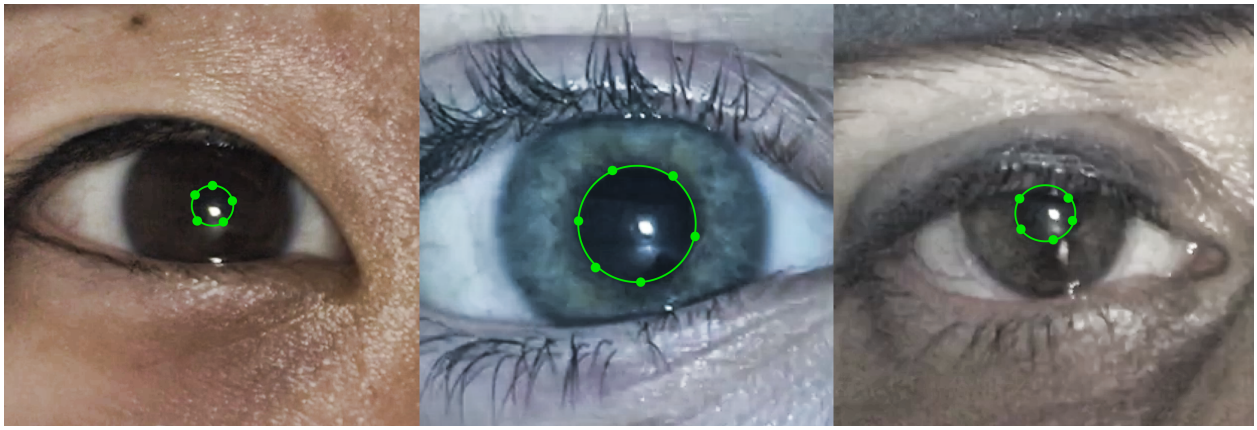

Figure 5.5: A selection of manually annotated images of pupils zoomed in on the region of interest. Note that although the pupil may seem indistinct from the iris in some of the images above, the labeling was performed on much larger monitors with better contrast than what appears in print.

circular. Since pupillometry is only concerned with a single pupil diameter, the ellipses were converted to circles by averaging their axes. With this method, the pupil diameters were labeled in pixels. The researchers labeled every fifth frame in the three videos from each user. Each video was 8 seconds long, but the first 3 seconds occur before the flash was turned on, resulting in 5 seconds $\times$ 30 frames/second $\times$ (1/5 frames) $\times$ 3 videos = 90 labeled frames per person. Frames were labeled independently of one another to avoid biases between frames; however, this led to greater variation between consecutive frames that can be primarily attributed to human error. A 3rd-order Savitzky-Golay filter was applied to temporally smooth the pupil center and diameter labels. To quantify the agreement of the labels across the researchers, both of them labeled a common set of 5 users (15 videos, 450 frames). The average difference between the smoothed pupil center labels was 3.46 px, which translates to 0.27 mm. The average difference between the smoothed pupil diameter labels was 2.00 px, which translates to 0.16 mm. Note that these variations are not independent; if a researcher underestimated the extent of an edge, the labeled center would move away from that edge and the labeled diameter

would be lower than the actual value. The degree of inter-researcher agreement can also be quantified using the intersection-over-union (IoU) measure, a standard metric for segmentation agreement between two regions. The mean IoU for the researchers' labels was 83.0%. Note that the IoU measure is calculated relative to the total area of the two labeled pupils. If the pupil center labels for a 3 mm pupil were only off by a single pixel, that difference alone would lead to an IoU score of 93.8%.

Although a clinical-grade pupillometer could have provided an alternative method for quantifying the PLR, its results would not have been directly comparable to PupilScreen. The two setups have light stimuli with different intensities, which would result in different magnitudes of pupil constriction. Furthermore, PupilScreen eliminates the effect of ambient lighting because the box completely encloses the patient's eyes, whereas pupillometers do not since they are used in hospitals with roughly standard lighting conditions. Infrared imaging could have been used to provide a comparative ground truth measurement of pupil diameter; however, an algorithm still would have been needed to turn those frames into pupil diameters, and that algorithm would have needed its own validation.

5.4 Algorithm

In this section, I will describe how the video data was pre-processed before being input to the CNNs. I then follow by describing the architecture of the CNNs used to estimate the pupil center and the pupil diameter, the post-processing of the CNN outputs, and the specifics of the CNN training.

5.4.1 Pre-processing

Videos were recorded at 30 fps with 1920×1080 resolution. Treating each pixel as an individual input feature produces a very large input layer with a significant amount of unnecessary information; pixels around the eye socket provide no information about the

Figure 5.6: Each frame was cropped to create two input images for the CNNs: one for the left eye and one for the right eye. The image of the right eye and its label were flipped to make the two images comparable.

pupil, and pixels on the left and right sides of the image should be considered independently in order to catch cases in which the pupils behave differently. I attempted to crop around the eyes using off-the-shelf eye detection algorithms, but found that they failed in many cases. This may have been because the detection algorithms rely on the presence of other facial features (e.g., nose) that are obscured by the PupilScreen box. Instead, the conservative cropping bounds in Figure 5.6 are used. The bottom third is cropped off because it only contains the box. The remainder of the video frame is split into two halves - left and right - to produce one image per pupil. To make the images comparable and allow a single CNN to handle each task, the image of the right eye and the coordinates of its pupil center label are flipped horizontally. To emphasize the pupil, the image is converted to the HSL color space and contrast-limited adaptive histogram equalization (CLAHE) [195] is applied to the lightness (L) channel. In short, CLAHE avoids the pitfalls of global histogram equalization by dividing an image into small tiles (88 px in this case) and then equalizing only within those individual tiles.

5.4.2 CNN Architectures

Two different architectures were tested for measuring the size of the pupil. I describe their inspiration and implementation details below.

First Architecture: Sequential CNNs

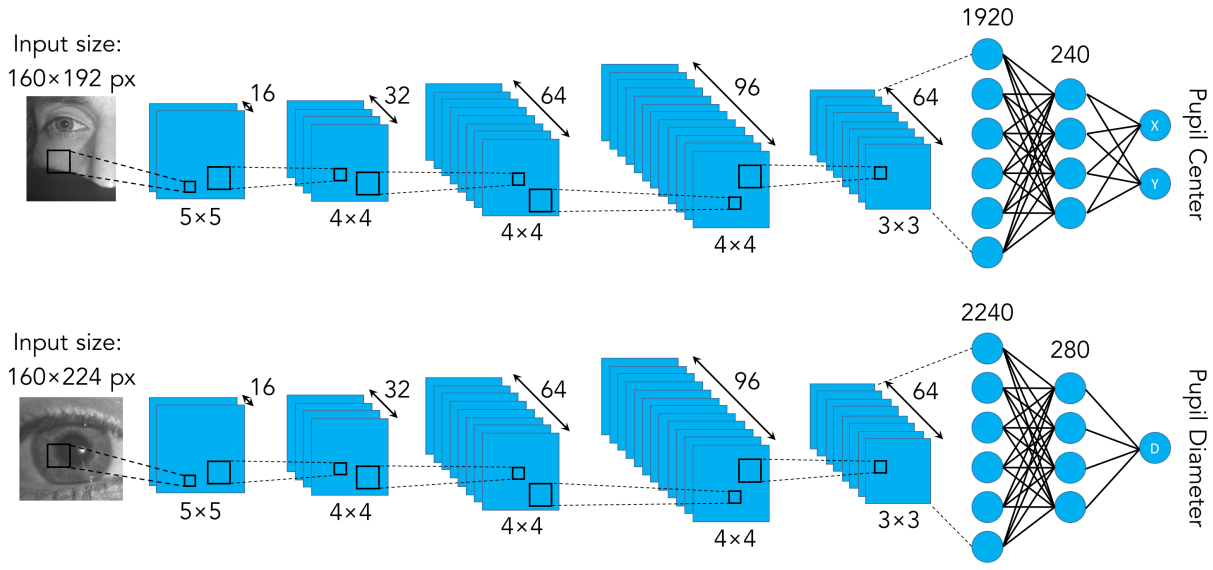

Figure 5.7: The first architecture that was explored for PupilScreen. The top numbers indicate the number of filters in the convolutional layers or neurons in the fully-connected layers. The bottom numbers specify filter dimensions. For example, the first convolutional layer in both networks applies 16 5×5 px filters. There are 2×2 px mean-pooling layers after each convolutional layer, but they are omitted for space. **(top)** The first CNN takes the original image as an input and returns an estimate of the pupil's location. **(bottom)** Given the location of the pupil center, a region of interest is cropped from the original image and provided to the second CNN to estimate pupil diameter.

The first architecture was similar to that of PupilNet by Fuhl et al. [78], which uses two networks in sequence to arrive at a precise estimate of the pupil center. The intuition behind their approach was that the first network reduces the search space for the pupil by

roughly localizing the pupil center, allowing for the second network to ignore irrelevant pixels and examine a specific region in more detail. Inspired by that intuition, I also explored the use of two networks for pupil measurement. The first network serves the same purpose, but the two applications differ in the second network. Rather than learning a finer pupil center measurement, I train the second network to learn the pupil diameter. I demonstrate that even if the pupil is not exactly centered using the output of the first network, the second network can be robust enough to handle those issues.

Figure 5.7 illustrates the details of the first architecture. The first network (Figure 5.7, top) is trained to accept an image from the pre-processing step as input and return the location of the pupil center. Before being input to the network, the image is downsampled by a factor of 4. The network has 5 convolutional layers, each with a rectified linear (ReLU) activation function followed by 2×2 px mean-pooling layers. Mean-pooling was chosen over max-pooling because max-pooling results in translation-independent behavior that would have been undesirable for capturing location information. The final layer of the first network is fully-connected to compress information across all filters and sub-regions to an x- and y-coordinate estimate. The output labels were normalized according to the mean and standard deviation of the pupil location across the entire dataset. This was done to ensure that the same error in either direction would equally affect the network's weights during backpropagation.

Using the output of the first network, a region of interest that is roughly $1/9^{\text{th}}$ of the original image's size is cropped and centered about the estimated pupil. That region is provided to the second network (Figure 5.7, bottom), which is trained to estimate the pupil diameter. The network has the same architecture as the first one except for the fact that it produces a single output: the pupil diameter.

The number of layers was determined empirically to balance the tradeoff between network size and accuracy. Smaller networks are desirable so that they can fit on the smartphone, but I found that using fewer layers did not yield satisfactory results. The other specifics of the networks (e.g., more smaller filters as the network gets deeper,

pooling after each set of convolutional filters) were based on suggestions from literature [88], but are certainly an area for future investigation.

Second Architecture: Fully Convolutional

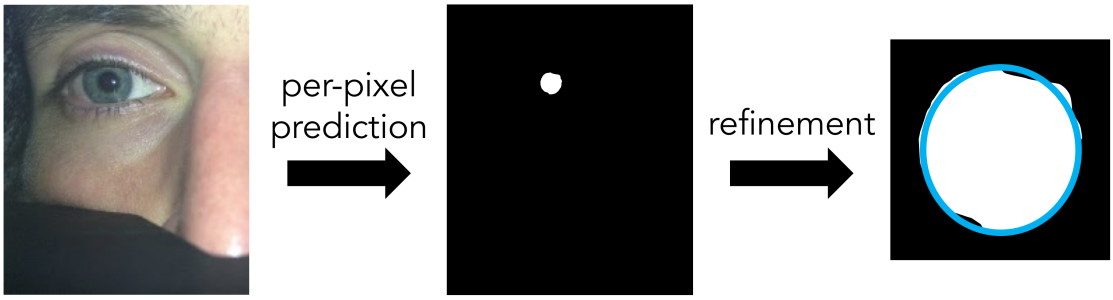

Figure 5.8: The second architecture assigns each pixel to one of two classes: "pupil" (white) or "non-pupil" (black). The largest contiguous cluster of "pupil" pixels is assumed to be the pupil, and its border is smoothed so that it can be fit to an ellipse.

The first network architecture learns the pixel indices of the pupil center and the diameter of the pupil, but treats them just like any other continuous outputs rather than explicit location and size information. The second network architecture takes a different approach, viewing the problem as one of explicit segmentation. The goal of segmentation is to produce a label for every single pixel that specifies the object to which it belongs; as illustrated in Figure 5.8, there are two classes for the purposes of PupilScreen: "pupil" and "non-pupil". I implemented FCN-8, a fully convolutional architecture proposed by Long et al. [147]. In short, fully convolutional networks are normally based on a pre-trained convolutional network for image classification (e.g., VGG16 [225]). The final classifier layer is removed and replaced by layers that deconvolve, or upsample, the downsampled predictions to their original resolution. For the sake of network size, I downsample images by a factor of 2 before inputting them to

the network.

Once pixel-wise predictions are produced, there is still the matter of measuring a pupil diameter. The largest contiguous cluster of pixels with the "pupil" label is treated as the pupil. The border of that cluster is smoothed using median blurring and then fit to an ellipse. The mean of the ellipse's two axes is treated as the pupil diameter for that frame.

5.4.3 Training

Both architectures were trained with backpropagation using batches composed of 10 images randomly sampled from the training set. To ensure that there was no overlap between training and testing data, the evaluation was conducted using 5-fold cross-validation across users; in other words, if there are N users, N/5 users are held out each time for testing and the remaining 4×N/5 users are used for training. Recall that three videos were recorded for each user. All networks were trained for 10 epochs per fold; this number was determined empirically based on the convergence of the smoothed loss function outputs across the training data. On average, training the first network architecture took 14 mins per fold, resulting in a total training time of 14 mins × 5 folds × 2 networks = 2 hours 20 mins. Training the second network architecture took 1 hours 59 mins per fold, resulting in a total training time of 119 mins × 5 folds = 9 hours 55 mins. Computation was carried out by a single NVidia GeForce Titan X GPU. Testing an individual frame through either network architecture took approximately 2 ms, which means that it would take the system roughly 2 ms × 30 frame/second × 5 seconds = 300 ms to test an entire video. The networks in the sequential CNN architecture were trained using batch gradient descent in order to minimize the L_2 loss. The fully convolutional network was trained in the same way to minimize the per-pixel multinomial logistic loss.

To ensure that the dataset was not significantly biased towards images of fully constricted pupils, only frames within the first 3 seconds of the light stimulus were used

for training. To both generate more training samples and further promote training data diversity, training images and their associated labels were randomly jittered together (i.e., translated by a small amount). That amount was at most 10% of the input image dimensions for the first network, which was determined based on the variation of the pupil center observed in the videos. The jitter amount was at most 15% of the input image dimensions for the second network in order to sufficiently cover the spread of pupil center predictions from the first network. In this latter case, jittering the input images allows the second network to be trained to tolerate such errors.

5.4.4 Extracting PLR Metrics

In the end, the consecutive CNNs in PupilScreen take an individual image as input and return the pupil's diameter as output. A PLR curve shows a patient's pupil diameter as a function of time following a light stimulus. To construct this, videos are passed through the networks frame-by-frame. From that point, there are three post-processing steps to make the resulting curve more comparable to the curves provided by pupillometers: (1) Extreme prediction outliers are removed using heuristics based on human physiology: pupils should not be smaller than 1 mm or larger than 10 mm, and the pupil diameter should not change by more than 10 mm/s [30]. (2) Like the ground truth labels, the predictions are smoothed using a 3rd-order Savitzky-Golay filter. This removes undesirable fluctuations between frames that occur because the pupil diameter is estimated from each frame individually. (3) Predictions are scaled from pixels to millimeters using a constant factor that was estimated through a device calibration procedure. A fiducial of known dimensions was placed in front of the camera at roughly the same distance as the user's eyes; its dimensions were measured in pixels and the calculated ratio was applied to all videos. This approach is not perfect since different people have different eye socket depths. Nevertheless, the ground truth labels used for analyses are all in pixels, so the conversion is primarily used to transform the results into

more relevant units.

Relevant clinical measures (Section 5.1.1) can be extracted from the smoothed and scaled PLR curve. Calculations for the constriction amplitude and the constriction percentage require the minimum and maximum pupil diameter. The maximum pupil diameter always occurs at the beginning of the video since the pupil is most dilated before the light stimulus. After the pupil constricts, its diameter can fluctuate as it reaches its final equilibrium size. Because of this, the minimum diameter is identified by taking the average diameter in the last second. The maximum constriction velocity is calculated by computing the maximum of the centered derivatives across the entire curve. Although PupilScreen is designed to measure the latency between the time of the light stimulus and when the pupil begins to constrict, I found that the frame rate limits the granularity of the calculation ($(30\text{ fps})^{-1} = 0.03$ s/frame) and the usefulness of that measure, so I ignore it for this study.

5.5 Results

Since PupilScreen is a data-driven algorithm, the diversity of the data used to train the algorithm is important. Section 5.3.1 details the diversity of the participants, but in Section 5.5.1, I describe the quantitative diversity of the pupil center and diameter. I then present the accuracy of PupilScreen's ability to localize and measure the pupil with the two different architectures that were explored, followed by an examination of how the errors manifest in the PLR curves and affect the PLR metrics. I conclude with a brief evaluation of PupilScreen's clinical efficacy, including how accurately clinicians can make diagnostic decisions based on PupilScreen's estimated PLR curves and their comments on PupilScreen's design.

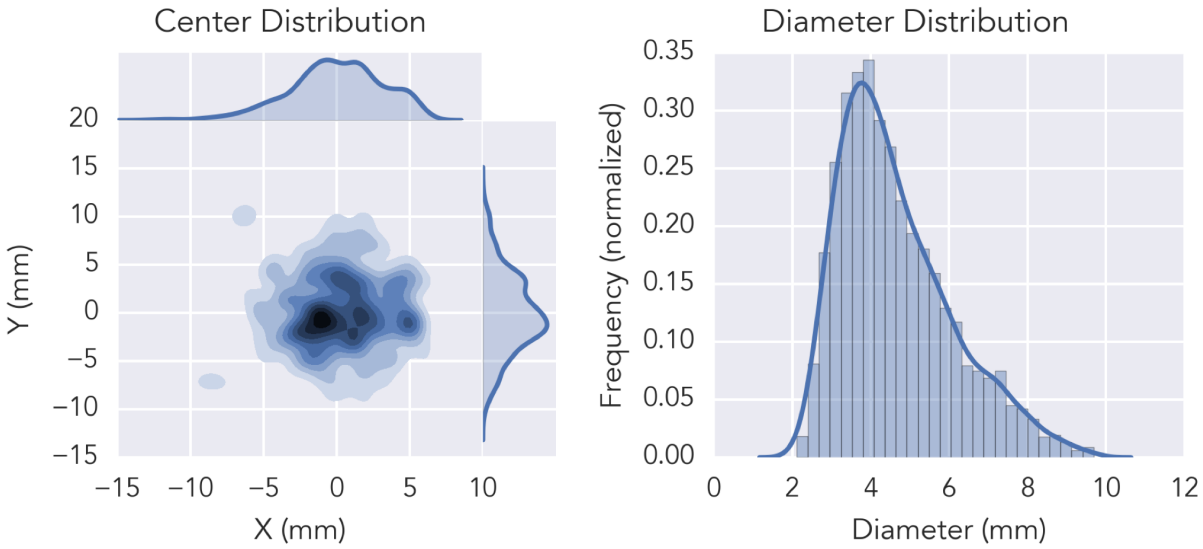

Figure 5.9: **(left)** The distribution of the pupil centers across all users. **(right)** The distribution of the pupil diameters across all users.

5.5.1 Data Distribution

The left side of Figure 5.9 shows the distribution of the pupil center location across all users after the video frames were cropped, flipped, and scaled to millimeters. The distribution is centered at the mean pupil center for reference. The distribution has a standard deviation of 3.22 mm in the x-direction. This spread can be attributed to variation in interpupillary distance and the fact that participants did not perfectly align their face within the PupilScreen box. The distribution has a standard deviation of 4.18 mm in the y-direction, which can also be attributed to different face shapes and the placement of the PupilScreen box relative to the participant's face.

The right half of Figure 5.9 shows the distribution of the pupil diameter scaled to millimeters. The distribution has a mean of 4.39 mm and a standard deviation of 1.38 mm. However, the distribution is non-normal because the pupil constricts in a logarithmic fashion, which means that the pupil only spends a small amount of time in its fully dilated

state.

5.5.2 CNN Results

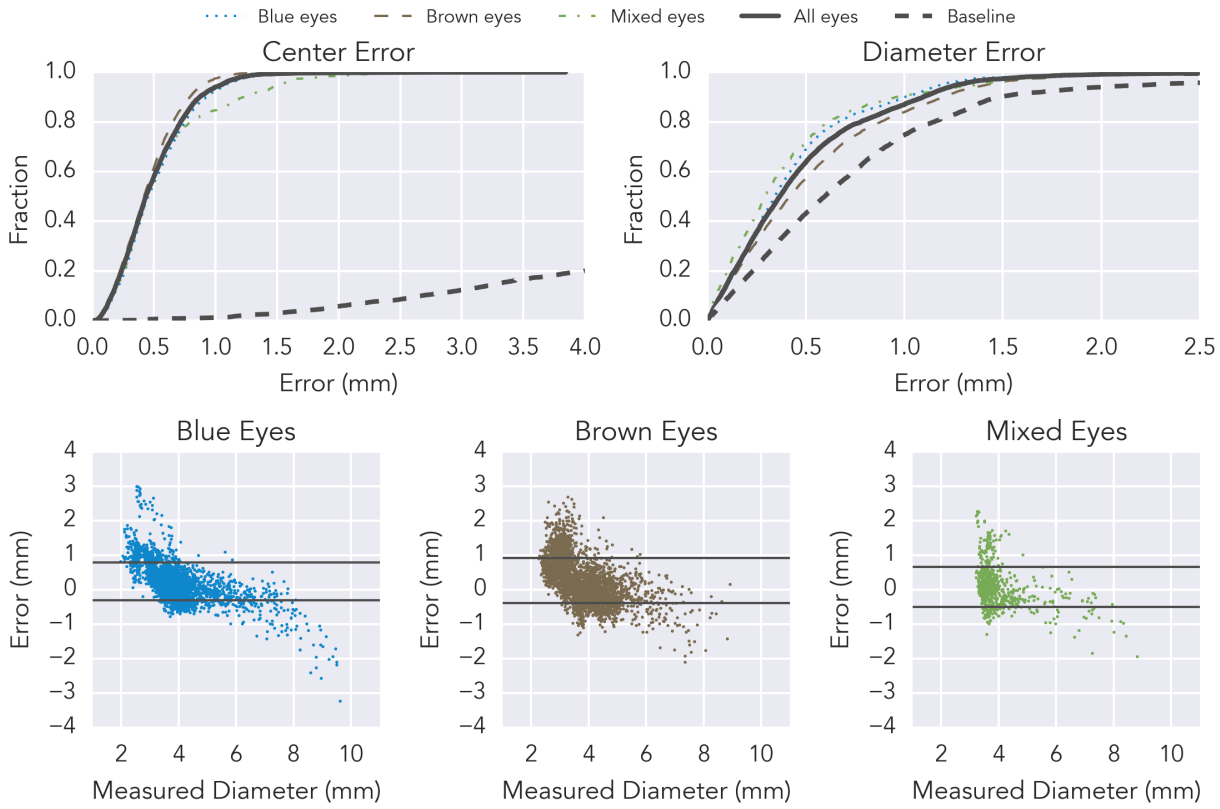

Figure 5.10: The accuracy results for the sequential network architecture. **(top-left)** The CDF of the pupil center prediction error. **(top-right)** The CDF of the pupil diameter prediction error. **(bottom)** Bland-Altmann plots showing the residuals of the pupil diameter predictions split across the different iris colors: blue, brown, and mixed from left to right. The black lines indicate one standard deviation from the mean.

The cumulative distribution functions (CDFs) at the top of Figure 5.10 show the distribution of the absolute errors for the sequential network architecture. The thick dashed line in both plots compares the results to a baseline that assumes the mean predictions for all users; this is not meant to serve as a comparable algorithm, but rather

ground the results relative to some other estimator. Improvement over the baseline demonstrates that the networks are learning more than just the mean value.

The top-left of Figure 5.10 shows the CDF for the errors of the first network, which estimates the pupil center for a cropped input video frame. Across all users, the distribution of Euclidean errors has a median of 0.43 mm and a 90^{th} percentile of 0.87 mm. The error distributions across the different iris colors are nearly identical. The magnitude of the error can partly be attributed to the pre-processing of the video frame. Input images are downsampled by a factor of 4, which reduces the resolution of the pupil center estimation to 0.31 mm. Despite the loss of resolution, the errors are well within the diameter of the iris (10-12 mm). In fact, most are within the smallest observed pupil diameters ($\sim$2 mm). Although it is ideal for the pupil to be centered in the image that is input to the second network, the most important result is that the eye always remains in the region of interest that is cropped around the center prediction. By jittering the training data, the second network is trained to handle shifted images.

The top-right of Figure 5.10 shows a similar CDF plot for the errors of the second network, which estimates the pupil diameter given an image cropped using the pupil center output by the first network. Across all users, the distribution of absolute errors has a median of 0.36 mm and a 90^{th} percentile of 1.09 mm. According to Meeker et al. [165], the error of PupilScreen's diameter estimation is better than that of manual examination (0.5 mm), but worse than that of a clinical pupillometer (0.23 mm). To determine if the error of the first network leads to greater errors in the second network, I examined the accuracy of the second network given input images cropped around the ground truth pupil center. I found that there was little difference between using the predicted pupil centers and the ground truth pupil centers (50^{th}: 0.36 mm, 90^{th}: 1.19 mm vs. 50^{th}: 0.36 mm, 90^{th}: 1.15 mm). The fact that using the ground truth centers did not improve the accuracy of the pupil diameter estimation may be a byproduct of the fact that the training data was jittered, leading the network to be invariant to exact pupil location.

The Bland-Altmann plots in the bottom half of Figure 5.10 show a different

representation of the diameter prediction errors split across the different iris colors. In all cases, the sequential network architecture tends to overestimate the pupil diameter. If the CNN relies upon convolutional filters that look for edges, overestimation could be happening because those filters are more likely to respond to regions outside of the pupil's actual boundary. The mean pupil diameter errors are +0.24 mm, +0.27 mm, and +0.07 mm for blue, brown, and mixed eyes, respectively. I find that the most extreme outliers belong to a small subset of participants who had particularly dark irises. I believe that this error can be reduced with more training data from participants with similarly dark irises.

Figure 5.11 shows the same performance measures for the fully convolutional architecture. The CDFs at the top of the figure show that the fully convolutional network was generally more accurate than using sequential networks. Across all users, the distribution of Euclidean errors for the pupil center has a median of 0.20 mm and a 90^{th} percentile of 0.50 mm. The distribution of absolute errors for the pupil diameter has a median of 0.30 mm, which is closer to the observed accuracy of a clinical pupillometer than the 0.36 mm median error of the sequential network architecture. Examining the Bland-Altmann plots in Figure 5.11, I find that the fully convolutional architecture tends to underestimate the pupil diameter. The mean pupil diameter errors are -0.11 mm, -0.20 mm, and -0.55 mm for blue, brown, and mixed eyes, respectively. Beyond the inherent differences between the two architectures from a deep learning standpoint, one reason for the improved results could be the fact that explicit morphological operations could be performed on the pixel labels; rather than hoping that the network could learn some attribute in regards to smooth edges, it is easier exercise domain-knowledge and enforce such rules afterwards. The post-processing could also explain why this architecture underestimated diameters; although smoothing can remove protrusions from a jagged pupil boundary estimate, it can also shrink an otherwise correct, smooth pupil boundary estimate.

There is a noticeable difference between the results for different iris colors. For both

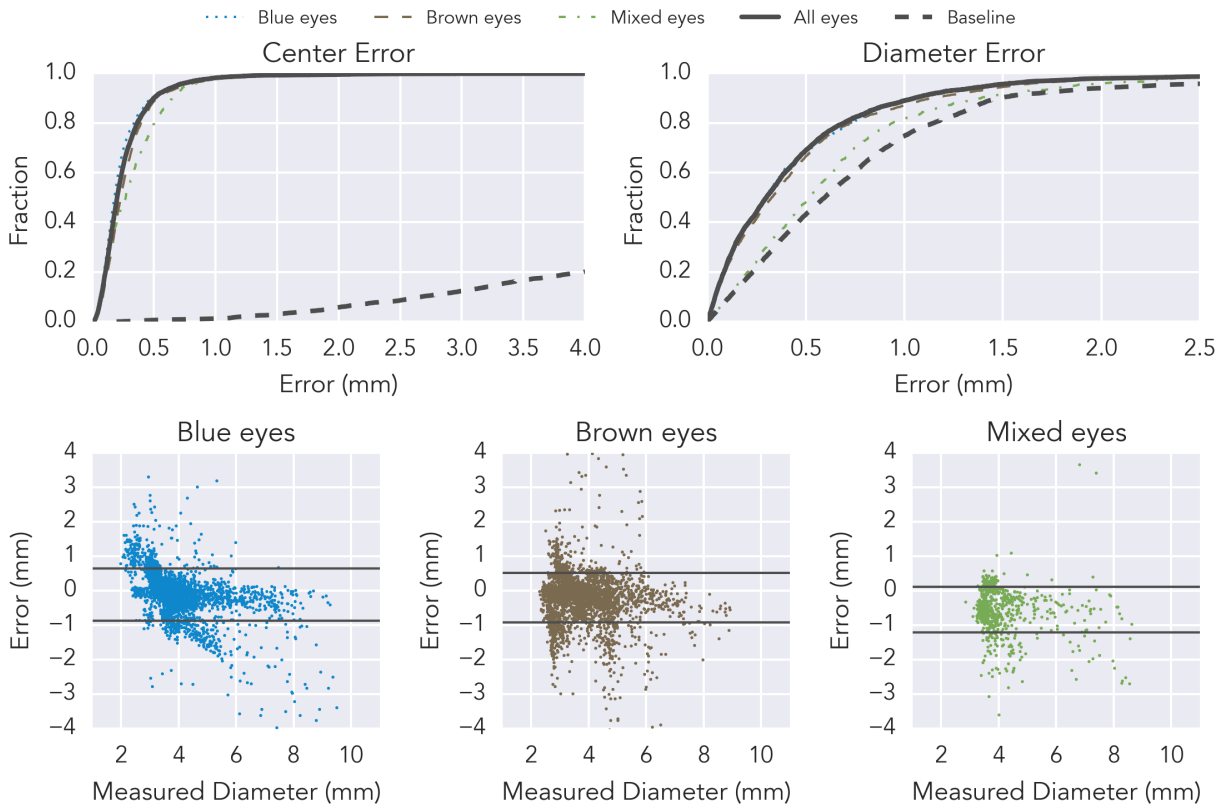

Figure 5.11: The accuracy results for the fully convolutional architecture. **(top-left)** The CDF of the pupil center prediction error. **(top-right)** The CDF of the pupil diameter prediction error. **(bottom)** Bland-Altmann plots showing the residuals of the pupil diameter predictions split across the different iris colors: blue, brown, and mixed from left to right. The black lines indicate one standard deviation from the mean.

architectures, images of brown eyes led to the worst results. The sequential network architecture had a median error of 0.41 mm and a 90th percentile error of 1.19 mm, and the fully convolutional architecture had a median error of 0.33 mm and a 90th percentile error of 1.14 mm. This may be because the boundary between the pupil and the iris is less noticeable for people with darker irises, so the convolutional filters in the networks are less likely to respond to the appropriate regions of the eye. I also hypothesize that this is the reason for why the measured diameter error for brown eyes does not correlate with

the pupil size as it does with the lighter iris colors, a phenomenon noted by Meeker et al. when pupils were manually examined.

5.5.3 Metric Evaluation

The outputs of PupilScreen's networks are irrelevant unless they are combined sequentially in PLR curves. For the sake of brevity, the results from here on out come from the fully convolutional architecture since it was slightly more accurate. To quantify how well the predicted PLR curves track the human-labeled PLR curves, their normalized cross-correlation was calculated. The average normalized cross-correlation across all videos is 0.91. Figure 5.12 compares several examples of PLR curves produced by PupilScreen with ground truth PLR curves from manual annotations.

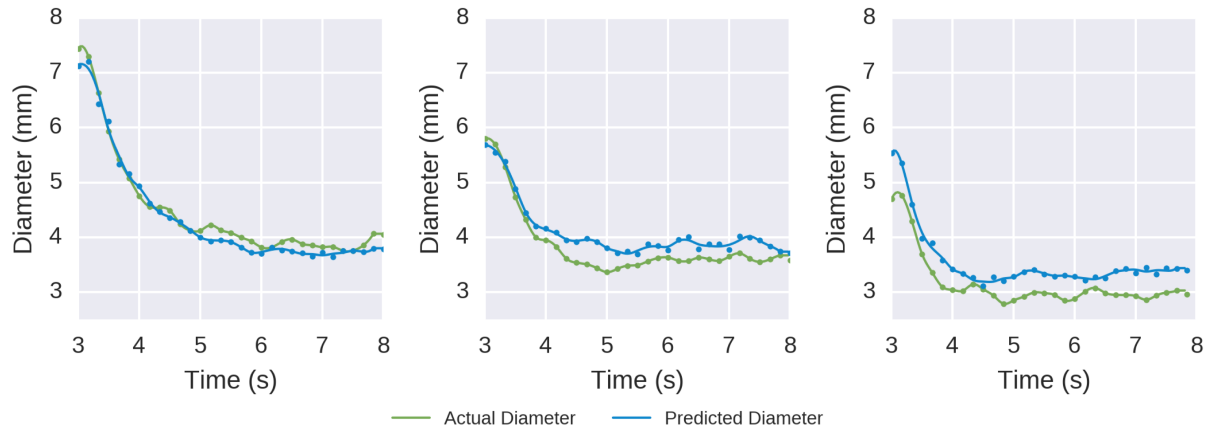

Figure 5.12: Examples of predicted and ground truth PLR curves. **(left)** An example where PupilScreen accurately estimates all PLR metrics. **(center)** An example where PupilScreen accurately estimates the max constriction velocity, but underestimates the constriction amplitude and percentage. **(right)** An example where PupilScreen accurately estimates the constriction amplitude and max constriction velocity, but underestimates the constriction percentage.

Table 5.2 describes how well PupilScreen is able to predict PLR metrics relative to

Table 5.2: PLR metric evaluation

CONSTRICTION AMPLITUDE - mm	
Ground truth range	0.32-6.02
Mean absolute error	0.62
Standard deviation of absolute error	0.72
CONSTRICTION PERCENTAGE - %	
Ground truth range	6.21-62.00
Mean absolute error	6.43
Standard deviation of absolute error	6.74
MAX CONSTRICTION VELOCITY - mm/s	
Ground truth range	1.37-8.99
Mean absolute error	1.78
Standard deviation of absolute error	0.67

those measured from the manually labeled dataset. Table 5.2 also shows the range of those metrics across all participants as a point of comparison for the error magnitude. PupilScreen can track constriction amplitude with a mean error of 0.62 mm, constriction percentage within a mean error of 6.43%, and max constriction velocity with a mean error of 1.78 mm/s. As a point of comparison from the literature, an evaluation of PupilWare by Rafiqi et al. [203] demonstrated that their system tracked constriction and dilation percentages with an accuracy such that 90% of their predictions fell within 10% of the ground truth. However, there are many differences between PupilWare and PupilScreen that make these results difficult to compare. PupilScreen was evaluated on many more participants than PupilWare (42 vs. 9), but the evaluation of PupilWare aggregated a time series of percent change values rather than the single summary statistic like PupilScreen. The two systems are also intended for different applications. PupilWare is designed to track changes in pupil size attributed to varying cognitive load, which tend to be smaller in amplitude than the changes induced in PupilScreen.

Examining the predicted PLR curves further provides insight into the nature of these errors. The center and right plots in Figure 5.12 show cases where a repeated error across frames led to the the inaccurate estimation of some PLR metrics, but not others. In the center, PupilScreen correctly tracks the pupil diameter during constriction, but then overestimates the final diameter of the pupil after constriction. The max constriction velocity is correctly estimated in these situations, but the constriction amplitude and percentage are not. On the right, PupilScreen follows the ground truth PLR curve with a roughly constant offset. This means that although the absolute estimate of the pupil diameter may be off, the change between the minimum and maximum pupil remains unchanged. This behavior only affects the constriction percentage since it relies on an absolute baseline; the constriction velocity and amplitude remain unaffected. Although not shown in Figure 5.12, errors in all three metrics can also be attributed to pupil diameter predictions that deviated from nearby frames in a manner that failed PupilScreen's outlier criteria but were significant enough to create a deflection in the

filtered PLR curve.

5.5.4 Preliminary Clinical Evaluation

To gauge PupilScreen's diagnostic efficacy, I supplemented my dataset with videos from six patients at Harborview Medical Center's trauma ward and neuro-intensive care unit (neuro-ICU). These individuals had sustained significant head trauma, but were stable enough at the time to be recruited for the study. Their doctors and nurses knew beforehand that they had non-reactive pupils. Non-reactive pupils are frequently observed in patients whose condition is unstable, making it difficult to use my research prototype without interfering with the clinician's workflow. As before, three videos were recorded for each patient; however, there were complications in collecting these videos, including the inability of the patients to keep their eyes open and the inability of the clinician to maintain the position of the box while recording the videos. Because of these issues, only 24 of the 36 possible PLR curves (3 videos per patient $\times$ 2 eyes per patient $\times$ 6 patients) were suitable for analysis.

To evaluate PupilScreen's accuracy on non-reactive pupils, I randomly selected one of the folds created during my initial training and analysis. The patient videos were processed using the CNNs that were trained on that fold's training data to produce pathologic PLR curves. An equal number of healthy PLR curves were generated using randomly selected videos from that fold's test set. Using the same network for both sets of videos guaranteed that the PLR curves were generated from networks that were trained on the same data. Figure 5.13 shows examples of both responsive and non-responsive pupils that were collected with PupilScreen. The PLR curves from healthy individuals have a noticeable exponential decay, whereas the PLR curves from the patients do not.

The PLR curves were anonymized, shuffled, and then sent to two clinicians familiar with pupillometery. The clinicians were asked to classify the PLRs as either "responsive"

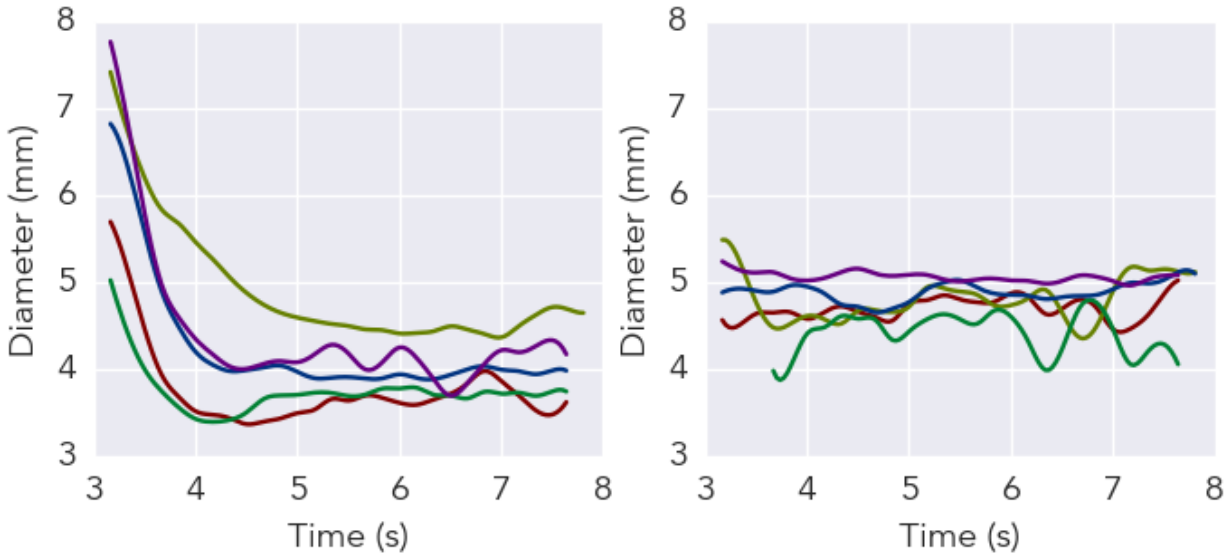

Figure 5.13: A subset of **(left)** responsive and **(right)** non-responsive PLR curves that were shown to clinicians for my preliminary clinical evaluation.

or "non-responsive". They were not told how many curves would be in each category, nor were they shown the video recordings themselves. The first clinician was able to correctly classify every curve in my dataset. The second clinician misclassified one non-responsive PLR curve as responsive. In that particular case, PupilScreen estimated that the person's pupil constricted in a slow and almost linear manner, but by a significant amplitude. The second clinician also misclassified one responsive PLR curve as non-responsive, again, due to the borderline pupil constriction amplitude.

5.5.5 Clinician Feedback

Throughout my design process, I asked clinicians about their personal experiences with pupillometry and for feedback on PupilScreen's design. These clinicians included surgeons, nurses, and other personnel at the Harborview Medical Center's neuro-ICU. Although PupilScreen is proposed as a tool to be used by team doctors and parents,

clinicians who work with TBI are far more familiar with existing pupillometry methods and their tradeoffs and could provide far more insight beyond novelty.

One of the surprising findings early on was that although the clinicians were familiar with the purpose of a pupillometer and its advantages over a penlight test, the pupillometer was hardly used in the clinical setting. The pupillometer was mainly used to track changes in PLR over a long period of time to identify worsening injuries as quickly as possible in otherwise unresponsive patients. For diagnosis or triage, penlights are strongly preferred for their simplicity and ease of access, despite the limited precision and lack of consistency they afford. As one clinician stated, "If whatever you ask an EMT to do adds twenty seconds or so, it's not worth it". In fact, I found that some clinicians use their smartphone's flash instead of a penlight, validating aspects of my idea.

When I asked the clinicians about the prospect of PupilScreen's convenience, they were excited by the idea of a smartphone app that would be in their pockets at all times. Unsurprisingly, clinicians pointed out that the PupilScreen box was still a bulky object that needed to be carried to conduct the test, although some reasoned that it would be far cheaper to place multiple boxes in the neuro-ICU than multiple pupillometers. One clinician recommended a foldable box that would be easier to transport. Another clinician suggested a monocular design that would record one eye at a time; such a system would still require a separate component from the phone, but it would be roughly half the size of the PupilScreen box. The most popular suggestion was a system where no box was required at all. Eliminating the box would make PupilScreen even more convenient than a penlight, but removing the box eliminates control over lighting, which is crucial for ensuring that the pupil is visible and that the light stimulus provided to the eyes is standardized. Nonetheless, I plan on exploring this possibility in the future and address this potential in Section 5.7.

Another issue raised about PupilScreen's design is the difficulty of using PupilScreen on patients who are unconscious. In the sports-related concussion scenario, the cases that most warrant the use of pupillometry are when the patients are conscious and can comply

with most verbal instructions. In the neuro-ICU, penlights and pupillometers are often used on unconscious patients, and clinicians must hold those patients' eyelids open with one hand in order to expose the pupil(s). This is a manageable, but clumsy maneuver to conduct with the PupilScreen box. Manipulating the patient's face in this manner can also allow extra light to seep in from the top of the box, which reduces the control over the lighting within it.

From my interviews, I believe that PupilScreen's design will be suitable for use by team doctors and parents, but requires further improvement for use by EMTs and other hospital clinicians.

5.6 Discussion

My goal was to develop a system that could quantitatively assess the severity of TBIs by measuring a person's pupillary light reflex. Furthermore, I imposed the requirements that the system should be automated and easy to deploy. I believe that PupilScreen is the first step toward these goals. The PupilScreen box allows anyone to use their phone as an inexpensive pupillometer. It does so by blocking out ambient lighting while allowing the smartphone to provide its own light stimulus from the flash. Using two sequential CNNs, PupilScreen measures the pupil center with a median error of 0.43 mm and the pupil diameter with a median error of 0.36 mm. Using a fully convolutional network, PupilScreen achieves median errors of 0.20 mm and 0.30 mm for those same two measures, respectively. Once I found that PupilScreen could track the PLR with reasonable accuracy, I conducted a preliminary clinical trial with six patients who had suffered a TBI. When clinicians were given PLR curves from both healthy and injured individuals, they were almost always able to reach the correct diagnosis.

5.6.1 Hardware

The low-fidelity nature of the PupilScreen box has advantages and disadvantages. The only requirements on the box were that it needed to block out light from the environment and that it allowed the smartphone's flash to illuminate the patient's eyes. A variety of materials for the box could have satisfied these requirements. I 3D-printed the box using PLA plastic for durability over the course of the study. The PupilScreen box could easily be mass-produced using injection molding for similar results. Since the box does not require any embedded electronics outside of the user's smartphone, people can even construct their own PupilScreen box using stiff cardboard. This last idea is particularly enticing because it could allow for the generalization of our system throughout the diverse smartphone ecosystem. The PupilScreen box used in the study was made specifically with iPhones in mind since they have a more unified design. Later models (iPhone 4 or after) have both the camera and flash on the top-left corner at the back of the phone, which lent itself to the design shown in Figure 5.4. Android phones come in all sorts of different configurations and shapes, which would require a dedicated box design for each model or a configurable box to cover all of them.

Beyond the design of the PupilScreen box, the diverse smartphone ecosystem could influence the diagnostic efficacy of PupilScreen, although I believe these effects would be minimal. Different smartphone models may have different flash LEDs, but most are bright enough to cause a similarly significant PLR. PupilScreen could tune its thresholds for various PLR metrics based on information about the flash LED that can be stored in a lookup table. There is larger variation in smartphone cameras across specifications, including sensitivity and resolution. Cameras can respond to various wavelengths of light in different ways. I believe this should have minimal impact on PupilScreen's CNN-based approach since the convolutional filters should still respond in a similar manner if preprocessing or calibration can be employed to standardize input frames. With regards to camera resolution, a higher resolution translates to a higher

pixel-per-mm ratio given a fixed camera placement and focus. A higher pixel-per-mm ratio allows PupilScreen to detect smaller changes in pupil diameter and measure the PLR with increased precision. In cases when the resolution is too low, PupilScreen could incorporate a zooming procedure that maximizes the pixel-per-mm ratio without sacrificing focus. However, too much variability in resolution could lead to issues since the filters in PupilScreen's CNNs have fixed pixel sizes and may be trained to only recognize contours within certain scales.

By relinquishing lighting control to the smartphone, the current PupilScreen design is limited in what kind of responses it can capture. In our evaluation, I only examined pupil constriction, not dilation. This is because there is no intermediate lighting state between the on and off stages of the smartphone's flash, and when the flash is off, the camera cannot see the patient's eyes. Some smartphone models are beginning to provide multiple flash LEDs (e.g., iPhone 6), but I found there was not enough of a difference between them to induce significant pupil diameter changes. Early in our design phase, I briefly experimented with using the smartphone's screen as the lighting source. I decided against this design because most smartphone screens are not sufficiently bright to make the eyes visible within the PupilScreen box. Furthermore, since most front-facing cameras are on the corner of the smartphone, the screen illuminates the patient's face at an angle when the camera is centered between their eyes. This can form a light gradient across the patient's face, or worse, a shadow on an eye, creating undesirable noise in the data.

5.6.2 Software

One might argue that I did not collect enough data to sufficiently train the networks' thousands of parameters. I attempted to mitigate some of these issues by jittering our data during training and starting with a pre-trained network in the case of the fully convolutional architecture; however, I recognize that more data is always better. Beyond collecting more data in the same manner as I have in the past, I plan to incorporate

synthetic datasets, such as SynthesEyes by Wood et al. [260], to develop a more diverse dataset. I may also explore ways of scraping the web for images to further bolster our dataset.

There is more exploration left to be done concerning the optimal CNN architecture for identifying the pupil. One drawback from using CNNs on individual video frames in general is that consecutive frames are treated independently until predictions are combined for the PLR curve. This approach does not account for the fact that the pupil changes size continuously and, therefore, nearby frames should have correlated pupil diameters. PupilScreen uses low-pass filtering to reduce unnecessary variation between nearby frames. Another way to account for frame continuity would have been to use an algorithm that trains on entire sequences, such as a continuous-time recurrent neural network. I chose not to do this because it requires a significant number of examples for both reactive and non-reactive pupils, which would only be feasible with a larger deployment. There is also the possibility that such an approach could bias towards learning the typical PLR, leading to diagnostic false negatives. Although using two sequential CNNs led to slightly worse results, the full range of possible structures for those networks was not explored. As pointed out by Chellappa [36], factors related to network size (e.g., memory footprint, number of parameters, training time) are still an open challenge in the deep learning community. Staying up to date with advancements in that field while focusing on the our specific task will be important for eventually moving PupilScreen to a configuration that does not require a server.

Most of our participants complied with PupilScreen's procedure, meaning that they blinked as little as possible and kept their gaze toward the camera. These constraints are also imposed by pupillometers; if the patient does not comply, the pupillometer rejects the trial and requests a retest. Both pupillometers and PupilScreen currently handle blinking in different ways that lead to similar results. Pupillometers explicitly localize the pupil using infrared light. If they cannot find the pupil, the PLR curve for those frames has a null value. As long are there are not too many null values in the PLR curve,

the pupillometer interpolates the pupil diameter for those frames. PupilScreen does not include an explicit blinking detection step, so all frames are tested through the CNNs regardless of the whether the pupil is visible in them or not. That being said, the CNNs are only trained on images where the pupil is visible, so cases when the pupil is not visible lead to outlier results that are handled through the post-processing described in Section 5.4.4. I found that cases of blinking were not a significant source of error in PupilScreen's results, but a blink detector [170] could be incorporated at the beginning of PupilScreen's pipeline so that irrelevant frames are accounted for sooner.

Handling different gaze directions is a simpler matter for both PupilScreen and pupillometers. Pupillometers fit an ellipse, not a circle, to the pupil. If the ellipse's eccentricity is too low (e.g., its axes are uneven), the frame is rejected just as a frame with a blink. The data for PupilScreen was also originally labeled as ellipses. The elliptical labels were converted to a circular representation where the diameter was defined as the average of the ellipse's axes, so the CNNs are trained to interpret the ellipses in that manner. The maximum of the ellipse's axes could have been a better summary of the pupil since the dimension parallel to the direction of the rotation decreases in size; however, I chose to use the mean as a compromise between this phenomenon and the fact that some pupils have small protrusions along their perimeter that artificially extend their clinically significant boundary.

5.6.3 Future Applications

PupilScreen is primarily targeted toward individuals interested in assessing the severity of head trauma, whether it be a high school coach checking for concussions or an EMT checking the extent of a more general TBI. Zafar and Suarez [264] note that most of the studies involving the diagnostic significance of pupillometry are limited due to small sample sizes. The clinical study I conducted has the same issue since it only included six individuals who had suffered significant head trauma. I was limited to individuals who

were in a stable condition because clinicians were hesitant of introducing yet another instrument into their workflow during time-critical situations. Following their suggestions, I plan to explore the possibility of removing the PupilScreen box. Rather than imagining PupilScreen as an inexpensive pupillometer, removing the box would turn PupilScreen into a more quantitative penlight exam, sacrificing consistency and standardization in favor of convenience. Ensuring that the penlight exam is conducted in a reasonable manner would become the responsibility of the user interface. Visual guides could show an inexperienced user how close the phone should be from the patient's face, and feedback could be provided if the pupils were not sufficiently stimulated by the light.

I believe that by making pupillometry more accessible in this manner, I can enable researchers to reassess previous studies with greater sample sizes. In fact, I plan to conduct a follow-up study looking at the correlation between PupilScreen, a clinical-grade pupillometer, and the tools currently used by American football teams for assessing concussions (e.g.,the King-Devick test and the SCAT). I also plan on examining how our technique can be used to check for other eye-related conditions that may indicate a TBI, such as involuntary eye movement [145] and poor visual tracking performance [158, 159].

5.7 PupilScreen v2.0

My research team and I engaged in conversations with clinical partners to determine the steps that would be needed to deploy PupilScreen in a larger scale study with more cases of TBI. We found that doctors and nurses were reluctant to introduce our prototype into clinics because of the reasons listed in Section 5.5.5. The lack of portability and ease-of-use far outweighed the repeatability and precision that the PupilScreen box enables. Given those issues, I am now working on the next version of PupilScreen that only requires a smartphone. Whereas PupilScreen with a box can be considered an inexpensive alternative to a clinical pupillometer, PupilScreen without a box would be

considered a more precise penlight exam. The version of PupilScreen without the box will be used in a study in Phitsanulok, Thailand where my clinical collaborators will be able to recruit emergency room patients who have suffered from TBIs. Below, I describe the challenges that must be overcome in order to make PupilScreen without the box reliable and the steps I have either taken so far or plan to take in the future to address those challenges.

5.7.1 Challenge #1: Varying Ambient Light

The PLR is not only a function of a person's cognition and ocular motor functions but also light intensity. The PupilScreen box controls the light intensity that is shone on patients' eyes to remove that degree of freedom. PLR metric thresholds are easier to set in that case for distinguishing normal and abnormal responses since the ideal response would look roughly similar in the same lighting environment. Without the box, however, varied ambient lighting introduces an additional parameter that must be considered when classifying PLR curves.

There are two approaches I plan on comparing to address this challenge, both of which involve measuring the amount of ambient light using newer smartphones' light sensor. The first takes advantage of prior work by Pamplona et al. [188], which presents a mathematical model for pupil size as a function of light intensity. The algorithm would generate an expected curve shape for the given lighting environment to set dynamic classification thresholds. The second approach is more data-driven. The light measurements would be combined with the observed PLR metrics as features in a machine learning classifier for separating normal and abnormal responses.

Despite the fact that ambient lighting complicates classification, it also provides an additional measurement opportunity. PupilScreen with the box is limited to only measuring pupil constriction; pupil dilation is impossible to measure because there is not a way of reducing the light stimulus without making it pitch-black inside the box.

Without the box, however, there is sometimes sufficient ambient lighting to see dilation as well. Although we cannot guarantee sufficient lighting to see the pupils in all cases, we plan on exploring dilation in the future.

5.7.2 Challenge #2: Varying Phone Position

The PupilScreen box fixes the position of the smartphone relative to the user's face, which serves two purposes. First, the box ensures that the smartphone is as close as possible to the user's face to increase the resolution of the pupils while keeping both pupils within the camera's view. Second, the mapping from pixels to millimeters remains constant because the scale of objects in the camera's view does not change during the PupilScreen test.

Without the box, there is nothing to constrain the position of the smartphone relative to the patient's face. One small user interface change I added to help in this regard was adding visual guides on the smartphone's screen where the patient's eyes should land within the camera frame. The separation between the guides is based on the average interpupillary distance for humans (62.9 mm) [89]. Even with the guides, we found during pilot testing that both users and patients were prone to moving while the camera was recording. To further address this issue, I have implemented an eye detection module that automatically finds the eyes, crops them out, and feeds those images to the fully convolutional neural network for segmentation. The eye detection module relies on Haar feature-based cascade classification [247] to produce candidate bounding boxes for the eyes within each video frame. When extra candidate boxes are produced, the algorithm eliminates erroneous boxes by leveraging the facts that there should only be one eye in each half of the image and they should occur at roughly the same height. If a candidate box is only produced on one side, the algorithm extrapolates the position of the opposite box using horizontal symmetry. If no candidate boxes are found, the algorithm assumes that the positions of the eyes are within the guides provided on the screen. The eye detection algorithm is improved by the fact that concurrent frames

should have the patient's eyes at roughly the same position. A majority vote is used to resolve conflicts in case an eye's bounding box jumps from one frame to another. Cropping directly around the eyes provides an additional benefit for the neural network training. The original PupilScreen algorithm assumes a broad region-of-interest to ensure that the eyes were always fed into the network; this led to larger input images, and thus a larger network. Using eye detection reduces the size of the input images by a factor of almost $\times 1/2$.

Even after automatically cropping the eyes in the frame, there is another challenge introduced by having a shifting camera. If the initial PupilScreen algorithm were to segment the pupil throughout a video and see the pupil shrink from frame-to-frame, there would be two possible explanations: (1) the pupil constricted or (2) the camera moved away from the patient. Innovations in augmented reality like ARKit[3] have made it possible to measure objects in 3D space while the camera is moving. I considered leveraging ARKit to measure the pupils in millimeters regardless of the smartphone's position. However, informal testing revealed that ARKit does not have the millimeter-level precision needed to accurately report the PLR.

Instead, the new algorithm takes advantage of the fact that the pupils are surrounded by irises that have constant size. The new fully convolutional network is trained to segment both the iris and the pupil simultaneously. The diameter of the iris is measured horizontally since the top and bottom of the iris are occluded by the eyelids, while the diameter of the pupil is measured as before. Rather than reporting the diameter of the pupil in millimeters, the algorithm reports the pupil diameter as a fraction of the iris' diameter. One limitation of this approach is that it complicates comparisons between PupilScreen and existing literature that measures the pupil in millimeters; nevertheless, I feel that it is necessary to take this approach for PupilScreen to be practical.

---

[3] `https://developer.apple.com/arkit/`

5.7.3 *Preliminary Results*

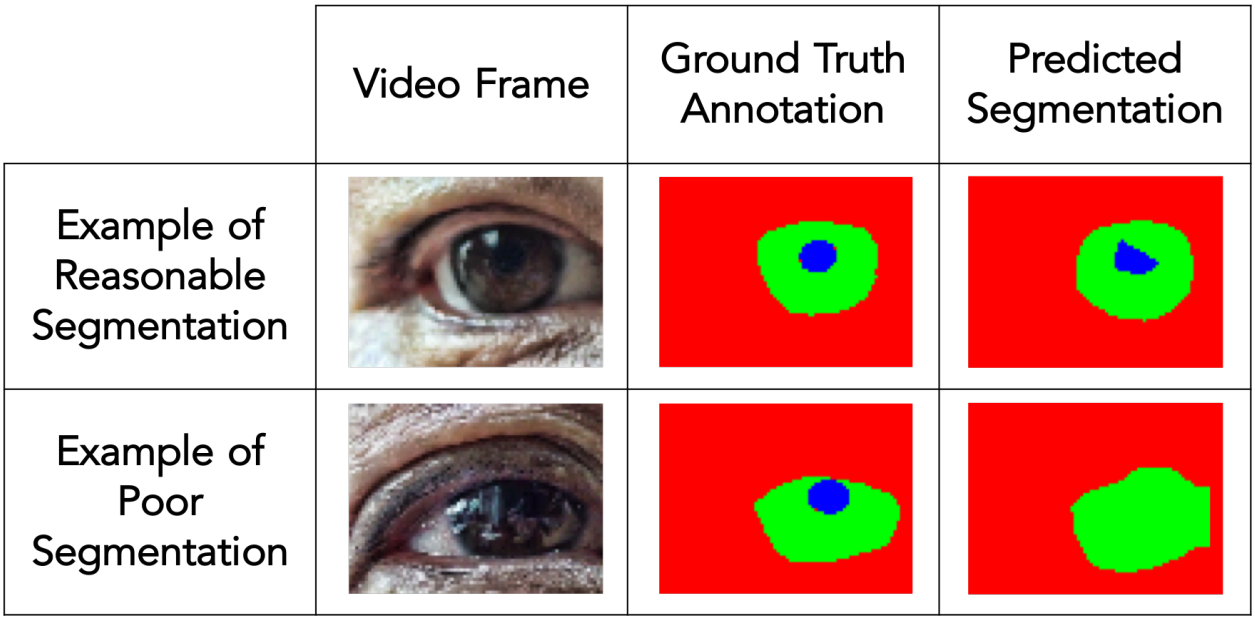

Figure 5.14: Examples of the segmentation results generated using the current version of PupilScreen.

Figure 5.14 shows some preliminary segmentation results using the updated PupilScreen algorithm, including automatic eye detection and segmentation of both the iris and pupil. These eye images were extracted from videos collected in Thailand as part of the training orientation given to nurses who will be conducting our study. Note that these eyes have brown irises, which are the most challenging for pupil detection algorithms. The neural network was trained using annotated frames from 20 videos in uncontrolled lighting environments. This is not nearly enough data to cover the variety of eye colors and shapes that one would expect; however, it was enough to give insight into some preliminary results.

The top row of Figure 5.14 illustrates a case when the algorithm was able to identify both the pupil and the iris. In both the ground truth and predicted images, the iris

segments have roughly the same shape, but the estimated pupil region is only a subset of the actual pupil region. Nevertheless, the pupil-to-iris ratios are similar between the ground truth annotation (31%) and the predicted segmentation (37%). The bottom row of Figure 5.14 illustrates a case when the algorithm was able to identify the iris with moderate accuracy, but unable to identify the pupil at all. Closer examination reveals that the image in that case has a corneal reflection, thereby obscuring the border between the pupil and the eye.

As with all data-driven models, the results will be improved with more training data. The 20 videos we used to train our model are simply not enough to cover the diversity of eye appearances we expect to see in the future. However, another step I will take to improve PupilScreen's robustness is to build a responsive, real-time system that alerts the user if the segmentation algorithm cannot properly see the pupil. If that is the case, the smartphone app would guide the user to either direct the camera at a slightly different angle or to reorient the patient if possible.

Chapter 6

CHALLENGES IN REALIZING SMARTPHONE-BASED HEALTH SENSING

Given my past experience with prototyping smartphone-based health sensing apps, I have come to realize that we are far from seeing them deployed in the real world. In this chapter, I present a list of challenges in bringing smartphone-based health sensing to fruition in today's medical and technological infrastructure.

6.1 Challenge #1: Limitations Fundamental to Smartphones

Most smartphone sensors are primarily focused on improving the user experience. IMUs measure the smartphone's orientation to determine how content should be presented, microphones record audio for communication, and cameras allow users to capture images and videos of their favorite moments. Because these user experiences are currently the primary driving force behind smartphone sales, sensor specifications do not exceed what is necessary to support them.

For example, CMOS image sensors used for smartphone cameras are sensitive to visible and near infrared wavelengths (400-1000 nm). However, most smartphone manufacturers place a thin film on top of the sensor to block infrared light, limiting the spectrum to 400-700 nm to ensure that photographs are visually correct. This design decision for common photography use-cases is counterproductive to specific use-cases like HemaApp that could benefit from an extended light spectrum. The design of the flash LED also poses challenges for both HemaApp and BiliScreen. The LED is intended for flash photography and torch lighting, so it is designed to produce intense light. For BiliScreen, that intensity can cause discomfort to someone who stares directly at the

light. The flash LED and camera also get hot if they are left on for too long, which can cause discomfort while using HemaApp.

6.1.1 Current Approaches

Although the IR blocking film presents difficulties for HemaApp, some IR light can still leak to the camera if enough is shone. The initial study of HemaApp exploited this fact by utilizing a custom IR and visible light LED array with an incandescent light bulb to augment the smartphone's limited spectral range. The study revealed that incorporating the extra LEDs improved the rank order correlation coefficient between HemaApp's estimates and the corresponding blood draw result from 0.69 to 0.82 when compared to only using the built-in white LED. The use of custom lighting is less attractive than being able to use what already exists on smartphones. Conveniently, newer models have an IR time-of-flight autofocus sensor positioned right next to the rear camera. The current Android API does not provide access to the raw data, but rather the end result of an algorithm that estimates distance. This data can be accessed through a custom kernel installation.

Fortunately, smartphone operating systems have begun to give low-level access to some sensors. In the context of HemaApp, standard white-balancing algorithms often suppresses blue and green channel fluctuation because the red channel fluctuation from the blood is so dominant. The Camera 2 API for Android allows for control over such gains, which HemaApp leverages for consistent variation across the color channels. Smartphone operating systems have also begun to provide access to raw image files, which are useful for apps like BiliCam and BiliScreen that require the truest representation of color directly from the camera sensor.

6.1.2 Future Directions

A smartphone operating system that offers more control over sensors and other smartphone components can accelerate exploration at the intersection of health and mobile sensing, especially when developers have access to raw sensor data. For example, the IR time-of-flight sensor can be used as a pulse sensor if an API exposes raw sensor values, avoiding the need for custom kernel solutions that cannot be widely deployed.

A loftier goal would be for smartphone manufacturers and researchers to come together and agree upon a concise set of sensors that together form a "dedicated health sensor". My approach to health sensing has been to push the limits of sensors that already exist on smartphones, yet history has shown that manufacturers are willing to support new sensors if their use has enough value proposition. Apple's M-series coprocessors offload the collection of accelerometer and gyroscope data from the main CPU for gesture recognition even when the smartphone is asleep, and dedicated depth sensors are beginning to appear on newer smartphones for augmented reality applications. Demonstrating the utility of new sensors often requires working with dedicated hardware and then identifying the minimum requirements needed to support the application. This approach can also uncover signals that may not have been discovered otherwise by limiting research to smartphone sensors.

6.2 Challenge #2: Smartphone Heterogeneity

HCI and ubiquitous computing researchers often cite the fact that smartphones are pervasive, but this statement only applies to the general category of smartphones; not all smartphones are created equal. There are multiple smartphone manufacturers (e.g., Samsung, Apple, Motorola) and software operating systems (e.g., Android, iOS), which lead to a diverse smartphone ecosystem. This poses challenges when someone wants to receive FDA approval for an app that relies on the built-in sensors of whichever

smartphone model they happened to use for prototyping. The FDA has spent years devising regulations about dedicated medical devices ranging from MRI machines to blood glucose monitors——devices that are assumed to be static and self-contained, performing only their prescribed function with a fixed hardware and software specification.

The studies presented in this article were conducted using a single smartphone model to avoid cross-device biases. Attaining FDA approval for those apps would require further studies with many different smartphone configurations. Camera-based apps like HemaApp or BiliScreen, for example, would have to work for a number of different camera modules, LEDs, and sensor arrangements. The flow detected by the microphone in SpiroSmart relies on the mechanical transduction of sound, which is affected by the position of the microphone and the physical casing surrounding it. If generalizability is not possible, developers must restrict potential users to a subset of devices or convince manufacturers to fulfill specific hardware and software requirements to support their app.

Figure 6.1: To account for different lighting conditions and camera sensors, both **(left)** BiliCam and **(right)** BiliScreen incorporate paper accessories with colored squares that can be used as calibration references.

6.2.1 Current Approaches

In BiliCam and BiliScreen, smartphone diversity is handled by performing a check on the camera's properties during data collection. Both apps include paper accessories for color calibration: a square card for BiliCam and glasses for BiliScreen (Figure 6.1). These accessories are inspired by a Macbeth ColorChecker, a professional tool for post-hoc color balancing. If an accessory's colored squares appear different from what was expected, whether due to ambient lighting or the camera's sensitivity to various wavelengths, then the same artefact is likely affecting the appearance of the skin. A calibration matrix that corrects the discrepancy can be applied to the rest of the image to standardize colors across images.

6.2.2 Future Directions

Requiring an accessory for standardization adds another potential point of failure that must be FDA-approved. If the BiliCam card's colors fade over time while the card is kept in a person's wallet, the app's performance worsens. The card must also be printed with the same ink and paper used to train the algorithm. In the end, a seemingly trivial addition requires so much consideration that people would probably not be allowed to print the card themselves. Although I posit that such accessories would be far less expensive than a dedicated device, requiring an extra component limits deployability.

Another solution is to create transfer functions based on sensor specifications. When a complete transfer function cannot be generated between sensors, such as two microphones with different sampling rates, compensation mechanisms can be introduced to cater to the common denominator. For developers to find detailed information on a particular sensor, they must currently either disassemble the smartphone and look up the sensor's part number online or dig through the software's kernel and hope the information is documented. Having part numbers accessible in a centralized database or API would help developers understand the capabilities of the

sensors at their disposal and account for the diversity in the market. At the minimum, this would allow developers to restrict their app's use to compatible models or software states.

6.3 Challenge #3: Quality Control of Data Collection Procedures

Clinical tests are conducted under the supervision of a trained professional. With spirometers, for example, pulmonologists can ensure that their patients use the mouthpiece properly by placing their lips around the tube rather than within it. Pulmonologists can also coach patients on how to properly perform the breathing maneuver so that a spirometer can properly measure their peak and total lung function. Going from using a spirometer in a clinic to using SpiroSmart at home removes that safety blanket of quality control. If a user does not push their lungs to the limit while using SpiroSmart, they can be left with nonsensical results that are not representative of their health. Environmental factors are also more controlled in clinical settings. Traditional spirometers are accurate because they measure flow directly and their mouthpieces block out ambient noise. For SpiroSmart, however, the microphone picks up all the sound that occurs during the measurement, adding unexpected noise to the data.

Enforcing quality control is not only important for the immediate results that people receive, but also for algorithm development. The more assumptions that can be made about the signal, the easier it is for a researcher to design a signal processing pipeline or a machine learning algorithm that arrives at an accurate model. Data collection with many edge cases leads to outliers that either impede system accuracy or need to be handled explicitly.

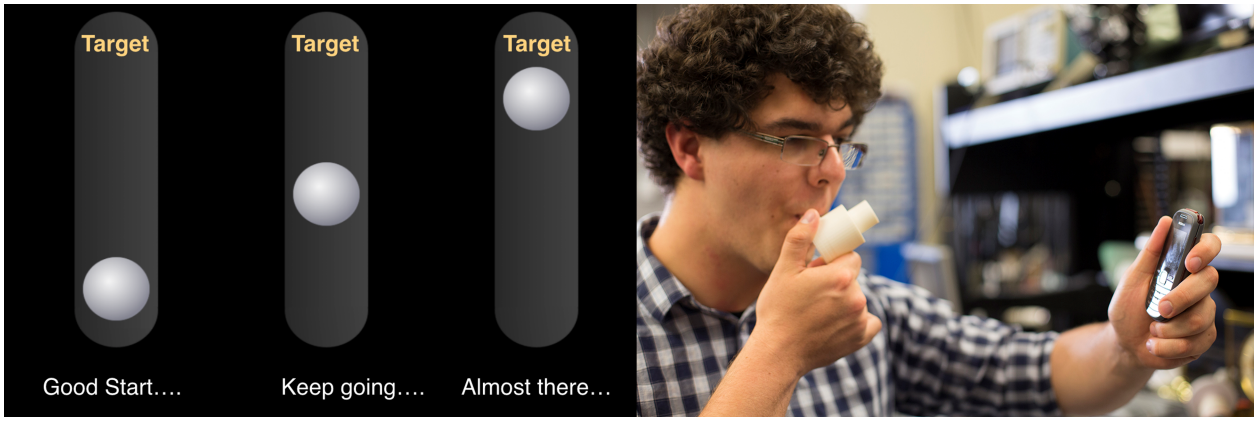

Figure 6.2: **(left)** As a person pushes more air out of their lungs, the ball in the SpiroSmart visualization rises to the top and encourages to the user to continue the maneuver. **(right)** The SpiroSmart vortex whistle can be used to control the diameter of the user's mouth, acting as a flow-to-pitch transducer.

6.3.1 Current Approaches

Automatic checks can be implemented to assess the ambient environment before data collection. For example, the BiliCam and BiliScreen apps check that there are no significant shadows or glare spots obstructing the color references. Real-time visualizations can also be made to coach users on how to improve data quality. The SpiroSmart app includes a dynamic visualization that reacts to the flow rate sensed by the microphone in order to encourage users to exhale as much air out of their lungs as possible (Figure 6.2, left).

When environmental factors or physical abilities impede a person's ability to comply with data collection, inexpensive accessories can improve the process. One of the observations from the first SpiroSmart deployment was that people with severely impaired lung function sometimes struggled to keep their mouths wide open as they performed the breathing maneuver. To help those people, a 3D-printable vortex whistle was developed to hold a person's mouth open like the mouthpiece does for a spirometer

(Figure 6.2, right). The vortex whistle has an extra useful property: the faster the flow of air that enters it, the higher the pitch that leaves it. In other words, the vortex whistle acts as a flow-to-pitch transducer that simplifies the sensing problem.

6.3.2 Future Directions

Another way quality control can be integrated into an app is by adding a classifier that decides whether or not data is "valid" before it goes to the main analysis component. In the case of spirometry, researchers had already categorized the mistakes made during spirometry maneuvers (e.g., coughing during the test, pursing lips while blowing) [17]. Work has been done to train a machine learning algorithm that identifies these errors for spirometer maneuvers in order to provide users with targeted feedback so that they can improve their technique [149]. This approach is currently being expanded to SpiroSmart, as well.

6.4 Challenge #4: Data Interpretation for Untrained Users

The acceleration of hypochondria due to information available on the Internet, also known as cyberchondria [255], is likely to be exacerbated by ubiquitous medical testing. Using BiliScreen as a worst-case scenario, users could interpret a positive test result as a pancreatic cancer diagnosis. However, not everyone with an elevated bilirubin has pancreatic cancer, and not everyone with pancreatic cancer has an elevated bilirubin. Even if users can internalize this subtlety, false positives and false negatives have significant repercussions, whether it be undue stress or a missed diagnosis.

Doctors receive years of training on how to apply Bayesian reasoning when accounting for a test result in the diagnostic process. This procedure requires calculating the patient's pre-test probability of having the condition and then updating that probability according to the accuracy and result of the test. Calculating the prior probability requires knowing the prevalence of the condition and the specific risk factors

that may increase a person's likelihood of having the condition, such as family history and environmental factors. Updating to a post-test probability given a positive test result entails calculating the positive predictive value (PPV) of a test——how often people with a positive test result actually have the medical condition. Calculating PPV requires knowing the test's sensitivity (SNS), the test's specificity (SPC), and the prevalence of the condition (P_0):

$$PPV = \frac{SNS \times P_0}{SNS \times P_0 + (1 - SPC) \times (1 - P_0)} \tag{6.1}$$

This calculation does not always lead to intuitive results. A test with a sensitivity and specificity of 80% for a disease that occurs in 15% of the population will have a PPV of 41.3%. A similar test for a disease that occurs in 5% of the population will have a PPV of only 17.4%. In both cases, the test performs worse than random chance despite having a seemingly decent accuracy.

Test results are never black-and-white; all models have uncertainty bounds that complicate decisions. For example, the current state of BiliScreen has a mean error of -0.09 $\pm$ 2.76 mg/dl. This is reasonable for a disease management scenario when a person's bilirubin may vary between 5-20 mg/dl. For a diagnostic scenario, where the threshold for concern is around 1.3 mg/dl, it is debatable whether or not a test result of 2 mg/dl should be considered elevated.

6.4.1 Future Directions

If smartphone-based health sensing apps are going to be freely distributed to the general public rather than prescribed and supervised by trained physicians, the routine of estimating a post-test probability should be as automated as possible. Apps should be able to calculate a pre-test probability by collecting risk factor information. Family history and demographic data can be explicitly recorded through digital forms. Sensors can also be used to infer risk factors. As an example, GPS data could reveal that a person is at a higher risk of a lung condition because of poor local air quality.

The weighing scale provides an interesting study of how important the presentation of results can be to the decision-making process. All scales have uncertainty, yet people tend to fixate on the number they see. Weight is also a function of how much the person is wearing and how much they ate and drank before the measurement. Kay et al. found that people often forget these factors, leading to stress over negligible weight changes [119]. One scale design they suggest graphically emulates a traditional analog scale with exaggerated needle movement to reflect uncertainty. Kay et al. also propose an "always-on" scale design that accounts for daily variance and incorporates information through low burden question prompts so that measurements can be automatically adjusted closer to their true value. Researchers in the machine learning community have actually trained models that learn how different clinical measurements vary over time to help clinicians identify high-risk patients [16, 217]. The same models could be used to help users extrapolate reasonable trends in their data if they feel the need to do so.

Chapter 7

A SURVEY INSTRUMENT FOR EVALUATING EARLY-STAGE UBIQUITOUS HEALTH SENSING TECHNOLOGIES

Early-stage research for a new ubiquitous health-screening technology often focuses on a subset of technology features, namely sensing accuracy or interface design. However, other factors become equally important to how a person perceives a technology. As a researcher wants to translate their technology from research into practice, they might ask questions like:

- How can the interface design and instructions make end-users confident that they will be able to conduct data collection properly?
- What kinds of results should be shown to end-users?
- Should positive and negative results be presented differently?
- How much technical information about the technology should be available to end-users, if any at all?
- Would providing specific information about the target medical condition sway end-users' decision-making?

Exhaustively implementing and evaluating all possible combinations of features to answer such questions can be a burdensome process, especially when it is done through a functional prototype. Evaluation methods like paper prototyping aim to reduce engineering effort for gathering feedback, yet some of a technology's credibility can be lost if the prototype is not sufficiently refined; awkward interface interactions and

hand-drawn sketches can distract end-users, causing them to react differently with such a prototype than they would with a final technology.

As a step towards addressing Challenge #4 in Section 6.4, I contribute a survey instrument that can be used by researchers who intend to translate a *ubiquitous health-screening technology* into practice. I define ubiquitous health-screening technologies as tools that support end-user decision-making in regards to seeking or ignoring treatment. My survey gauges two intertwined outcomes: (1) people's willingness to use the technology (i.e., its *acceptability*), and (2) how the technology might affect people's decision-making (i.e., its *effectiveness*). Giving a ubiquitous health-screening technology to individuals who do not have a sophisticated knowledge of proper diagnostic decision-making can lead to undesirable outcomes. Falsely leading a person to believe they have a health issue can lead to unnecessary stress, while falsely leading them to believe they do not have a health issue can inhibit timely treatment.

I use the Health Belief Model (HBM) [102, 109] to provide a common language with which UbiComp and HCI researchers can evaluate and examine potential ubiquitous health-screening technologies. The survey itself presents respondents with a hypothetical scenario regarding their health and probes constructs under the HBM. The survey then introduces a hypothetical health-screening technology that claims to screen for the target medical condition and asks the respondent how the technology would change their answers, if at all. Researchers can modify features of their technology in screenshots or text descriptions within the survey instrument rather than building multiple versions of a prototype, lessening the required engineering burden relative to pilot testing. The results of the survey are analyzed using structural equation modeling (SEM) to determine the importance of manipulated variables.

To demonstrate how my survey instrument can be used, I focus specifically on health-screening applications (apps) that use the built-in sensors on smartphones (e.g., accelerometer, camera, microphone) to measure a symptom. I used a formative study with 96 online respondents to select scenarios and apps that were both believable and

distinct from one another. I then deployed my survey instrument to 263 online respondents, varying the types of scenarios and the classification accuracy of the apps they were shown. After using SEM to verify that responses aligned with my expectations from the HBM, I used SEM again to uncover interactions with my manipulated variables. I expected to find that respondents would be more willing to use sensor-based health-screening apps with higher reported accuracy, and my data confirmed that fact. However, increased accuracy did not always lead to an increased likelihood in changing a person's course of action. When respondents were shown positive test results for a serious medical condition, they were more willing to take health-promoting actions regardless of the apps' reported accuracy. When respondents were shown negative test results for common or socially stigmatizing medical conditions, they were less willing to take health-promoting actions regardless of the apps' reported accuracy.

My research contributes:

1. A survey instrument that can be used to evaluate the perception of a ubiquitous health-screening technology at the early stages of its development (Chapter A),

2. A confirmatory analysis of responses from 263 online participants demonstrating that my survey instrument aligns with expectations based on the HBM, and

3. An exploratory analysis of the same responses that uncovers the potential effects that the medical condition in question and an app's classification accuracy can have on acceptability and effectiveness.

I conclude by discussing opportunities for researchers to use and extend my instrument in future research to better understand people's perception of ubiquitous health-screening technologies.

Table 7.1: The constructs of the HBM and their definitions.

HBM Construct	**Definition**
Perceived Seriousness	A person's subjective assessment of the severity of the health problem and its potential consequences
Perceived Susceptibility	A person's subjective assessment of their risk of developing the health problem
Perceived Benefits	A person's subjective assessment of the value in taking a certain action
Perceived Barriers	A person's subjective assessment of the obstacles to taking a certain action
Modifying Variables	Individual characteristics (demographic, psychosocial) that can impact a person's perception of a health problem
Self-Efficacy	A person's subjective assessment of their ability successfully perform a behavior
Cues to Action	Internal or external triggers that prompt a certain action

7.1 Related Work

I use the HBM as the theoretical foundation of my survey instrument. I therefore describe the HBM in detail below. I then discuss methodologies for the support of health-related decision-making and for evaluating the perception of sensor-based technology.

7.1.1 The Health Belief Model

The Health Belief Model (HBM) was first developed in the 1950s by a group of social psychologists at the US Public Health Service to explain the failure of tuberculosis screening programs [102, 109]. Table 7.1 lists the definitions of the HBM's constructs, and Figure 7.1 shows how they are related. The HBM posits that a person will undergo an

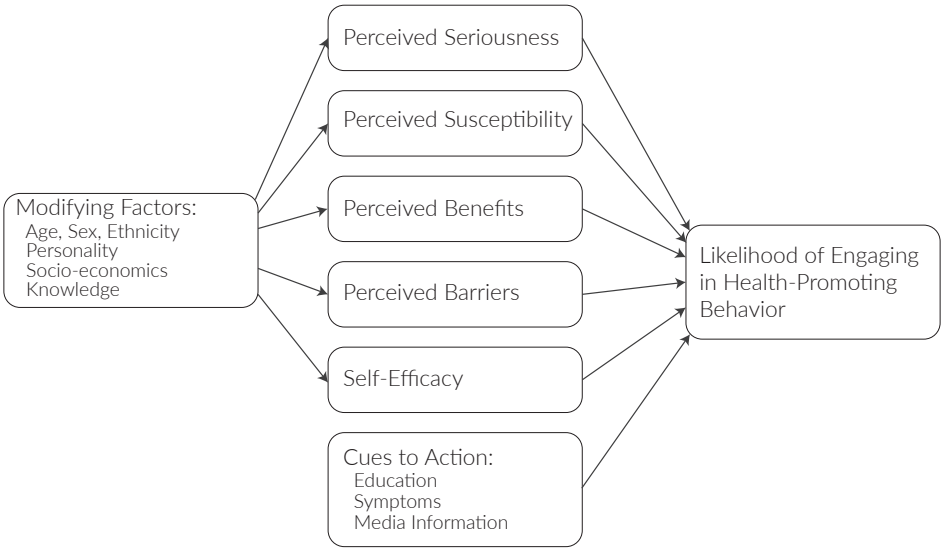

Figure 7.1: The Health Belief Model

.

action to improve or maintain their health if the perceived seriousness and perceived susceptibility of the health problem (together known as the perceived threat) combined with the perceived benefits of the action outweigh the perceived barriers to that action. All of these constructs are affected by modifying variables—demographic information that can influence a person's decisions. For instance, someone who is well-educated may understand the benefits of early screening, and a person who is wealthy may not view the cost of a screening exam as a burden. Decisions are made when there is at least one cue to action. The cue can be internal (e.g., discomfort, fear due to family history) or external (e.g., appointment reminder, advertisement).

The HBM was originally intended for one-time actions like screening exams. However, it has also been applied to actions that require adherence, such as diet modification or smoking cessation. In addition to the initial barriers that may impede a person's ability to take such actions, the person must also believe in their own ability to successfully maintain the behavior; for this reason, self-efficacy was later added to the

HBM [210].

Past studies in medicine and psychology have applied the HBM in various health contexts. For example, Wagner et al. [249] used the HBM to explore the perception of vaccines in China. They found that caregivers were more likely to get their children vaccinated for measles than pneumonia due in part to the higher perceived benefits of receiving the measles vaccine. Champion [34] used the HBM to investigate the factors that influence the frequency of breast self-examination. She found that lower perceived barriers, higher perceived susceptibility, and higher familiarity with breast cancer were all correlated with more frequent breast self-examination; she also found that women who received instruction from their doctor or nurse tested themselves more frequently. My survey instrument provides a standardized platform with which researchers can conduct investigations like these to determine what factors play a role in how people perceive and react to health-screening technologies.

Of course, the HBM is not without criticism. Taylor et al. [230] and Azjen [2] both argue that the HBM is specifically framed around health, but frameworks like the theory of planned behavior (TPB) [69] and the transtheoretical model (TTM) [199] can be applied to other behavioral domains. Having been more broadly utilized, more generalizable evidence has been generated to support the TPB and the TTM. Nevertheless, I am comfortable using the HBM because my focus is strictly on health-related outcomes. Another criticism of the HBM is that it has many constructs with inconsistent definitions, leading to weaker predictive power [230, 7, 97]. Such criticism calls for an instrument that standardizes the assessment of HBM constructs for a broad, yet defined category of interventions in order to generate greater confidence in their specification. My survey aims to fill that gap by providing such an instrument for ubiquitous health-screening technologies.

7.1.2 Evaluating Health-Related Decision-Making Support Technologies

To the best of my knowledge, there has not been prior commentary on evaluation methods for health-related decision-making technologies, but there has been such commentary in the related field of behavior change. Behavior change aims to change a person's habits to prevent disease, whereas decision-making support focuses on the similar goal of getting a person to take a single health-promoting action (e.g., going to the doctor, stopping drinking coffee). Klasnja et al. [123] provide a thorough meta-analysis on different evaluation approaches for health behavior change, including interviews, field studies, and randomized control trials. They come to the conclusion that system evaluations should be tailored to their specific intervention strategies (e.g., self-monitoring, conditioning, tunneling [72]). Although Klasnja et al.'s commentary concentrates on evaluation strategies for after a technology is ready to be deployed to end-users, their call for additional evaluation strategies motivates my survey instrument for early-stage technologies.

Hekler et al. [100] urge HCI researchers to utilize and contribute to behavioral science theories. In particular, Hekler et al. call for the development of new strategies for investigating design recommendations that balance abstraction with contextual relevance. They note that many design guidelines for behavior change technologies are often tied to assumptions about the specific technology that was studied, leading to findings that are less generalizable than intended.

One way to provide abstraction is through vignettes: brief, carefully written situations that include a subset of key features to simulate a real-world scenario [3, 9]. My survey instrument uses hypothetical scenarios and technology descriptions to probe people's decision-making; however, I am not the first to do so. Evans et al. [65] and Bachmann et al. [10] both provide systematic reviews on this field of research. Two of the prominent vignette-based methods they describe are conjoint analysis [92] and judgment analysis [95, 44]. In conjoint analysis, participants are asked to rank or select among

different versions of an object with slight variations across a feature set. As more of these decisions are made, the influence of each feature on the participant choices can be elicited. As an example of health-related conjoint analysis, Ryan [216] used conjoint analysis to examine the values that are important to people pursuing in vitro fertilization. In judgment analysis, participants are asked to decide whether they would take action in a series of scenarios with different features. Participant decisions are compared to the optimal decisions according to an oracle, producing correlations between the weighting of the features in both cases. As an example of health-related judgment analysis, Kee et al. [120] used the method for evaluating prioritization decisions within a dialysis program.

My work diverges from existing vignette-based methods in several ways. First, my survey instrument not only elicits preferences between different feature combinations, but also examines how those features influence people's health-related decision-making. Second, I do not assume that an optimal decision exists for my hypothetical scenarios. The fact that a person may change their course of action at all is an interesting result that I believe should be studied further.

7.2 Survey Instrument Design

My survey instrument elicits measurements of HBM constructs for hypothetical scenarios involving ubiquitous health-screening technologies. I describe the structure of the survey instrument in this section. An abridged version of the survey instrument itself, used for the analysis I conduct in Section 7.5, can be found in Chapter A. For the rest of this paper, I focus on sensor-based health-screening apps as a specific instantiation of ubiquitous health-screening technology to illustrate a particular use of my survey instrument. Note that italicized terms in the following sections designate variables that are examined in the analysis.

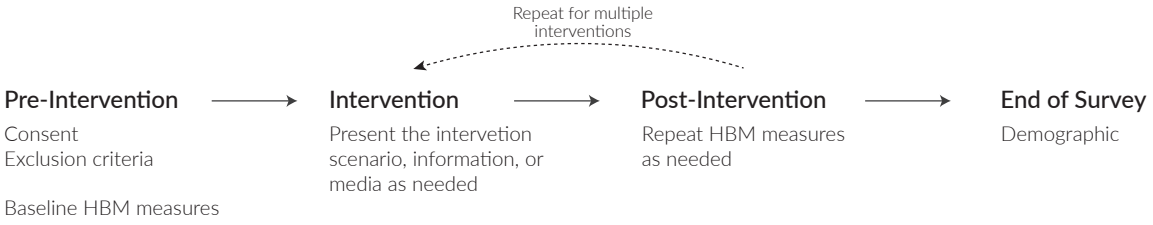

Figure 7.2: The organization of my survey instrument.

7.2.1 Design

The progression of my survey instrument is illustrated in Figure 7.2. My survey starts by asking the respondent about their familiarity with the medical condition they will read about in the survey and with the hardware platform associated the associated hypothetical ubiquitous health-screening technology. Because my survey demonstration involves sensor-based health-screening apps, I ask about the respondent's familiarity with smartphones.

Respondents are then presented with a hypothetical scenario where they are asked to consider that they may have been afflicted with a medical condition. This prompt serves as the cue to action. For example, a scenario about a sinus infection could be presented as follows:

> *A number of your friends have recently stayed home sick with a sinus infection. You have started to have a "stuffed-up" nose and pain in your sinuses as well. After looking up information online, you now suspect that you might be developing a sinus infection.*

After reading the scenario, the respondent is asked to complete an instructional manipulation check [183] that involves selecting symptoms that are associated with the

Table 7.2: The set of questions used to probe the HBM constructs.

HBM Construct	**Survey Question**
Perceived Seriousness	If you had [[medical condition]] in this scenario, how impactful do you believe it would be on your long-term health?
	If you had [[medical condition]] in this scenario, how impactful do you believe it would be on your finances?
	If you had [[medical condition]] in this scenario, how impactful do you believe it would be socially and/or professionally?
Perceived Susceptibility	How likely do you think you are to have [[medical condition]] in this scenario?
Perceived Benefits	How beneficial do you believe each of these actions would be towards helping you recover from your symptoms?
Perceived Barriers	How easy do you think it would be for you to take each of the following actions to help you recover from your symptoms?

described medical condition. Besides checking that the respondent actually read the scenario, the instructional manipulation check forces the respondent to spend extra time reflecting on the scenario.

The respondent is then asked to answer a series of questions related to their general perception of the scenario. The questions probe *PerceivedSeriousness*, *PerceivedBenefits*, and *PerceivedBarriers* (Table 7.2). *PerceivedSeriousness* is broken into three questions because a person may be concerned about how a medical condition impacts different aspects of

their life: their health, finances, and social standing. *PerceivedBenefits* and *PerceivedBarriers* each correspond to a single question, but are asked for various potential actions. All of the responses are recorded along a 7-point scale.

On the next page, the respondent is asked about their *PerceivedSusceptibility* to the medical condition, which is recording along a 7-point scale. The respondent is also asked whether or not they would take various actions as a series of yes-or-no questions—a variable I call *ActionTaken*. The actions can vary depending on the target medical condition, but can include options like scheduling an appointment with a doctor or searching for information online. The respondent is allowed to take none of actions or multiple actions, if they so choose.

After a respondent has reported which actions they would take, they are informed about a sensor-based health-screening app that claims to detect the target medical condition. The text includes a high-level description of what symptom the app is detecting, how the app conducts the measurement, and the source of the app itself. Continuing with the sinus infection example (note that the respondent's phone company is automatically filled in using data from the survey instrument's screening questionnaire),

> *A smartphone app named SinusCheck analyzes the sound your nose makes as you inhale to determine whether or not it is congested due to a sinus infection. To use the app, you are asked to inhale through your nose close to the smartphone's built-in microphone. The app guides you through the recording process so that it can hear the sound properly.*
>
> *SinusCheck comes with your smartphone by default as part of a new mobile health initiative by [[Phone Company]]. SinusCheck provides text-based and audio-based instructions to help you perform the test. The app also checks that the test was performed correctly. You can repeat the test until the app determines the*

image to be "valid". The results of the test are available instantly.

After reading the app description, the respondent is asked whether or not they would use the app on a 7-point scale of *AppInterest*. If the respondent says that they would use the app beyond the neutral score, they are taken to two new pages that ask the respondent how they would react to "normal" and "abnormal" test results. For clarity, the remainder of this paper refers to "normal" as a positive test result and "abnormal" as a negative test result; however, I showed the former in the survey instrument because test results are never definitive. The order between the two test results is randomized. I posit *PerceivedSeriousness* is only dependent on a person's perception of a medical condition and should not change because of a test result. Similarly, *PerceivedBenefits* and *PerceivedBarriers* are appraised characteristics of the actions and also should not change because of a test result. Therefore, when asking the respondent to react to the test results, I only repeat the questions related to *PerceivedSusceptibility* and *ActionTaken*.

Given the respondent's projected course of action before and after test results, I can generate outcome variables for each action that describe whether or not the app would have changed the respondent's plan. *ActionChangePositive* is true whenever the respondent would not have taken an action before a positive test result and would have taken an action after it, false whenever the respondent would not have taken an action before or after the positive test result, and undefined otherwise. Inversely, *ActionChangeNegative* is true whenever the respondent would have taken an action before a negative test result but would not have taken an action after it, false whenever the respondent would have taken an action before and after the negative test result, and undefined otherwise.

At the end of the survey instrument, the respondent is asked for information related to *ModifyingVariables* within the HBM (Table 7.3). The demographic *ModifyingVariables* capture aspects of the respondent's background and living circumstances. The

Table 7.3: Modifying Variables.

Modifying Variable Type	**Survey Question**
Demographic	Gender
	Age
	Race/Ethnicity
	Marital status
	Children
	Country of residence
	Education
	Estimated household income
Smartphone-Specific	Opinion of smartphone vs. clinical test regarding time to results
	Opinion of smartphone price vs. clinical test regarding price
	Opinion of smartphone vs. clinical test regarding privacy
	Smartphone brand
	Years with smartphone
Experience-Specific	Familiarity with the medical condition
	Number of statistics courses
	Frequency of statistics usage
	Numeracy

smartphone-specific *ModifyingVariables* probe the respondent's experience with smartphones. The questions asking for the respondent's opinion of smartphones versus clinical tests are answered on a 7-point scale. The experience-based *ModifyingVariables*

measure the respondent's perceived knowledge of the various skills that are required to interpret a diagnostic test result rationally, including statistics and familiarity with the condition. To measure statistical experience, the respondent is asked how many statistics courses they have taken, how often they use statistics, and the Berlin Numeracy Test [40]—a word problem that challenges a person's ability to reason about numbers. Note that familiarity with the medical condition is probed at the beginning of the survey, which is necessary because the respondent is told about the condition's symptoms within the survey itself.

7.2.2 Summary

To summarize, responses to each of the HBM constructs are recorded along a 7-point scale. *PerceivedSeriousness*, *PerceivedBenefits*, and *PerceivedBarriers* are recorded once, before the app is described. *PerceivedSusceptibility* is recorded three times: before the app is described, after a positive test result, and after a negative test result. Four outcome variables are recorded throughout the survey instrument: (1) *AppInterest*, the likelihood that the respondent would use the app along a 7-point scale; (2) *ActionTaken*, whether or not the respondent would take action as yes-no responses; (3) *ActionChangePositive*, whether a person who was not going to take action would change their mind based on a positive test result; and (4) *ActionChangeNegative*, whether a person who was going to take action would change their mind based on a negative test result. The latter three variables are recorded per action.

7.3 Research Questions

In this work, I use my survey instrument for both confirmatory and exploratory investigations. I enumerate the questions behind these investigations below.

7.3.1 Confirmatory

RQ.1 Does my survey instrument follow the expectations of the HBM for a person's likelihood of using a sensor-based health-screening app?

RQ.2 Does my survey instrument follow the expectations of the HBM for a person's intended course of action?

According to the HBM, higher *PerceivedSeriousness*, *PerceivedSusceptibility*, and *PerceivedBenefits* should increase a person's likelihood of taking health-promoting actions, and higher *PerceivedBarriers* should decrease a person's likelihood of taking action. My survey instrument needs a firm theoretical base in order to examine how manipulated variables affect respondent decision-making. Therefore, my confirmatory investigation is needed to support my expectations of the HBM.

7.3.2 Exploratory

RQ.3 Does **(a)** the target medical condition, **(b)** the app's sensitivity, and **(c)** the app's specificity affect a person's likelihood of using a sensor-based health-screening app?

RQ.4 Can the result of a sensor-based health-screening app change a person's intended course of action? If so, how is that affected by **(a)** the target medical condition, **(b)** the app's sensitivity, and **(c)** the app's specificity?

I manipulate three variables in my exploration. The first is the target medical condition. The HBM dictates that medical conditions with increased *PerceivedSeriousness* are more likely to result in a person taking health-promoting action. A person's likelihood of taking action should also be increased by a positive test result, seeing as how it serves as a cue to action according to the HBM. However, it is unclear how the two constructs interact with one another. The second and third variables I manipulate relate to the accuracy of the sensor-based health-screening app. People will always

prefer classifiers with higher accuracy, but people internalize a trade-off between false positives and false negatives that varies from scenario to scenario [118]. Thus, I break down accuracy into sensitivity (i.e., the percentage of sick people who are correctly identified as having the condition) and specificity (i.e., the percentage of healthy people who are correctly identified as not having the condition).

7.4 Scenario and App Selection

Before I could explore the aforementioned research questions, I first needed to create prompts for health-related scenarios and corresponding sensor-based health-screening apps that would be sufficiently believable and distinct from one another. I selected these prompts using an abridged version of my survey instrument, which I describe below.

7.4.1 Participants

I recruited survey participants through Facebook, Reddit, and a mailing list within the Institute of Translational Health Sciences (ITHS), a center sponsored by the NIH's Clinical and Translational Science for connecting clinicians, patients, and other communities throughout the northwest United States. Respondents who reported no smartphone experience were excluded from participation. Respondents who completed the survey were eligible for a raffle in which 1-in-20 people would win a $20 Amazon gift card. In total, 96 respondents completed the survey from start to finish. A subset of their demographic information is provided in Table 7.4.

7.4.2 Apparatus

To select my scenario and app prompts, I deployed an abridged version of my survey instrument (Figure 7.3). Respondents were not asked about how they would react to positive and negative test results. Instead, they were asked to rate the plausibility of the scenario (*ScenarioPlausibility*) and the plausibility of the app (*AppPlausibility*) on a 7-point

Table 7.4: Respondent demographics for the scenario and app selection survey.

Survey Demographics (N=96)	
Source	Facebook (56), ITHS (37), Reddit (3)
Gender	Male (26), Female (68), Transgender Male (1), Gender Variant / Non-conforming (1)
Age	18-24 (42), 25-34 (38), 35-44 (8), 45-54 (4), 55-64 (3), 65+ (1)
Country of residence	United States (93), India (2), United Kingdom (1)
Smartphone operating system	iOS (60), Android (36)
Self-reported smartphone experience	Expert/Advanced (60), Intermediate (34), Novice/Beginner (2)

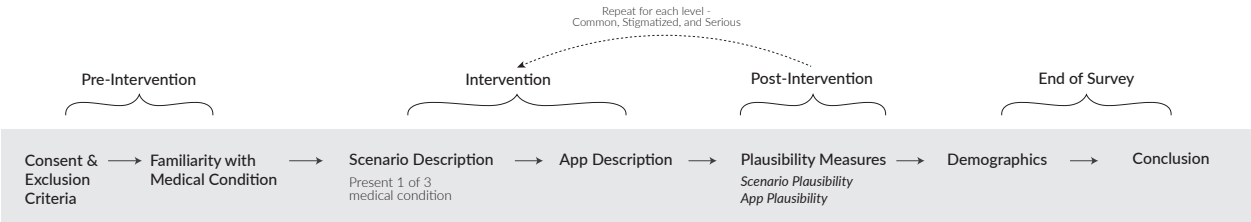

Figure 7.3: The structure of the survey given to respondents to select the most believable scenarios and apps. The structure builds on my survey instrument (Figure 7.2), including many different target medical conditions and excluding the notion of test results.

scale.

To create the scenario prompts and app prompts, I selected three categories of medical conditions that I believed could elicit different reactions from people: *Common* conditions, *Serious* conditions, and *Stigmatizing* conditions (Table 7.5). The categories are neither meant to be comprehensive nor definitive, but merely a formalized effort towards investigating different situations. Many conditions could have fallen within

Table 7.5: The categories of medical conditions that were explored through the survey.

Category	**Characteristics**	**Medical Conditions**	**App Inspiration**
Common	Relatively well-known; only requires short-term treatment; infectious	Sinus infection	Chandra et al. [35]
		Strep throat	Nall and Charles [176]
		Pink eye	Bhadra et al. [19]
Serious	Possibly fatal; requires long-term treatment	Pancreatic cancer	Mariakakis et al. [155]
		Skin cancer	Wadhawan et al. [248]
		Anemia	Wang et al. [252]
Stigmatizing	Could lead to uncomfortable social interactions if discovered by someone else	Halitosis	Seshan and Shwetha [220]
		Irritable bowel syndrome	Lewis and Heaton [140]
		Psoriasis	Shrivastava et al. [223]

these categories. I used two criteria that led to my final selections: (1) the condition had to involve a symptom that a person could either recognize with their senses or perceive within their body, and (2) the condition had to involve a symptom that could possibly be detected with a sensor-based health-screening app using standard built-in smartphone sensors. The apps corresponding to the conditions I selected did not exist, but most were inspired by past publications in mobile health research.

There are also many possible health-promoting actions a person might take based on

Table 7.6: The list of health-promoting actions that were proposed for each medical condition category.

Action	**Common**	**Serious**	**Stigmatizing**
Schedule an appointment with your doctor/physician	✓	✓	✓
Contact your doctor/physician for advice	✓	✓	✓
Stay at home and avoid contact with other people	✓	✗	✓
Purchase over-the-counter medication	✓	✗	✓

a health-screening app, but I restricted my studies to the four listed in Table 7.6. One reason I selected these particular actions was because they can typically be taken within the same day of receiving a test result. Not all actions make sense for all kinds of medical conditions; for instance, there is no over-the-counter medication that can be purchased for most *Serious* medical conditions. Therefore, not all actions were shown for each scenario.

Out of the 169 respondents who opened the survey, 96 completed it (56.8%), 51 partially completed it (30.2%), and 22 were disqualified because they did not own a smartphone (13.0%). Ignoring the two cases when participants completed the survey a day after starting it, the median survey completion time was 11.0 minutes. The average completion time was 13.4 ± 10.9 minutes.

7.4.3 Design and Analysis

The survey was deployed in a 3×3 nested factorial design. The within-subject factor was the different categories of medical conditions (*ConditionType*), while the across-subject factor was the specific medical conditions within the categories (*Condition*). In other words, each respondent was randomly shown one medical condition from each category. The assignment of the conditions was counterbalanced, and the presentation

order of the conditions in the survey was randomly shuffled. Responses from respondents who failed the instructional manipulation checks were removed.

To determine the most representative medical conditions for each *ConditionType*, the HBM construct ratings were compared within the same category using the Kruskal-Wallis test [128]. This test was selected because ratings like Likert scores are generally treated nonparametericallyand each of the factors had more than two levels. When statistical significance was found, post-hoc Mann-Whitney U tests [151] with the Bonferroni-Holm correction [103] were used for pairwise comparisons. After the representative medical conditions were selected, a similar analysis was performed to compare HBM construct ratings across *ConditionType* to ensure that there was sufficient separation between them.

7.4.4 Results: Within ConditionType

Common Conditions

Statistically significant differences were found across the three *Common* conditions for both *ScenarioPlausibility* ($H(2) = 9.091, p < .05$) and *AppPlausibility* ($H(2) = 8.247, p < .05$). The sinus infection scenario was significantly less believable than the other two conditions ($p < .05$ vs. both strep throat and pink eye). The pink eye app was significantly more believable than the sinus infection app ($p < .05$). 100.0% of the respondents stated that the pink eye scenario was at least slightly believable, and 70.0% of the respondents stated that the pink eye app was at least slightly believable. Given these results, I selected pink eye as my representative *Common* condition.

Serious Conditions

Statistically significant differences were found across the three *Serious* conditions for both *ScenarioPlausibility* ($H(2) = 15.264, p < .001$) and *AppPlausibility* ($H(2) = 8.832, p < .05$). The pancreatic cancer scenario was significantly less believable than the other two conditions ($p < .01$ vs. both skin cancer and anemia). The skin cancer

app was significantly more believable than the anemia app ($p < .01$). 93.3% of the respondents stated that the skin cancer scenario was at least slightly believable, and 80.0% of the respondents stated that the skin cancer app was at least slightly believable. Given these results, I selected skin cancer as my representative *Serious* condition.

Stigmatizing Conditions

A statistically significant difference was only found across the three *Stigmatizing* conditions for *AppPlausibility* ($H(2) = 6.420, p < .05$). The psoriasis app was slightly more believable than the IBS app ($p = .06$). However, there was a statistically significant difference between the three *Stigmatizing* conditions regarding the impact they would have on a person's social life and professional standing ($H(2) = 14.892, p < .01$). In particular, psoriasis was deemed significantly less impactful than the other two conditions ($p < .005$ vs. both halitosis and IBS). Since halitosis was rated at least as high as the other *Stigmatizing* conditions in terms of *ScenarioPlausibility*, *AppPlausibility*, and *PerceivedSeriousness*, I selected halitosis as my representative *Stigmatizing* condition. 90.0% of the respondents stated that the halitosis scenario was at least slightly believable, and 46.7% of the respondents stated that the halitosis app was at least slightly believable. Although the latter number is low compared to the other condition categories, 26.7% of the respondents stated that they found the halitosis app to be neither believable nor unbelievable.

7.4.5 Results: Across ConditionType

Figure 7.4 shows the distribution of *ScenarioPlausibility* and *AppPlausibility* ratings for the three selected medical conditions. No statistically significant difference was found across the conditions for *AppPlausibility* ($H(2) = 3.067, p = .21$), but there was a statistically significant difference for *ScenarioPlausibility* ($H(2) = 9.068, p < .05$). The scenario about pink eye was significantly more believable than the scenario about halitosis ($p < .01$).

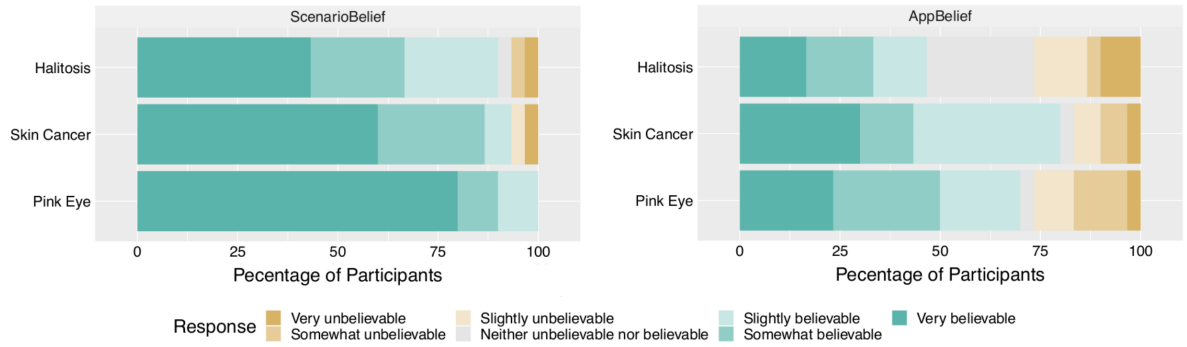

Figure 7.4: The distribution of ratings for **(left)** *ScenarioPlausibility* and **(right)** *AppPlausibility*.

Nevertheless, I were satisfied with the selected conditions since they all had high median *ScenarioPlausibility* ratings.

Statistically significant differences were found across the three conditions for all of the HBM constructs, including *PerceivedSeriousness* regarding long-term health ($H(2) = 47.352, p < .001$), *PerceivedSeriousness* regarding finances ($H(2) = 49.162, p < .001$), *PerceivedSeriousness* regarding social standing ($H(2) = 16.128, p < .001$), and *PerceivedSusceptibility* ($H(2) = 34.218, p < .001$). There were no statistically significant ordering effects for these tests.

My definition of a *Serious* medical condition suggests that skin cancer should have a higher impact on a person's long-term health and finances than the other two medical conditions. The definition also suggests that people should believe that they are less prone to having skin cancer than the other medical conditions. My results supported both of these hypotheses. Skin cancer was rated as having a significantly higher *PerceivedSeriousness* regarding long-term health ($p < .001$), higher *PerceivedSeriousness* regarding finances ($p < .001$), and lower *PerceivedSusceptibility* ($p < .001$) compared to pink eye and halitosis.

My definition of a *Stigmatizing* medical condition suggests that halitosis should have

a higher impact on a person's social life and professional standing than the other two medical conditions. Halitosis was rated as having a significantly higher *PerceivedSeriousness* on social standing than pink eye ($p < .05$); however, there was not a significant difference between halitosis and skin cancer ($p = .18$). Nevertheless, the other characteristics that were unique to skin cancer as a *Serious* condition provided enough separation between them.

The combination of these results indicates that pink eye was viewed as having low *PerceivedSeriousness* and high *PerceivedSusceptibility*. Therefore, pink eye was deemed suitable as a *Common* condition.

7.5 Evaluation of Research Questions

With my scenarios and sensor-based health-screening apps selected, I were prepared to deploy my survey instrument and investigate my research questions. In this section, I describe the survey deployment, my use of structural equation modeling (SEM) for analysis, and my findings.

7.5.1 Participants

This study was advertised through the same outlets as the previous study for scenario and app selection. Respondents who reported no smartphone experience were excluded from participation. Respondents who completed the survey were eligible for a raffle in which 1-in-20 people would win a $20 Amazon gift card and 1-in-100 people would win a $100 Amazon gift card. In total, 263 respondents completed the survey from start to finish. A subset of their demographic information is provided in Table 7.7.

7.5.2 Apparatus

Figure 7.5 illustrates the structure of the survey that was shown to respondents. To address my exploratory research questions, I varied the app description to include

Table 7.7: Participant demographics for main evaluation

Survey Demographics (N=263)	
Source	Facebook (16), ITHS (240), Reddit (3), Other (4)
Gender	Male (45), Female (204), Transgender Male (5), Gender Variant / Non-conforming (7), Self-Identify (1), Undisclosed (1)
Age	18-24 (145), 25-34 (84), 35-44 (17), 45-54 (8), 55-64 (3), 65+ (3), Undisclosed (3)
Country of residence	United States (257), Belgium (1), Hong Kong (1), Indonesia (1), Netherlands (1), Rwanda (1), Spain (1)
Smartphone operating system	iOS (170), Android (93)
Self-reported smartphone experience	Expert/Advanced (146), Intermediate (115), Novice/Beginner (2)

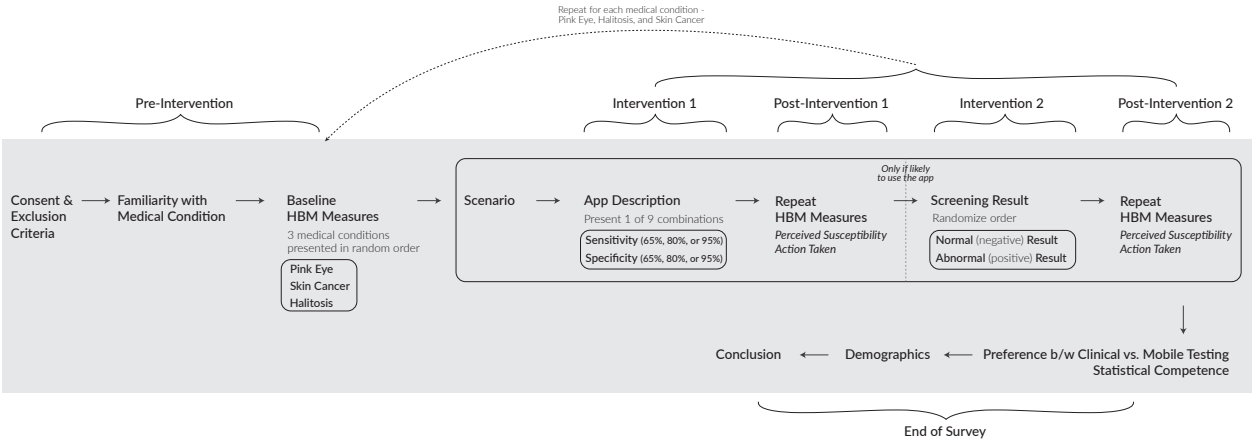

Figure 7.5: The structure of the survey given to respondents to investigate my confirmatory and exploratory research questions. The structure builds on my survey instrument (Figure 7.2), including many different target medical conditions and the sensing accuracy in the technology descriptions.

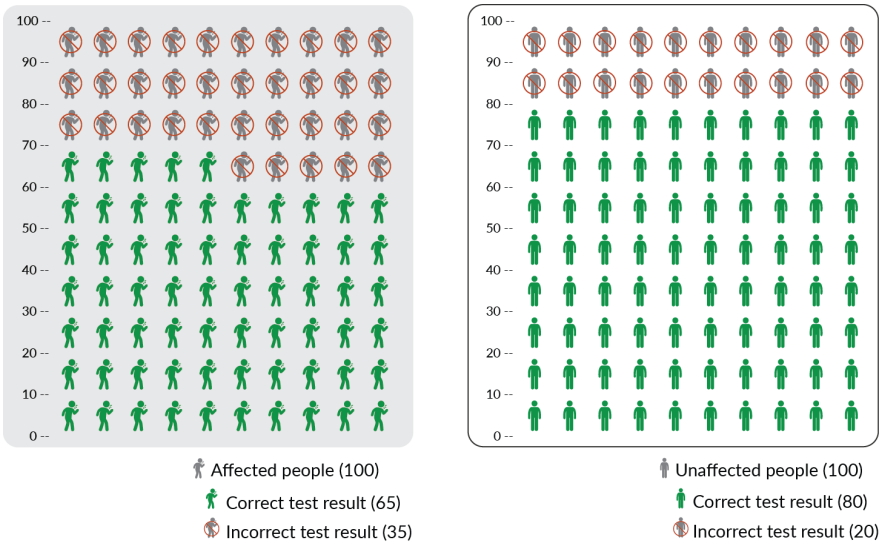

Figure 7.6: An example of an icon array provided in the survey instrument to illustrate the sensitivity and specificity of an app. This array illustrates 65% sensitivity and 80% specificity.

sensitivity and specificity information for the sensor-based health-screening apps. Sensitivity and specificity rates were presented with counts, rather than probabilities or fractions, because prior work has found that the general public is more adept at reasoning about counts [236]. The rates were also presented in graphical form using icon arrays (Figure 7.6), again using counts for readability. An example of the additional text is provided below:

> *Out of every 100 people who **have** a sinus infection, SinusCheck correctly told 65 people that they had a sinus infection. Out of every 100 people who **do not have** a sinus infection, SinusCheck correctly told 80 people that they did not have a sinus infection.*

Out of the 361 respondents who opened the survey, 265 completed it (73.4%), 94

partially completed it (26.0%), and 2 were disqualified because they did not own a smartphone (0.6%). Ignoring the five cases when participants completed the survey a day after starting it, the median survey completion time was 16.1 minutes. The average completion time was 23.7 ± 29.0 minutes.

7.5.3 Design and Analysis

The survey instrument was used in a $3 \times 3 \times 3$ mixed factorial design study. Each respondent read all three scenarios that were selected from the previous study—pink eye (*Common*), skin cancer (*Serious*), and halitosis (*Stigmatizing*)—making *ConditionType* a within-subjects factor. The presentation order of the scenarios was counterbalanced across all subjects. Three equally-spaced levels of sensitivity and specificity were investigated—65%, 80%, and 95%—producing 9 possible combinations that described the overall accuracy of the apps. Each app for each respondent was assigned one of those 9 combinations at random, making *Sensitivity* and *Specificity* between-subjects factors. Responses from respondents who failed the instructional manipulation checks were removed.

The survey responses were analyzed using structural equation modeling (SEM) [110]. Prior work has used SEM and its variants to find evidence that supports the HBM framework and the effectiveness of health-related interventions [28, 169]. SEM revolves around a graphical model known as a path diagram. A path diagram describes hypothesized causal relationships between variables. The nodes of the path diagram can either represent single observable constructs or quantities that are not directly observable, the latter of which are known as latent variables. The nodes are connected by directed edges that describe the causal interactions. A series of regressions are performed on a path diagram to generate a model where each edge is assigned a path coefficient and p-value. The coefficient is not a correlation coefficient, but rather an indication of how one variable influences another. For example, if the coefficient

between A and B is 0.10, an increase in A by one standard deviation from its mean would be expected to cause an increase in B by 0.10 of its own standard deviation from its mean while holding all other factors constant. Chin [37] proposes that meaningful path coefficients have an absolute magnitude greater than 0.20, though other guidelines exist as well.

There are a number of fit statistics that researchers use to assess the overall quality of a model, each with their own trade-offs and no generally agreed upon standard [124, 105, 172]. I use comparative fit index (CFI) as my main indicator of fit goodness. CFI compares the fit of a target model to the fit of an independent model in which the variables are assumed to be uncorrelated. CFI ranges between 0 and 1.00, where 1.00 is the best result. As with path coefficients, there is little agreement on the recommended cut-off that indicates an acceptable fit. As a point of reference, Hooper et al. [105] believe that a CFI > 0.90 dictates a strong fit.

Figure 7.7 shows the complete path diagrams for each of the survey instrument's outcome variables. Since three questions were used to probe *PerceivedSeriousness*, those responses are consolidated into a single latent variable. *AppInterest* is independent of the action-specific HBM constructs—*PerceivedBenefits* and *PerceivedBarriers*—so they are excluded its path diagram. The variables related to the app—*Sensitivity*, *Specificity*, and smartphone-specific *ModifyingVariables*—are only connected to variables that can change after the person is shown a test result: *PerceivedSusceptibility*, *ActionTaken*, *ActionChangePositive*, and *ActionChangeNegative*. *ConditionType* and *ActionType* are used as grouping variables and are thus not included in the path diagrams.

The analyses in this work were conducted using the R library `lavaan` [211]. `Lavaan` is equipped to handle binary and ordinal variables, but not multi-level categorical variables like race or gender; such variables were reformulated as binary dummy variables. I used a robust variant of NLMINB [84] as my model estimator since my data included non-continuous variables. For some of the analyses, the same model was fit across groups (e.g., different actions or medical conditions) to answer the same question. This was done

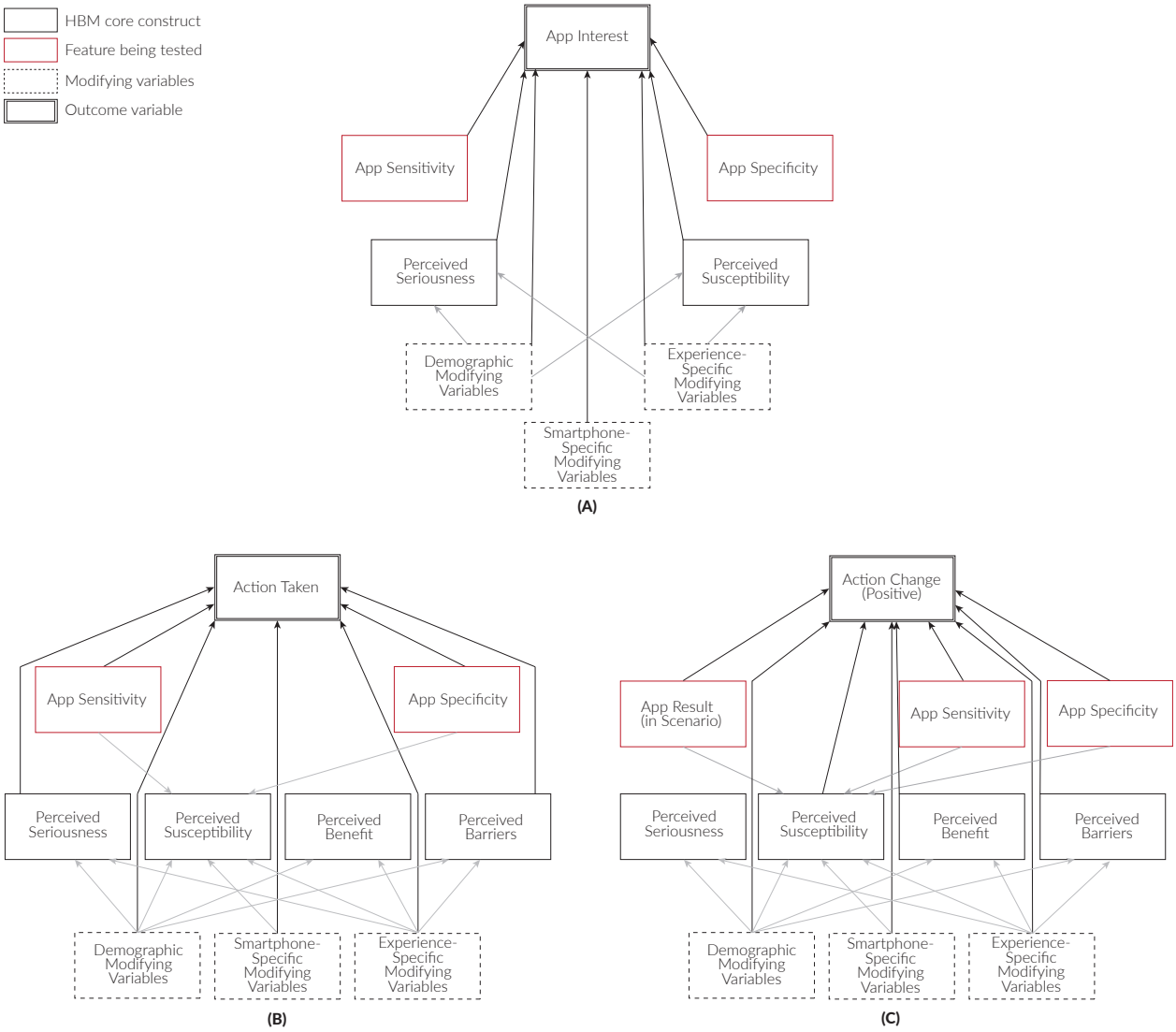

Figure 7.7: The complete path diagrams for the different analyses conducted in the study: (a) *AppInterest*, (b) *ActionTaken*, and (c) *ActionChangePositive/ActionChangeNegative*

.

using an extension of SEM called multi-group SEM, where each group is assigned its own path coefficients, but a single CFI is calculated across the entire model. By splitting the data into groups, there were cases when certain factor levels disappeared; in those cases,

Table 7.8: Path coefficients for the confirmatory analysis of *AppInterest* without *ModifyingVariables* (CFI = 0.961).

	AppInterest
∼Seriousness	0.132***
∼Susceptibility	0.057*

*$p < .05$, **$p < .01$, ***$p < .001$

the missing level was clustered with a nearby level across all groups.

Multi-group SEM and *ModifyingVariables* produced too many results to include in this paper. I surface the most important results for the sake of brevity.

7.5.4 RQ.1: Confirmatory Analysis Results for AppInterest

The confirmatory analysis for **RQ.1** was conducted on the *AppInterest* path diagram (Figure 7.7a) across all levels of *ConditionType* without paths connected to the manipulated variables—*Sensitivity* and *Specificity*. Table 7.8 shows the causal path coefficients for the model fit without *ModifyingVariables*, which was strong compared to baseline independent model (CFI = 0.961).

According to the HBM, I expected to see statistically significant positive coefficients from *PerceivedSeriousness* and *PerceivedSusceptibility* to *AppInterest*. Significant positive coefficients were found in both cases, although the coefficient from *PerceivedSeriousness* was more than double that from *PerceivedSusceptibility*. The model with *ModifyingVariables* was also strong compared to the baseline independent model (CFI = 0.959). However, a statistically significant coefficient was only found from *PerceivedSeriousness* to *AppInterest* in that case (*PerceivedSeriousness* $\rightarrow$ *AppInterest* $= 0.114, p < .01$). The lack of a significant effect from *PerceivedSusceptibility* could be because the introduction of *ModifyingVariables* incorporated many more degrees of freedom. Nevertheless, the fact that the effect from

Table 7.9: Path coefficients for the confirmatory analysis of *ActionTaken* without *ModifyingVariables* (CFI = 0.965).

	Action Taken (Schedule Appt)	Action Taken (Contact Physician)	Action Taken (Stay at Home)	Action Taken (Purchase Meds)
∼Seriousness	0.161***	0.203***	-0.05	0.156**
∼Susceptibility	0.361***	0.345***	0.207***	0.235***
∼Benefits	0.224***	0.14***	0.248***	0.318***
∼Barriers	-0.076***	-0.046*	-0.129***	-0.104

*$p < .05$, **$p < .01$, ***$p < .001$

PerceivedSeriousness was stronger and more significant than the effect from *PerceivedSusceptibility* both with and without *ModifyingVariables* shows that people cared more about the seriousness of a medical condition than their likelihood of getting getting a medical condition when it came to *AppInterest*.

The model with *ModifyingVariables* had other relationships with statistically significant coefficients. Respondents who were more familiar with the medical conditions were more likely to believe that they posed a serious threat to their health, although the magnitude of the effect was small (*Familiarity* $\rightarrow$ *PerceivedSeriousness* $= 0.087, p < .01$). Although increased *PerceivedSeriousness* is not always desirable and could indicate hypochondria in extreme cases, this result could show that respondents with more intimate knowledge about the medical conditions took the repercussions of inaction more seriously.

7.5.5 RQ.2: Confirmatory Analysis Results for ActionTaken

Similar to **RQ.1**, the confirmatory analysis for **RQ.2** was conducted on the *ActionTaken* path diagram (Figure 7.7b) across all levels of *ConditionType* without paths connected to the manipulated variables. Multi-group SEM was applied with *ActionType* as the group variable to analyze each action separately. Table 7.9 shows the causal path coefficients

for the model fit without *ModifyingVariables*, which was strong compared to the baseline independent model (CFI = 0.965)

According to the HBM, I expected to see statistically significant positive coefficients from *PerceivedSeriousness*, *PerceivedSusceptibility*, and *PerceivedBenefits* to *ActionTaken*. I also expected to see a statistically significant negative coefficient from *PerceivedBarriers* to *ActionTaken*. These hypotheses were all supported by the model fit. The same could be said about the model with *ModifyingVariables* (CFI = 0.907). The lower CFI with the inclusion of *ModifyingVariables* could be because multi-group SEM splits the data into smaller subsets, thereby reducing statistical power while maintaining degrees of freedom.

As before, the model with *ModifyingVariables* presented intriguing statistically significant relationships. For example, respondents who were more familiar with the conditions were more likely to consider health-promoting actions to be beneficial (*Familiarity* $\rightarrow$ *PerceivedBenefits* $= 0.103, p < .001$) and less likely to consider them difficult (*Familiarity* $\rightarrow$ *PerceivedBarriers* $= -0.089, p < .001$); this was true for all actions except "purchasing over-the-counter medication". These results could show that respondents who were more familiar with the medical conditions were more aware of the repercussions of inaction, thus viewing those actions as easy to do and worth their time.

7.5.6 Summary of Confirmatory Analyses

For both *AppInterest* and *ActionTaken*, my results generally followed my expectations according to the HBM. All of the path coefficients from the core HBM constructs to the outcome variables had the sign that I expected them to have. With the exception of the path coefficient between *PerceivedSusceptibility* and *AppInterest*, those coefficients were also statistically significant. The magnitude of the path coefficients for the *AppInterest* model did not exceed Chin's recommended threshold of 0.2, but those for the *ActionTaken* model did. Therefore, I had confidence to move forward and explore the rest

Table 7.10: Path coefficients for the exploratory analysis of *AppInterest* (CFI = 0.997).

	AppInterest (Common)	AppInterest (Serious)	AppInterest (Stigmatizing)
∼Seriousness	-0.120	0.101	0.129
∼Susceptibility	0.206**	0.120*	0.104
∼Sensitivity	0.416***	0.357***	0.268**
∼Specificity	0.461***	0.300**	0.292***

*$p < .05$, **$p < .01$, ***$p < .001$

of my research questions.

7.5.7 RQ.3: Exploratory Analysis Results for AppInterest

The exploratory analysis for **RQ.3** was conducted on the *AppInterest* path diagram (Figure 7.7a) with the manipulated variables—*Sensitivity* and *Specificity*. Since the *ModifyingVariables* added many degrees of freedom with few significant relationships, they were excluded from these analyses. Multi-group SEM was applied with *ConditionType* as the group variable to analyze the perception of each app separately. Table 7.10 shows the causal path coefficients for the model fit. The model produced a very strong fit against the baseline independent model (CFI = 0.997).

The path coefficients from *Sensitivity* and *Specificity* to *AppInterest* were sizable and positive across all levels of *ConditionType*, confirming that increased accuracy made the sensor-based health-screening apps more attractive. In fact, the effect was so strong that those coefficients were much more statistically significant than those from the core HBM constructs. This suggests that when a person is presented with the opportunity to use a sensor-based health-screening app, they may be willing to use it regardless of their health condition as long as they know that the app is accurate.

When examining the different levels of *ConditionType* individually, accuracy was most

valued for the *Common* condition, then the *Serious* condition, and then the *Stigmatizing* condition. *Specificity* was preferred over *Sensitivity* for the *Common* and *Stigmatizing* condition, while *Sensitivity* was preferred over *Specificity* for the *Serious* condition. In some sense, this finding demonstrates that respondents had an inherent knowledge about the notion of prevalence and how it relates to the diagnostic decision making process. The *Common* and *Stigmatizing* conditions are fairly prevalent, so prioritizing *Specificity* indicates that respondents wanted to use an app's test result to "rule out" having the condition. The *Serious* condition is less prevalent, so prioritizing *Sensitivity* indicates that respondents wanted to "rule in" having the condition.

7.5.8 RQ.4: Exploratory Analysis Results for ActionTaken

The exploratory analysis for **RQ.4** was conducted on the *ActionChangePositive* and *ActionChangeNegative* path diagrams (Figure 7.7c) with the manipulated variables and without the *ModifyingVariables*. Multi-group SEM was applied with the combination of *ConditionType* and *ActionType* as the group variables, producing 10 model fits (4 actions for *Common* + 2 actions for *Serious* + 4 actions for *Stigmatizing*). Separate analyses were conducted for positive and negative test results.

Positive Test Results

Table 7.11 shows the path coefficients for *ActionChangePositive*. Across all combinations of *ConditionType* and *ActionType*, there were 1,074 cases when respondents said that they would not take an action before using the app. Of those 1,074 cases, 417 (38.8%) changed their mind after receiving a positive test result. The number of action changes was roughly evenly distributed across the three levels of *ConditionType*. The model produced a strong fit against the baseline independent model (CFI = 0.959).

Across all scenarios, there was a large, positive coefficient between *AppResult* and *ActionChangePositive*. This result was expected since respondents had to see a test result

Table 7.11: Path coefficients for the exploratory analysis of *ActionChangePositive*, specifically for "scheduling an appointment" (CFI = 0.959).

	Common		Serious		Stigmatizing	
	Action Change	Suscept-ibility	Action Change	Suscept-ibility	Action Change	Suscept-ibility
∼Seriousness	0.129		-0.297		0.192	
∼Susceptibility	0.426**		0.508***		0.474***	
∼Benefits	0.226**		0.273		0.055	
∼Barriers	-0.137		-0.061		0.038	
∼AppResult	6.962***	0.398*	6.524***	1.516***	5.902***	0.471**
∼Sensitivity	-0.263	0.283**	-0.009	0.01	-0.004	0.083
∼Specificity	-0.185	0.093	-0.204	-0.107	0.033	-0.032

*$p < .05$, **$p < .01$, ***$p < .001$

in order to change their opinion on *ActionTaken*. There was also a strong positive coefficient between *AppResult* and *PerceivedSusceptibility* across all scenarios, which verified my intuition that a positive test result should increase a person's perceived likelihood of having a medical condition. However, the magnitude and significance of that coefficient varied across the different medical conditions. The coefficient from *AppResult* to *PerceivedSusceptibility* for the *Serious* medical condition was three times as large and more significant than the corresponding coefficients for the other medical conditions. Again, this result hints at the fact that respondents were willing to use the positive test result from an app related to "rule in" having a *Serious* condition.

Specificity corresponds to a test's true negative rate and is thus more directly linked to negative test results, but *Specificity* does impact how a positive test result should be interpreted according to Bayesian statistics. Nevertheless, *Specificity* did not have a statistically significant effect on either *ActionChangePositive* or *PerceivedSusceptibility*.

Table 7.12: Path coefficients for the exploratory analysis of *ActionChangeNegative*, specifically for "scheduling an appointment" (CFI = 0.949).

	Common		Serious		Stigmatizing	
	Action Change	Suscept-ibility	Action Change	Suscept-ibility	Action Change	Suscept-ibility
∼Seriousness	-0.451*		-0.182		0.169	
∼Susceptibility	-0.311**		-0.358***		-0.212	
∼Benefits	0.001		0.125		-0.196	
∼Barriers	0.105		0.083		-0.176	
∼AppResult	6.230***	-1.952***	6.144***	-0.970***	6.833***	-2.191**
∼Sensitivity	0.022	-0.034	-0.022	-0.026	0.003	0.056
∼Specificity	-0.103	-0.001	0.063	-0.185*	-0.192	-0.119

*$p < .05$, **$p < .01$, ***$p < .001$

There were, however, cases when *Sensitivity* had a statistically significant effect on *ActionChangePositive* and *PerceivedSusceptibility* for the *Common* and *Stigmatizing* condition scenarios. Since statistically significant effects were not found from *Sensitivity* to *ActionChangePositive* or *PerceivedSusceptibility*, one could surmise that respondents were willing to accept a positive test result for a *Serious* condition regardless of the app's reported accuracy.

Negative Test Results

Table 7.12 shows the path coefficients for *ActionChangeNegative*. Across all combinations of *ConditionType* and *ActionType*, there were 982 cases when respondents said that they would take an action before using the app. Of those 982 cases, 451 (45.9%) changed their mind after receiving a negative test result. Almost half of those action changes occurred in the *Common* condition scenario. As with the model for positive test results, the model

for negative test results produced a similarly strong fit against the baseline independent model (CFI = 0.949).

As before, there was a large, positive coefficient between *AppResult* and *ActionChangeNegative* across all scenarios. On the other hand, there were strong negative coefficients between *AppResult* and *PerceivedSusceptibility* across all scenarios. Negative coefficients were expected since a negative test result should decrease a person's perceived likelihood of having a medical condition. This result was statistically significant across all scenarios, but the magnitude varied. The coefficients from *AppResult* to *PerceivedSusceptibility* for the *Common* and *Stigmatizing* condition were nearly double the corresponding coefficient for the *Serious* condition, indicating that respondents used the negative test result from those apps to "rule out" having those conditions.

Sensitivity did not have a statistically significant effect on either *ActionChangeNegative* or *PerceivedSusceptibility*, which aligns with the reasoning before for *Specificity* and positive test results. In the *Serious* condition scenario, statistically significant negative coefficients were found from *Specificity* to *ActionChangeNegative* and *PerceivedSusceptibility* for both of the possible actions. The same could not be said for the other two medical conditions, indicating that respondents were willing to accept a negative test result in those cases regardless of that app's reported accuracy.

7.5.9 Summary of Exploratory Analyses

My exploratory analysis revealed that respondents greatly valued *Sensitivity* and *Specificity* when considering whether they would use a sensor-based health-screening app. However, my data suggests that the way that people value those quantities depends on the target medical condition. Respondents valued *Sensitivity* over *Specificity* for the *Serious* condition, yet *Specificity* over *Sensitivity* for the *Common* and *Stigmatizing* conditions.

When presented with a positive test result, respondents seemed willing to change their mind in the *Serious* condition scenario regardless of the app's reported *Sensitivity* or *Specificity*. Respondents did, however, cared more about *Sensitivity* when determining how likely they were to have a *Common* condition. When presented with a negative test result, respondents seemed willing to change their mind in the *Common* and *Stigmatizing* condition scenarios regardless of the app's reported *Sensitivity* or *Specificity*, yet respondents cared more about *Specificity* when determining how likely they were to have a *Serious* condition.

7.6 Discussion

My goal was to develop a survey instrument that researchers can use as they consider translating their ubiquitous health-screening technology from research into practice to maximize its potential effectiveness at supporting health-related decision-making. To that end, I demonstrated that the coefficients from the core HBM constructs to *ActionTaken* and the overall model fit were statistically significant. I discuss the findings and design implications from the specific investigation I ran, and then I delve into survey design considerations for researchers who may use my instrument.

7.6.1 Design Implications for Sensor-Based Health-Screening Apps

Through my exploration into sensor-based health-screening apps, I uncovered that respondents were willing to use positive test results to "rule in" a serious medical condition regardless of whether the test had 65% sensitivity or 95% sensitivity. Similarly, I found that respondents were willing to use negative test results to "rule out" common and stigmatizing medical conditions regardless of test specificity. These results show that sensor-based health-screening apps can have a large and even undue influence on a person's course of action, a finding that should be accounted for in the design of such technologies.

Researchers often design their data collection mechanisms and classifiers to optimize overall accuracy, but my results suggest that researchers should consider the trade-off between sensitivity and specificity for their target medical condition. For example, part of my survey in Section 7.5 asked respondents to place themselves in a scenario where they might believe they have skin cancer. It is currently common practice for a physician to encourage a patient to monitor a mole with an ambiguous appearance to see how it develops over time. Likewise, a smartphone app that aims to minimize the unnecessary costs associated with false positives might employ such an approach, asking a person to test themselves over time to generate greater decision confidence. Because my data suggests that people do not always properly consider accuracy metrics when deciding their next course of action, this approach might be preferable to reporting a result as "abnormal with 65% confidence". Other health-screening apps may take different approaches according to the the implications of the trade-off between sensitivity and sensitivity.

7.6.2 Increasing Model Complexity

SEM is a powerful tool for evaluating causal models. The path diagrams for my analyses consisted of regressions and one latent variable for *PerceivedSeriousness*, but SEM can accommodate many other constraints. My model could have included more latent variables with the demographic information I collected, such as "statistical understanding" or "comfort with technology". SEM also allows researchers to specify expected intercepts and covariance between variables in order to further constrain the models. Because my core contribution in this paper is a demonstration of the survey instrument and not an exhaustive evaluation of the survey, I specified neither in my analyses. Future work can be done to examine the impact and utility of these specifications.

7.6.3 Accuracy Metrics and Prevalence

Not all positive test results indicate that a person has a medical condition, nor do all negative test results indicate that a person does not have a medical condition. My evaluation was concerned with situations when a sensor-based health-screening app might incite a change in action. However, a change in action is not always a good result. The nuance that is missing from my survey instrument lies in the notion of a prior probability.

Prior probabilities are influenced by a number of factors, including prevalence (i.e., the rate at which a medical condition manifests in a population), medical history, and habits. Sensitivity (SNS) and specificity (SPC) rates are the standard metrics used to describe the accuracy of a medical test because they exclude the notion of prevalence. Sensitivity and specificity can be used to compute the diagnostic power of a positive test result, called its positive likelihood ratio (LR_+):

$$LR_+ = \frac{SNS}{1 - SPC} \tag{7.1}$$

According to Bayesian reasoning, the prior probability (P_0) is required for using sensitivity and specificity to actually make a diagnosis. When a person receives a positive test result, that prior probability is updated to a posterior probability (P) as follows:

$$P = \frac{SNS * P_0}{(1 - SPC)(1 - P_0) + SNS * P_0} = \frac{LR_+ * P_0}{1 - P_0 + LR_+ * P_0} \tag{7.2}$$

I excluded the notion of prior probabilities in my survey instrument because it is difficult to accurately quantify P_0 without detailed information about the respondent's environment and medical history. Prevalence rates also varied across the different medical conditions I investigated, which would have complicated the interpretation of my results. Future deployments of my survey instrument that focus on a single medical condition could select a fixed prevalence rate and go as far as calculating a posterior probability for respondents to see how their reactions change.

In fact, an alternative survey format could emerge from the use of Bayesian reasoning. Rather than asking respondents to commit to action or inaction, my survey could have asked respondents to express the likelihood that they would take action as a probability percentage. Using the percentage before the introduction of the app as P_0 and the percentage after as P, the diagnostic power a person attributes to the test can be calculated by solving Equation 7.2 for LR_+. The test's actual positive likelihood ratio can be calculated using the sensitivity and specificity rates shown to the respondent. Comparing the respondent's perceived positive likelihood ratio versus the test's actual positive likelihood ratio can indicate whether the respondent underestimated or overestimated an app's diagnostic power. In a way, this approach would have mimicked prior work in judgment analysis [95, 44] where an optimal decision can be made. I chose to avoid probabilistic responses because prior research has found that people tend to apply an inherent weighting function to probability values [113], but finding an effective format to elicit such information is a future research opportunity.

7.6.4 Exploration of Other Potential Attributes

I explored the influence of the target medical condition and app accuracy on people's reactions to sensor-based health-screening apps, but my survey instrument can be easily adapted to explore a variety of other attributes.

Interface Design

Interface design is a major attribute that warrants exploration. I chose to exclude screenshots of potential interfaces so that respondents would focus on the overall scenario rather than specific design decisions like fonts, instructions, or screen layouts. This decision coincides with Truong et al.'s [241] finding that the exclusion or abstraction of elements like text can control where a participant's attention is drawn when prototyping with storyboards. Nevertheless, a technology's interface is critical to its

credibility. Something as small as a typo may make an app seem unprofessional, leading a potential end-user to distrust the app. Researchers can add screenshots to the technology descriptions and loop through different interface designs in the survey flow as I did for target medical conditions in Figure 7.5. Unique features within those screenshots would be encoded as binary variables within the SEM analysis.

Technology Pricing

One interesting attribute that arose during my pilot testing was app pricing. At first, I stated that apps cost $0.99 because I worried that a free app would appear illegitimate and unregulated while an expensive app would diminish interest. When potential participants were shown descriptions of a $0.99 app, they felt that a less expensive app was less legitimate than a free one because the $0.99 price was viewed as "cheap". With my survey instrument, researchers can simply mention the price when they are describing their technology. Price would be treated as an ordinal variable during the SEM analysis.

Endorsement

Another factor that influences the legitimacy of a technology is endorsements. Apps stores, smartphone manufacturers, special interest groups, and physicians can all endorse technologies, serving as a "seal of approval" that may imbue end-users with confidence in a technology. A limitation of my survey instrument is that it is difficult to convey an endorsement to respondents without explicitly drawing the respondent's attention to it. Endorsements can appear in many places—commercials, supplemental materials, or websites—that may not be as conspicuous as mentioning would be done in the survey. Determining a more natural way of introducing endorsements could be a potential avenue for future work.

Chapter 8
IMPLICATIONS AND CONCLUSIONS

8.1 Summary

In this dissertation, I provided evidence in support the following thesis statement:

> *Technological and scalability barriers to some medical assessments can be addressed through smartphone-based sensing tools; moreover, the acceptability of these tools can be addressed through surveys that reveal how these tools and their results are likely to be regarded by potential users.*

First, I provided three examples of how smartphone sensors can be used to reduce **technological and scalability barriers** to medical testing. I posited that this could be achieved by using smartphone sensors to make medical observations accurate, precise, repeatable, and pervasive. I supported this supposition with three examples focused on visual observations of the eye, showing how cameras can be used to outperform human sight. Although some mobile health technologies for basic biometrics are already seeing usage by average consumers, technologies like the ones I presented are far from being accessible to all. This dissertation describes challenges that I believe either hinder mobile health from reaching complete ubiquity (e.g., smartphone heterogeneity) or serve as a warning to make sure that mobile health actually improves medical outcomes (e.g., proper result interpretation). I describe recent steps that have been taken to address those challenges and offer up potential areas of future exploration. To address one of those challenges, I presented a survey instrument that provides a low-cost way for researchers to examine **acceptability barriers**. The survey instrument provides a way for researchers and designers to modify characteristics of their technology and gauge how

their target audience would react to those changes when deciding if they would take health-promoting action.

8.2 Implications

I firmly believe that mobile health will continue to grow over the next decade, allowing people to monitor aspects of their health in the comfort of their homes. Regardless of whether mobile health continues to be delivered through current mobile devices like smartphones and smartwatches or emerging devices like augmented reality headsets, there must be more research effort in this space to ensure that mobile health does more good than harm. I proposed some potential research directions in Chapter 6. Here, however, I discuss the broader implications of my work outside of the HCI and UbiComp communities.

8.2.1 Medicine and Physiology

Work in the mobile health space can accelerate discoveries in medicine and physiology by enabling clinical researchers to rapidly scale up deployment and reach populations that were previously unreachable. Although clinicians called PupilScreen with the box "inconvenient", they were far more willing to conduct clinical studies on TBI with a smartphone than with a pupillometer. This fact was shocking to me considering that clinicians already have access to pupillometers and pupillometers are have been supported by more evidence than PupilScreen. This shows that the convenience of a smartphone cannot be taken for granted. The clinicians we spoke to were very excited about using their own smartphones that they carry in their pockets to conduct a test rather than searching for a separated device on their floor. Beyond convenience, clinicians preferred smartphones for their familiar user interface (i.e., a touchscreen on display) rather than the extra buttons and switches that are included on a pupillometer.

Nevertheless, I believe that there must be a clear understanding between technical

and clinical researchers about the algorithmic development process, particularly when the algorithms involve data-driven models. Machine learning requires lots of data before results can be trusted, yet some people are only willing to participate in research if they can receive results in real-time. Striking a balance in this regard is a critical issue that must be addressed in a case-by-case manner.

8.2.2 User Experience and Design

I believe that user experience researchers and designers are going to be critical to the adoption of mobile health apps. As I describe in Section 6.3, smartphone-based health-screening apps often require proper compliance by users to ensure that the results they receive are relevant to their health and not their environment or abilities. Clever user experiences can be engineered to encourage proper data collection, whether those experiences involve guiding users through proper sensor placement or checking users' environments. I also foresee an opportunity for user experience researchers to create apps for the sole purpose of collecting high-quality sensor datasets. For instance, a smartphone app that guides users through taking clear and consistently positioned pictures of their face could power a number of applications that examine the progression of facial features.

As I describe in Section 6.4, well-designed user interfaces are also critical for ensuring that people properly interpret their test results. Interfaces should highlight the information most important to users without overwhelming them with unnecessary details; however, interfaces should also include enough information to instill the system with legitimacy. The survey instrument I presented in Chapter 7 provides a way for researchers to quickly explore that balance without requiring a functioning prototype. Eventually, I hope enough researchers examine this problem so that we someday reach more generalizable guidelines as the data visualization community has done for presenting visual information to non-experts.

8.2.3 Computer Architecture

When prototyping smartphone-based health-screening apps, I usually start by running code offline with all of the computational power available to me to explore complex machine learning models and to push accuracy as high as possible. The most immediate path to a full-blown deployment would involve setting up a centralized server that houses those models and accepts HTTP requests. However, such a system requires wireless connectivity, which is not always guaranteed in the developing world. Wirelessly transmitting data also introduces vectors for invading users' privacy, which is particularly important given the sensitive nature of health-related data. Recent efforts have been made to support deep learning on edge devices; it is my hope that these efforts continue so that mobile health can remain at the edge.

8.3 Reflection

To conclude my dissertation, I would like to take the opportunity to make two final points based on my experience in developing smartphone-based health-screening apps.

First, I would like to bring up a philosophical question about whether researchers should actually be striving towards at-home health screening through a smartphone altogether. I have had friendly arguments with other researchers who believe that we should instead be focusing on custom hardware. They go on to say that by restricting ourselves to the sensors that are available on smartphones, we are ignoring decades of research that has been dedicated to developing sensors that make our research problems much easier. A prime example of this tension is PupilScreen. I focused on segmenting the pupil using smartphone cameras, which detect visible wavelength; however, accurate gaze tracking and pupil measurement devices have already been commercialized using infrared cameras since those cameras emphasize the pupil regardless of iris color.

One issue I find with using custom hardware is that it limits the potential scalability

of a solution. Even if the novel hardware proves to be extremely useful, its uptake will be impeded by the rate at which people purchase the new device, assuming they can afford it in the first place. Smartphones are now beginning to include infrared cameras for augmented reality applications, only people who can afford those phones would benefit from a tool like PupilScreen. Mobile health is meant to reduce barriers to healthcare access, not make them worse. I believe that targeting existing smartphones early in the design process allows for more scalable prototypes that can uncover potential issues sooner with a ubiquitous device. Had we not used a smartphone to prototype PupilScreen, for example, we would never considered how important ambient lighting or camera resolution would be for measuring the PLR without a controlled testing environment. Although today's smartphones may not be always be suitable for the solutions developed in research labs, tomorrow's ubiquitous devices can be better informed by such investigations.

The second point I want involves how results from mobile health research are communicated to non-academic researchers. I have had discussions with clinicians who overestimate the power of machine learning and computer vision, claiming that algorithms should always able to pick on "invisible" signals. I have had conversations with media coordinators who have wanted to stretch the truth about what a project is capable of, talking about what the technology may be able to do in the future rather than what it is currently able to do. I have had people send me their medical health record after seeing those press releases, begging that they can use my projects to validate their concerns. Unfortunately, news outlets are incentivized to write catch headlines that grab readers' attention, even if that means stretching the truth. As researchers, we tend to be complicit with this because more coverage means that we are demonstrating broader impact with our research.

I believe that the blame does not rest on a particular group of people. Instead, I believe that we should all reflect on how research is communicated to broader audiences, especially when that research has the potential for real-world impact. There

are many benefits to teaching people about what is possible with technology, such as getting children excited about engineering and preparing people for what may become more mainstream in the future. However, we need to strike a balance between making content accessible and staying true what has actually been achieved.

To avoid ending on a pessimistic note, I want to re-iterate the strong potential I see in the mobile health space. There is a clear demand for at-home health screening tools. Talking with clinicians, industry leaders, and policymakers has confirmed my excitement in the potential that mobile health has towards changing people's lives and improving health outcomes. We are on the verge of finding ways to support telemedicine and community healthcare workers in ways that may have seemed impossible 50 years ago. It is my hope that researchers continue to join the mobile health movement so that, one day, medical screening will be just a download away.

Appendix A

SURVEY INSTRUMENT FOR ASSESSING PERCEPTION OF HEALTH-SCREENING TECHNOLOGIES

Note: This survey instrument can be used to explore ubiquitous health sensing technologies intended for a variety of medical conditions. To illustrate how the survey instrument is meant to be used, however, this instantiation uses skin cancer as the medical condition of interest. Other details that are meant to be completed by the researcher are indicated with double brackets (e.g., [[text]]*).*

Consent

[[CONSENT DETAILS AS REQUIRED BY INSTITUTION]]

By clicking next, you agree:

- That you are at least 18 years of age,
- That you are participating in this study,
- That you understand you can withdraw from the survey at any time, and
- That you should refrain from providing identifiable data in open-ended questions.

Required questions will be marked with an asterisk (*) sign.

To begin the survey, **please click the "Next" button**.

Exclusionary Criteria

1. What platform does your primary smartphone run on?

- ○ iOS (iPhone)
- ○ Android
- ○ Windows
- ○ Blackberry
- ○ Other
- ○ I do not own a smartphone

2. How would you rate your expertise with using a smartphone (e.g., using various apps, changing settings, etc.)?

- ○ Novice / Beginner
- ○ Intermediate
- ○ Expert / Advanced

Survey Introduction

[[PURPOSE OF THE STUDY]]

Disclaimer: Any information you see in this survey (e.g., accuracy numbers, costs, diagnostic tests) should not be taken as medical advice as this information may not match the standard for any current or future procedures. Furthermore, this survey involves hypothetical scenarios with diagnostic aid smartphone apps that may not currently exist. For any medical concerns, consult your physician or doctor.

To confirm that you have read and understood the disclaimer above, please click the box below:

☐ I have read and understood that the medical information presented in this survey should not be taken as advice.

Medical Condition Familiarity

3. How familiar, if at all, are you with skin cancer?

Not familiar at all: only heard the name, if at all
Familiar: familiar with some of the causes, symptoms, or treatments
Extremely familiar: familiar with all of the causes, symptoms, or treatments

Not familiar at all	Slightly familiar	Somewhat familiar	Familiar	Moderately familiar	Very familiar	Extremely familiar
○	○	○	○	○	○	○

4. To the best of your knowledge, have you, a family member, or a close acquaintance had skin cancer in the past 3 years?

Yes, I have or have had it in the past 3 years	Yes, a family member or close acquaintance has or has had it in the past 3 years	Yes, I and other family members or close acquaintances have or have had it in the past 3 years	No, neither me nor nobody I am close to has or has had it in the past 3 years	I am not sure	Prefer not to answer
○	○	○	○	○	○

Baseline HBM Measurements

Skin cancer is the most common form of cancer in the United States. Overexposure to ultraviolet (UV) light from the sun is the major cause of skin cancer. More information about the condition can be found at: https://www.cdc.gov/cancer/skin/basic_info/what-is-skin-cancer.htm.

5. If you had skin cancer, how much effect, if at all any, do you think it would have on your **long-term health**?

No effect at all	Very little effect	Little effect	Moderate effect	Large effect	Very large effect	Extreme effect
○	○	○	○	○	○	○

6. If you had skin cancer, how much effect, if at all any, do you think it would have on your **finances**?

No effect at all	Very little effect	Little effect	Moderate effect	Large effect	Very large effect	Extreme effect
○	○	○	○	○	○	○

7. If you had skin cancer, how much effect, if at all any, do you think it would have on you **socially and/or professionally**?

No effect at all	Very little effect	Little effect	Moderate effect	Large effect	Very large effect	Extreme effect
○	○	○	○	○	○	○

8. How **beneficial**, if at all, do you think each of these actions would be towards improving your skin cancer?

	Not at all	Slightly beneficial	Somewhat beneficial	Beneficial	Moderately beneficial	Very beneficial	Extremely beneficial
Schedule an appointment with a doctor / physician	○	○	○	○	○	○	○

Contact a doctor's / physician's office for advice	○	○	○	○	○	○	○

9. How **easy or difficult** do you think it would be for you to take each of the following actions?

	Very difficult	Somewhat difficult	Slightly difficult	Neither difficult nor easy	Slightly easy	Somewhat easy	Very easy
Schedule an appointment with a doctor / physician	○	○	○	○	○	○	○
Contact a doctor's / physician's office for advice	○	○	○	○	○	○	○

Scenario Description

Imagine yourself in the following scenario and answer the following questions accordingly.

You recently noticed a new mole (beauty mark) on your arm that is oddly colored and misshapen. After looking up information online, you worry that you might be developing skin cancer.

10. Please check all of the symptoms that were mentioned in the scenario you just read:

- ☐ Oddly colored mole
- ☐ Knee swelling
- ☐ Stiff neck
- ☐ Frequent urination

11. How likely or unlikely do you think you are to have skin cancer in this scenario?

Very unlikely	Somewhat unlikely	Slightly unlikely	Neither unlikely nor likely	Slightly likely	Somewhat likely	Very likely
○	○	○	○	○	○	○

12. Given the possibility that you might have skin cancer, which of the following actions would you plan to take **on the same day** as when you discovered your symptoms?

	No / Probably No	Yes / Probably Yes
Schedule an appointment with a doctor/physician	○	○
Contact a doctor's/physician's office for advice	○	○

App Description

A smartphone app named SkinCheck analyzes a picture of a mole to determine whether or not it is cancerous. To use the app, you are asked to take a picture of the mole so that it is clearly visible. The app guides you through taking a picture so that it can see the mole clearly and at a proper distance.

SkinCheck comes with your smartphone by default as part of a new mobile health initiative by [[RESPONDENT'S PHONE COMPANY]]. SkinCheck provides text-based and audio-based instructions to help you perform the test. The app also checks that the test was performed correctly. You can repeat the test until the app determines the image to be "valid". The results of the test are available instantly.

13. How likely or unlikely would you be to use SkinCheck to check your symptoms from the scenario?

Very unlikely	Somewhat unlikely	Slightly unlikely	Neither unlikely nor likely	Slightly likely	Somewhat likely	Very likely
○	○	○	○	○	○	○

"Positive" Test Result

Note: This screen is only shown if the user's answer to Question 13 was one of the following: "Neither unlikely nor likely", "Slightly likely", "Somewhat likely", or "Very likely".

Imagine that you used SkinCheck and it said that the mark on your arm was **normal**. The app emphasizes that although it cannot guarantee that you do not have skin cancer, the test result suggests that the mark on your arm is inconsistent with those of people with skin cancer. With this test result in mind, please answer the same questions we asked you before.

14. How likely or unlikely do you think you are to have skin cancer in this scenario?

Very unlikely	Somewhat unlikely	Slightly unlikely	Neither unlikely nor likely	Slightly likely	Somewhat likely	Very likely
○	○	○	○	○	○	○

15. Given the possibility that you might have skin cancer, which of the following actions would you plan to take **on the same day** as when you discovered your symptoms?

	No / Probably No	Yes / Probably Yes
Schedule an appointment with a doctor/physician	○	○
Contact a doctor's/physician's office for advice	○	○

"Negative" Test Result

Note: This screen is only shown if the user's answer to Question 13 was one of the following: "Neither unlikely nor likely", "Slightly likely", "Somewhat likely", or "Very likely".

Imagine that you used SkinCheck and it said that the mark on your arm was **abnormal**. The app emphasizes that although it cannot guarantee that you have pink eye, the test result suggests that you may want to seek medical care. With this test result in mind, please answer the same questions we asked you before.

16. How likely or unlikely do you think you are to have skin cancer in this scenario?

Very unlikely	Somewhat unlikely	Slightly unlikely	Neither unlikely nor likely	Slightly likely	Somewhat likely	Very likely
○	○	○	○	○	○	○

17. Given the possibility that you might have skin cancer, which of the following actions would you plan to take **on the same day** as when you discovered your symptoms?

	No / Probably No	Yes / Probably Yes
Schedule an appointment with a doctor/physician	○	○
Contact a doctor's/physician's office for advice	○	○

Post-Survey Questionnaire Part 1

We have a few more questions before the survey is complete. Please answer the following questions so we can understand the factors that may have influenced your decisions.

18. In your opinion how would you rate **clinical tests** and **diagnostic aid apps** on the following factors?

	Very much so			No difference			Very much so	
I believe clinical tests are faster for getting results	○	○	○	○	○	○	○	I believe diagnostic aid apps are faster for getting results
I believe clinical tests are less expensive	○	○	○	○	○	○	○	I believe diagnostic aid apps are less expensive
I believe clinical tests keep records more private	○	○	○	○	○	○	○	I believe diagnostic aid apps keep records

								more private

19. Can you think of any other information which might have changed the way you made your decisions in the previous scenarios?

20. How many courses have you taken involving statistics (pick the highest level)?

- No courses
- A high school course
- Multiple high school courses
- A single college / university course
- Multiple college / university courses

21. How often do you use statistics in your daily life?

- Never
- Rarely
- Sometimes
- Often
- All of the time

22. Out of 1,000 people in a small town, 500 are members of a choir. Out of these 500 members in the choir, 100 are men. Out of the 500 inhabitants that are not in the choir, 300 are men. What is the probability that a randomly drawn man is a member of the choir?

Please indicate the probability as a percentage without the percent sign (0-100).

Post-Survey Questionnaire Part 2

Please answer the following demographic questions about yourself to the best of your abilities. Your answers will not be connected to any names, emails, or other personally-identifiable information.

23. What gender do you identify as?

- Male
- Female
- Transgender male
- Transgender female
- Gender variant / non-conforming
- Self-identify
- Prefer not to answer

24. What is your age?

- 18-24
- 25-34
- 35-44
- 45-54
- 55-64
- 65+
- Prefer not to answer

25. What race and/or ethnicity do you identify as?

- White (e.g., German, Irish, English, Italian, Polish, French, etc.)
- Hispanic, Latino, or Spanish origin (e.g., Mexican or Mexican America, Puerto Rican, Cuban, Salvadoran, Dominican, Columbian, etc.)
- Black or African American (e.g., African America, Jamaican, Haitian, Nigerian, Ethiopian, Somalian, etc.)
- Asian (e.g., Chinese, Filipino, Asian Indian, Vietnamese, Korean, Japanese, etc.)
- American Indian or Alaska Native (e.g., Navajo Nation, Blackfeet Tribe, Mayan, Aztec, Nome Eskimo Community, etc.)
- Middle Eastern or North African (e.g., Lebanese, Iranian, Egyptian, Syrian, Moroccan, Algerian, etc.)
- Native Hawaiian or Other Pacific Islander (e.g., Native Hawaiian, Samoan, Chamorro, Tongan, Fijian, Marshallese, etc.)

- Some other race, ethnicity, or origin
- Prefer not to answer

26. What is your current marital status?

- Married / domestic partner
- Widowed
- Divorced
- Separated
- Single / never married
- Prefer not to answer

27. Do you have any children?

- Yes
- No
- Prefer not to answer

28. In which country do you reside?

29. What is your highest level of education completed?

- Less than high school
- Graduated high school
- Trade/technical school
- Some college, no degree
- Associate degree
- Bachelor's degree
- Advanced degree (Masterís, PhD, MD)
- Prefer not to answer

30. What is your current estimated annual household income?

- Less than $25,000

- ○ $25,000-$34,999
- ○ $35,000-$49,999
- ○ $50,000-$74,999
- ○ $75,000-$99,999
- ○ $100,000-$124,999
- ○ $125,000-$149,999
- ○ $150,000 or more
- ○ Prefer not to answer

BIBLIOGRAPHY

[1] Alireza Abdolvahabi et al. "Colorimetric and longitudinal analysis of leukocoria in recreational photographs of children with retinoblastoma." In: *PloS one* 8.10 (2013). Ed. by Sanjoy Bhattacharya, e76677. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0076677. URL: http://dx.plos.org/10.1371/journal.pone.0076677http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3813630{\&}tool=pmcentrez{\&}rendertype=abstract{\%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/24204654.

[2] Icek Ajzen. "Models of human social behavior and their application to health psychology". In: *Psychology and Health* 13.4 (1998), pp. 735–739. ISSN: 08870446. DOI: 10.1080/08870449808407426. URL: http://www.tandfonline.com/doi/abs/10.1080/08870449808407426.

[3] Cheryl S Alexander and Henry Jay Becker. "The Use of Vignettes in Survey Research". In: *Public Opinion Quarterly* 42.1 (1978), p. 93. ISSN: 0033362X. DOI: 10.1086/268432. URL: https://academic.oup.com/poq/article-lookup/doi/10.1086/268432.

[4] AliveCor. *KardiaMobile*. 2018. URL: https://www.alivecor.com/?gclid=CMnU-ZbLwtkCFcEFfwodHOICpQ{\&}gclsrc=ds (visited on 02/25/2018).

[5] American Cancer Society. *Cancer Facts & Figures 2016*. Tech. rep. Atlanta, GA: American Cancer Society, 2016, pp. 1–72.

[6] John C Andrefsky, Jeffrey I Frank, and Douglas Chyatte. "The ciliospinal reflex in pentobarbital coma." In: *Journal of neurosurgery* 90.4 (1999), pp. 644–646. ISSN:

0022-3085. DOI: 10.3171/jns.1999.90.4.0644. URL: http://thejns.org/doi/abs/10.3171/jns.1999.90.4.0644.

[7] Christopher J Armitage and Mark Conner. *Social cognition models and health behaviour: A structured review*. 2000. DOI: 10.1080/08870440008400299. URL: http://www.tandfonline.com/doi/abs/10.1080/08870440008400299.

[8] Caroline Asiimwe et al. "Use of an innovative, affordable, and open-source short message service-based tool to monitor malaria in remote areas of Uganda". In: *American Journal of Tropical Medicine and Hygiene* 85.1 (2011), pp. 26–33. ISSN: 00029637. DOI: 10.4269/ajtmh.2011.10-0528. URL: http://www.ajtmh.org/content/85/1/26.short.

[9] Christiane Atzmüller and Peter M Steiner. "Experimental vignette studies in survey research". In: *Methodology* 6.3 (2010), pp. 128–138. ISSN: 16141881. DOI: 10.1027/1614-2241/a000014. URL: https://econtent.hogrefe.com/doi/10.1027/1614-2241/a000014.

[10] Lucas M Bachmann et al. "Vignette studies of medical choice and judgement to study caregivers' medical decision behaviour: Systematic review". In: *BMC Medical Research Methodology* 8.1 (2008), p. 50. ISSN: 14712288. DOI: 10.1186/1471-2288-8-50. URL: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-8-50.

[11] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation". In: *CVPR '15*. 2015, p. 5. DOI: 10.1103/PhysRevX.5.041024. arXiv: 1505.0729. URL: http://arxiv.org/abs/1511.00561http://arxiv.org/abs/1505.0729{\%}5Cnhttp://mi.eng.cam.ac.uk/projects/segnet/.

[12] Leonard Banco and Daniel Veltri. "Ability of mothers to subjectively assess the presence of fever in their children." In: *American journal of diseases of children* 138.10 (1984), pp. 976–8. ISSN: 0002-922X. DOI: 10.1001/archpedi.1984.02140480078024. URL: http://archpedi.jamanetwork.com/article.aspx?doi=10.1001/archpedi.1984.02140480078024.

[13] K Banitsas et al. "A simple algorithm to monitor HR for real time treatment applications". In: *2009 9th International Conference on Information Technology and Applications in Biomedicine* 44 (2009), pp. 1–5. DOI: 10.1109/ITAB.2009.5394308. URL: https://ieeexplore.ieee.org/abstract/document/5394308/http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5394308.

[14] Barbara J Barchiesi, Robert H Eckel, and Phillip P Ellis. "The cornea and disorders of lipid metabolism published erratum appears in Surv Ophthalmol 1992 Jan-Feb;36(4):324". In: *Survey of Ophthalmology* 36.1 (1991), pp. 1–22. URL: http://www.sciencedirect.com/science/article/pii/003962579190205Thttp://fox.novo.dk/netacgi/getref.pl?ref=M-92022995.

[15] Andrew Bastawrous et al. "Development and Validation of a Smartphone-Based Visual Acuity Test (Peek Acuity) for Clinical Practice and Community-Based Fieldwork". In: *JAMA Ophthalmology* 133.8 (2015), p. 930. DOI: 10.1001/jamaophthalmol.2015.1468. URL: http://archopht.jamanetwork.com/article.aspx?doi=10.1001/jamaophthalmol.2015.1468.

[16] David W Bates et al. "Big data in health care: Using analytics to identify and manage high-risk and high-cost patients". In: *Health Affairs* 33.7 (2014), pp. 1123–1131. ISSN: 15445208. DOI: 10.1377/hlthaff.2014.0041. URL: http://content.healthaffairs.org/cgi/doi/10.1377/hlthaff.2014.0041.

[17] Lu-Ann F Beeckman-Wagner and Diana Freeland. "Spirometry quality assurance: common errors and their impact on test results". In: (2012). URL: https://stacks.cdc.gov/view/cdc/11759.

[18] Matthias Behrends, Claus U Niemann, and Merlin D Larson. "Infrared pupillometry to detect the light reflex during cardiopulmonary resuscitation: A case series". In: *Resuscitation* 83.10 (2012), pp. 1223–1228. ISSN: 03009572. DOI: 10.1016/j.resuscitation.2012.05.013. URL: http://www.sciencedirect.com/science/article/pii/S0300957212002638.

[19] Amit Asish Bhadra, Manu Jain, and Sushila Shidnal. "Automated detection of eye diseases". In: *Proceedings of the 2016 IEEE International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2016*. IEEE, 2016, pp. 1341–1345. ISBN: 9781467393379. DOI: 10.1109/WiSPNET.2016.7566355. URL: http://ieeexplore.ieee.org/document/7566355/.

[20] Vinod K Bhutani, Lois Johnson, and Emidio M Sivieri. "Predictive ability of a predischarge hour-specific serum bilirubin for subsequent significant hyperbilirubinemia in healthy term and near-term newborns". In: *Pediatrics* 103.1 (1999), pp. 6–14. URL: http://pediatrics.aappublications.org/content/103/1/6.short.

[21] Barry P Boden et al. "Catastrophic Head Injuries in High School and College Football Players". In: *The American Journal of Sports Medicine* 35.7 (2007), pp. 1075–1081. ISSN: 0363-5465. DOI: 10.1177/0363546507299239. URL: http://ajs.sagepub.com/lookup/doi/10.1177/0363546507299239.

[22] Angel N. Boev et al. "Quantitative pupillometry: normative data in healthy pediatric volunteers". In: *Journal of Neurosurgery: Pediatrics* 103.6 (2005), pp. 496–500. DOI: 10.3171/ped.2005.103.6.0496. URL: http://thejns.org/doi/abs/10.3171/ped.2005.103.6.0496.

[23] Giles Bond-Smith et al. "Pancreatic adenocarcinoma". In: *BMJ* 344 (2012). DOI: 10.1136/bmj.e2476.

[24] Brian M Bot et al. "The mPower study, Parkinson disease mobile data collected using ResearchKit". In: *Scientific Data* 3 (2016), p. 160011. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.11. URL: http://www.ncbi.nlm.nih.gov/pubmed/26938265http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4776701http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4776701{\&}tool=pmcentrez{\&}rendertype=abstract.

[25] Yuri Y Boykov and Marie-Pierre Jolly. "Interactive graph cuts for optimal boundary & region segmentation of objects in ND images". In: *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*. Vol. 1. July. IEEE, 2001, pp. 105–112. ISBN: 0-7695-1143-0. DOI: 10.1109/ICCV.2001.937505. URL: http://ieeexplore.ieee.org/document/937505/.

[26] Steven P Broglio et al. "Test-retest reliability of computerized concussion assessment programs". In: *Journal of Athletic Training* 42.4 (2007), pp. 509–514. ISSN: 10626050. DOI: 10.1016/S0162-0908(09)79464-4. URL: http://www.ncbi.nlm.nih.gov/pubmed/18174939.

[27] Egon Brunswik. "The conceptual framework of psychology". In: *Psychological Bulliten* 49.6 (1952), pp. 654–656. URL: https://insights.ovid.com/plbul/195211000/00006823-195211000-00014.

[28] Angela Bryan, Sarah J Schmiege, and Michelle R. Broaddus. "Mediational Analysis in HIV/AIDS Research: Estimating Multivariate Path Analytic Models in a Structural Equation Modeling Framework". In: *AIDS and Behavior* 11.3 (2007), pp. 365–383. ISSN: 1090-7165. DOI: 10.1007/s10461-006-9150-2. URL: http://link.springer.com/10.1007/s10461-006-9150-2.

[29] Carla. Cantor and Brian Fallon. *Phantom illness : recognizing, understanding, and overcoming hypochondria*. Houghton Mifflin, 1996, p. 351. ISBN: 0395859921.

[30] Jose E Capó-Aponté et al. "Pupillary Light Reflex as an Objective Biomarker for Early Identification of Blast-Induced mTBI". In: *Journal of Spine* (2013). ISSN: 21657939. DOI: 10.4172/2165-7939.S4-004. URL: http://www.omicsgroup.org/journals/pupillary-light-reflex-as-an-objective-biomarker-for-early-identification-of-blast-induced-mtbi-2165-7939.S4-004.php?aid=19402.

[31] Centers for Disease Control. "Nonfatal traumatic brain injuries related to sports and recreation activities among persons aged 19 years—United States, 2001–2009". In: *MMWR: Morbidity and mortality weekly report* 60.39 (2011), pp. 1337–1342. URL: http://www.safetylit.org/citations/index.php?fuseaction=citations.viewdetails{\&}citationIds{\%}255B{\%}255D=citjournalarticle{_}325399{_}23.

[32] Centers for Disease Control. *TBI: Get the Facts*. 2016. URL: http://www.cdc.gov/traumaticbraininjury/get{_}the{_}facts.html.

[33] Victoria L. Champion. "Revised susceptibility, benefits, and barriers scale for mammography screening". In: *Research in Nursing and Health* 22.4 (1999), pp. 341–348. ISSN: 01606891. DOI: 10.1002/(SICI)1098-240X(199908)22:4<341::AID-NUR8>3.0.CO;2-P. URL: http://doi.wiley.com/10.1002/{\%}28SICI{\%}291098-240X{\%}28199908{\%}2922{\%}3A4{\%}3C341{\%}3A{\%}3AAID-NUR8{\%}3E3.0.CO{\%}3B2-P.

[34] Victoria L. Champion. "The relationship of breast self-examination to health belief model variables". In: *Research in Nursing & Health* 10.6 (1987), pp. 375–382. ISSN: 1098240X. DOI: 10.1002/nur.4770100605. URL: http://doi.wiley.com/10.1002/nur.4770100605.

[35] Rakesh K Chandra, Monica O Patadia, and Joey Raviv. "Diagnosis of Nasal Airway Obstruction". In: *Otolaryngologic Clinics of North America* 42.2 (2009), pp. 207–225. ISSN: 00306665. DOI: 10.1016/j.otc.2009.01.004. URL: https://www.sciencedirect.com/science/article/pii/S0030666509000073http://www.embase.com/search/results?subaction=viewrecord{\&}from=export{\&}id=L354355842{\%}0Ahttp://dx.doi.org/10.1016/j.otc.2009.01.004.

[36] Rama Chellappa. "The changing fortunes of pattern recognition and computer vision". In: *Image and Vision Computing* 55.1 (2016), pp. 3–5. ISSN: 02628856. DOI: 10.1016/j.imavis.2016.04.005. URL: http://www.sciencedirect.com/science/article/pii/S026288561630066X.

[37] Wynne W Chin. "Issues and opinion on structural equation modelling". In: *Management Information Systems quarterly* 22.1 (1998), pp. 1–8.

[38] Kun Woo Cho et al. "Gaze-Wasserstein: a quantitative screening approach to autism spectrum disorders". 2016. URL: http://ieeexplore.ieee.org/abstract/document/7764551/.

[39] Rumi Chunara et al. "Flu Near You: An Online Self-reported Influenza Surveillance System in the USA". In: *Online Journal of Public Health Informatics* 5.1 (2013). ISSN: 1947-2579. DOI: 10.5210/ojphi.v5i1.4456. URL: https://ojphi.org/ojs/index.php/ojphi/article/view/4456http://journals.uic.edu/ojs/index.php/ojphi/article/view/4456.

[40] Edward T Cokely et al. "Measuring risk literacy: The Berlin numeracy test". In: *Judgment and Decision Making* 7.1 (2012), p. 25.

[41] Heather Cole-Lewis and Trace Kershaw. *Text messaging as a tool for behavior change in disease prevention and management*. 2010. DOI: 10.1093/epirev/mxq004. URL: https://academic.oup.com/epirev/article-lookup/doi/10.1093/epirev/mxq004.

[42] Sunny Consolvo et al. "Activity Sensing in the Wild: A Field Trial of UbiFit Garden". In: *Proc. CHI '08*. 2008, pp. 1797–1806. ISBN: 9781605580111. DOI: 10.1145/1357054.1357335. URL: http://dl.acm.org/citation.cfm?id=1357335http://portal.acm.org/citation.cfm?doid=1357054.1357335{\%}5Cnhttp://portal.acm.org/citation.cfm?id=1357335.

[43] Sunny Consolvo et al. "Design requirements for technologies that encourage physical activity". In: *Proc. CHI '06*. 2006, p. 457. ISBN: 1595933727. DOI: 10.1145/1124772.1124840. arXiv: 9809069v1 [arXiv:gr-qc]. URL: http://dl.acm.org/citation.cfm?id=1124840http://portal.acm.org/citation.cfm?doid=1124772.1124840.

[44] Ray W Cooksey. *Judgment analysis: Theory, methods, and applications*. 1996. URL: https://psycnet.apa.org/record/1996-97447-000.

[45] David Couret et al. "Reliability of standard pupillometry practice in neurocritical care: an observational, double-blinded study". In: *Critical Care* 20.1 (2016), p. 99. DOI: 10.1186/s13054-016-1239-z. URL: http://www.ncbi.nlm.nih.gov/pubmed/27072310http://www.ncbi.nlm.nih.gov/pubmed/27072310http://www.ncbi.nlm.nih.gov/pubmed/27072310.

[46] Simona Crihalmeanu and Arun Ross. "Multispectral scleral patterns for ocular biometric recognition". In: *Pattern Recognition Letters* 33.14 (2012), pp. 1860–1869. ISSN: 01678655. DOI: 10.1016/j.patrec.2011.11.006.

[47] S Crowe et al. "Development of a Rapid Point-of-Care Immunochromatographic Test for Measurement of CD4 T-cells". In: 185 (2008).

[48] Joan M Daisey, William J Angell, and Michael G Apte. *Indoor air quality, ventilation and health symptoms in schools: An analysis of existing information*. 2003. DOI: 10.1034/j.1600-0668.2003.00153.x. URL: http://doi.wiley.com/10.1034/j.1600-0668.2003.00153.x.

[49] John Danias et al. “Method for the noninvasive measurement of intraocular pressure in mice”. In: *Investigative Ophthalmology and Visual Science* 44.3 (2003), pp. 1138–1141. ISSN: 01460404. DOI: 10.1167/iovs.02-0553. URL: http://arvojournals.org/article.aspx?articleid=2124404.

[50] Abhijit Das et al. “A new efficient and adaptive sclera recognition system”. In: *2014 IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM)*. IEEE, 2014, pp. 1–8. ISBN: 978-1-4799-4533-7. DOI: 10.1109/CIBIM.2014.7015436. URL: http://ieeexplore.ieee.org/document/7015436/.

[51] Abhijit Das et al. “Sclera recognition using dense-SIFT”. In: *International Conference on Intelligent Systems Design and Applications, ISDA*. IEEE, 2014, pp. 74–79. ISBN: 9781479935161. DOI: 10.1109/ISDA.2013.6920711. URL: http://ieeexplore.ieee.org/document/6920711/.

[52] Nediyana Daskalova et al. “SleepCoacher: A Personalized Automated Self-Experimentation System for Sleep Recommendations”. In: *Proc. UIST '16* (2016), pp. 347–358. DOI: 10.1145/2984511.2984534. URL: https://dl.acm.org/citation.cfm?id=2984534http://doi.acm.org/10.1145/2984511.2984534.

[53] Fred D. Davis. “Perceived ease of use, and user acceptance of information technology”. In: *MIS Quarterly* 13 (1989), pp. 319–340. ISSN: 02767783. DOI: 10.2307/249008. URL: http://www.jstor.org/stable/249008.

[54] Wändi Bruine De Bruin et al. “Verbal and Numerical Expressions of Probability: "It's a Fifty-Fifty Chance"”. In: *Organizational Behavior and Human Decision Processes* 81.1 (2000), pp. 115–131. ISSN: 07495978. DOI: 10.1006/obhd.1999.2868. URL: https://www.sciencedirect.com/science/article/pii/S0749597899928686.

[55] Maria Syl D De La Cruz, Alisa P Young, and Mack T Ruffin. "Diagnosis and management of pancreatic cancer." In: *American family physician* 89.8 (2014), pp. 626–32. ISSN: 1532-0650. URL: http://www.ncbi.nlm.nih.gov/pubmed/24784121.

[56] Nicola Dell, Dean Stevens, and Paul Yager. "Towards a Point-of-Care Diagnostic System : Automated Analysis of Immunoassay Test Data on a Cell Phone". In: *Networked Systems for Developing Regions*. June. New York, New York, USA: ACM Press, 2011, pp. 3–8. ISBN: 9781450307390. DOI: 10.1145/1999927.1999931. URL: http://portal.acm.org/citation.cfm?doid=1999927.1999931.

[57] Brian DeRenzi et al. "e-IMCI: Improving Pediatric Health Care in Low-Income Countries". In: *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*. 2008, p. 753. ISBN: 9781605580111. DOI: 10.1145/1357054.1357174. URL: http://dl.acm.org/citation.cfm?id=1357174http://dl.acm.org/citation.cfm?id=1357054.1357174{\%}5Cnhttp://portal.acm.org/citation.cfm?doid=1357054.1357174{\%}5Cnhttp://dl.acm.org/citation.cfm?id=1357174{\%}5Cnhttp://portal.acm.org/citation.cfm?doid=1357054.1357174.

[58] Tawanna R Dillahunt and Amelia R Malone. "The Promise of the Sharing Economy among Disadvantaged Communities". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. 2015, pp. 2285–2294. ISBN: 9781450331456. DOI: 10.1145/2702123.2702189. URL: http://dl.acm.org/citation.cfm?id=2702189http://dl.acm.org/citation.cfm?doid=2702123.2702189.

[59] H Dubey et al. "EchoWear: Smartwatch technology for voice and speech treatments of patients with Parkinson's disease". In: *Proceedings - Wireless Health 2015, WH 2015*. 2015. ISBN: 9781450338516. DOI: 10.1145/2811780.2811957. URL: http://dl.acm.org/citation.cfm?id=2811957.

[60] Mary T Dzindolet et al. "The role of trust in automation reliance". In: *International Journal of Human-Computer Studies* 58.6 (2003), pp. 697–718. ISSN: 10715819. DOI: 10.1016/S1071-5819(03)00038-7. URL: https://www.sciencedirect.com/science/article/pii/S1071581903000387.

[61] Matthew S Eastin. "Credibility Assessments of Online Health Information: The Effects of Source Expertise and Knowledge of Content". In: *Journal of Computer-Mediated Communication* 6.4 (2010), pp. 0–0. ISSN: 10836101. DOI: 10.1111/j.1083-6101.2001.tb00126.x. URL: https://academic.oup.com/jcmc/article/4584226.

[62] C J Ellis. "The pupillary light reflex in normal subjects." In: *The British journal of ophthalmology* 65.11 (1981), pp. 754–9. ISSN: 0007-1161. DOI: 10.1136/bjo.65.11.754. URL: http://bjo.bmj.com/content/65/11/754.shorthttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1039657{\&}tool=pmcentrez{\&}rendertype=abstract.

[63] Lennart A Ericson. "Twenty-four hourly variations of the aqueous flow; examinations with perilimbal suction cup." In: *Acta ophthalmologica. Supplementum* 37.Suppl 50 (1958), pp. 1–95. ISSN: 0065-1451. URL: https://www.mysciencework.com/publication/show/058d758800844c1da83a90e859f9271f.

[64] E. R. Ettinger, H. J. Wyatt, and R London. "Anisocoria: Variation and clinical observation with different conditions of illumination and accommodation". In: *Investigative Ophthalmology and Visual Science* 32.3 (1991), pp. 501–509. ISSN: 01460404. URL: http://www.ncbi.nlm.nih.gov/pubmed/2001925.

[65] Spencer C. Evans et al. "Vignette methodologies for studying clinicians' decision-making: Validity, utility, and application in ICD-11 field studies". In: *International Journal of Clinical and Health Psychology* 15.2 (2015), pp. 160–170. ISSN: 16972600. DOI: 10.1016/j.ijchp.2014.12.001. URL: https://www.

sciencedirect.com/science/article/pii/S1697260014000660{\#}bib0005.

[66] Mark Faul et al. *Traumatic Brain Injury in the United States*. Tech. rep. 2010. URL: http://origin.glb.cdc.gov/traumaticbraininjury/pdf/blue{_}book.docx.

[67] Quentin Ferry et al. "Diagnostically relevant facial gestalt information from ordinary photos". In: *eLife* 2014.3 (2014). ISSN: 2050084X. DOI: 10.7554/eLife.02020.001. URL: https://elifesciences.org/content/3/e02020v1.

[68] JD Fischer and DJ van den Heever. "Portable video-oculography device for implementation in side-line concussion assessments: a prototype". In: *Proc. EMBC '16*. 2016.

[69] Martin Fishbein. "Attitude and the prediction of behavior". In: *Readings in attitude theory and measurement* (1967), pp. 477–492.

[70] C. M. Fisher. "Oval pupils." In: *Archives of neurology* 37.8 (1980), pp. 502–503. ISSN: 0003-9942. DOI: 10.1001/archneur.1980.00500570050007. URL: http://archneur.jamanetwork.com/article.aspx?articleid=578873http://www.ncbi.nlm.nih.gov/pubmed/7417041.

[71] W A Fletcher and J A Sharpe. "Tonic pupils in neurosyphilis." In: *Neurology* 36.2 (1986), pp. 188–92. ISSN: 0028-3878. URL: http://www.neurology.org/content/36/2/188.shorthttp://www.ncbi.nlm.nih.gov/pubmed/3945389.

[72] Brian J Fogg. *Persuasive Technology: Using Computers to Change What We Think and Do*. Vol. 5. 1. 2003, p. 283. ISBN: 1558606432. DOI: 10.4017/gt.2006.05.01.009.00. arXiv: 9780201398298. URL: http://dl.acm.org/citation.cfm?id=763957.

[73] Susannah Fox and Maeve Duggan. *Mobile health 2012*. Tech. rep. 2012. URL: http://emr-matrix.org/wp-content/uploads/2012/12/PIP{_}MobileHealth2012.pdf.

[74] Orrin I Franko, Christopher Bray, and Peter O Newton. "Validation of a scoliometer smartphone app to assess scoliosis". In: *Journal of Pediatric Orthopaedics* 32.8 (2012), e72–e75. URL: http://journals.lww.com/pedorthopaedics/Abstract/2012/12000/Validation{_}of{_}a{_}Scoliometer{_}Smartphone{_}App{_}to.13.aspx.

[75] Caroline Free et al. *Smoking cessation support delivered via mobile phone text messaging (txt2stop): A single- blinded, randomized trial.* 2011. DOI: 10.1016/S0140-6736(11)60701-0. URL: http://www.sciencedirect.com/science/article/pii/S0140673611607010http://ovidsp.ovid.com/ovidweb.cgi?T=JS{\&}PAGE=reference{\&}D=psyc8{\&}NEWS=N{\&}AN=2011-13911-033.

[76] Wolfgang Fuhl et al. "ElSe : Ellipse Selection for Robust Pupil Detection in Real-World Environments". In: *Eye Tracking Research & Applications*. 2016, pp. 123–130. ISBN: 9781450341257. DOI: 10.1145/2857491.2857505. arXiv: arXiv:1511.06575v2. URL: http://dl.acm.org/citation.cfm?id=2857505.

[77] Wolfgang Fuhl et al. "ExCuSe: Robust Pupil Detection in Real-World Scenarios". In: *International Conference on Computer Analysis of Images and Patterns*. Springer International Publishing, 2015, pp. 39–51. ISBN: 978-3-319-23192-1. DOI: 10.1007/978-3-319-23192-1_4. URL: http://link.springer.com/chapter/10.1007/978-3-319-23192-1{_}4http://link.springer.com/10.1007/978-3-319-23192-1{_}4.

[78] Wolfgang Fuhl et al. "PupilNet: Convolutional Neural Networks for Robust Pupil Detection". In: 1-10 (2016). arXiv: 1601.04902.

[79] Belinda J Gabbe et al. "Comparison of mortality following hospitalisation for isolated head injury in England and Wales, and Victoria, Australia". In: *PloS one* 6.5 (2011). DOI: 10.1371/journal.pone.0020545. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0020545.

[80] Kristin M Galetta et al. "Journal of the Neurological Sciences The King – Devick test and sports-related concussion : Study of a rapid visual screening tool in a collegiate cohort". In: *Journal of the neurological sciences* 309.1-2 (2011), pp. 34–39. ISSN: 1878-5883 (Electronic). DOI: 10.1016/j.jns.2011.07.039. URL: http://www.sciencedirect.com/science/article/pii/S0022510X11004576.

[81] Kristin M Galetta et al. "The King-Devick test of rapid number naming for concussion detection: meta-analysis and systematic review of the literature". In: *Concussion* 1.2 (2015), cnc.15.8. ISSN: 2056-3299. DOI: 10.2217/cnc.15.8. URL: http://www.futuremedicine.com/doi/abs/10.2217/cnc.15.8.

[82] Matthew S. Galetta et al. "Saccades and memory: Baseline associations of the King-Devick and SCAT2 SAC tests in professional ice hockey players". In: *Journal of the Neurological Sciences* 328.1-2 (2013), pp. 28–31. ISSN: 0022510X. DOI: 10.1016/j.jns.2013.02.008. URL: http://www.sciencedirect.com/science/article/pii/S0022510X13000853.

[83] Rocio Garcia-Retamero and Edward T Cokely. "Communicating Health Risks With Visual Aids". In: *Current Directions in Psychological Science* 22.5 (2013), pp. 392–399. ISSN: 14678721. DOI: 10.1177/0963721413491570. URL: http://journals.sagepub.com/doi/10.1177/0963721413491570.

[84] David M Gay. "Usage summary for selected optimization routines". In: *Computing science technical report* 153 (1990), pp. 1–21. URL: https://ms.mcmaster.ca/{~}bolker/misc/port.pdf.

[85] Mario E Giardini et al. "A smartphone based ophthalmoscope". In: *Proc. EMBC '14* 2014 (2014), pp. 2177–2180. ISSN: 1557170X. DOI: 10.1109/EMBC.2014.6944049. URL: http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=6944049.

[86] Mayank Goel et al. "SpiroCall: Measuring Lung Function over a Phone Call". In: *Proc. CHI '16*. 2016, pp. 5675–5685. URL: http://homes.cs.washington.edu/{~}mayank/Papers/SpiroCall.pdf.

[87] Hans Goldmann. "Applanation tonometry". In: *Transactions of the Second Glaucoma Conference*. Josiah Macy Jr. Foundation, 1957, pp. 167–219. URL: https://scholar.google.com/scholar?q=Goldmann+H.+{\%}2522Applanation+Tonometry{\%}2522.+Transactions+Second+Glaucoma+Conference.+New+York{\%}252C+Josiah+Macy{\%}252C+Jr+Foundation.+1957.{\&}btnG={\&}hl=en{\&}as{_}sdt=0{\%}252C48{\#}0.

[88] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016, p. 1. ISBN: 9781491925614. DOI: 10.1038/nmeth.3707. arXiv: arXiv:1312.6184v5. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.672.7118{\&}rep=rep1{\&}type=pdfhttp://files.sig2d.org/sig2d14.pdf{\#}page=5.

[89] Claire C Gordon et al. *2012 Anthropometric Survey of U.S . Army Personnel : Methods and Summary Statistics*. Tech. rep. December 2014. 1988, p. 640. DOI: 10.1017/S0022112088000242. URL: https://apps.dtic.mil/docs/citations/ADA611869http://tools.openlab.psu.edu/publicData/ANSURII-TR15-007.pdf.

[90] P J Grattan-Smith and W Butt. "Suppression of brainstem reflexes in barbiturate coma." In: *Archives of disease in childhood* 69.1 (1993), pp. 151–152. ISSN: 0003-9888.

DOI: 10.1136/adc.69.1.151. URL: http://adc.bmj.com/content/69/1/151.abstract.

[91] Lilian de Greef et al. "Bilicam: using mobile phones to monitor newborn jaundice". In: *Proc. UbiComp '14*. 2014, pp. 331–342. URL: http://dl.acm.org/citation.cfm?id=2632076.

[92] Paul E Green. "On the Design of Choice Experiments Involving Multifactor Alternatives". In: *Journal of Consumer Research* 1.2 (1974), pp. 61–68. ISSN: 0093-5301. DOI: 10.1086/208592. URL: https://academic.oup.com/jcr/article-lookup/doi/10.1086/208592.

[93] Dorothy M Greig, Bruce T Porteous, and Allan H Seheult. "Exact maximum a posteriori estimation for binary images". In: *Series of the Royal Statistical Society* 51.2 (1989), pp. 271–279. ISSN: 00359246. DOI: 10.2307/2345609. URL: http://www.jstor.org/stable/2345609http://www.jstor.org/stable/2345609{\%}5Cnhttp://www.jstor.org/stable/pdfplus/2345609.pdf?acceptTC=true.

[94] Domenico Grimaldi et al. "Photoplethysmography Detection by Smartphone ' s Videocamera". In: *The 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*. September. 2011, pp. 488–491. ISBN: 9781457714252. URL: http://ieeexplore.ieee.org/abstract/document/6072801/.

[95] Kenneth R Hammond, Carolyn J Hursch, and Frederick J Todd. "Analyzing the components of clinical inference". In: *Psychological Review* 71.6 (1964), pp. 438–456. ISSN: 0033295X. DOI: 10.1037/h0040736. URL: http://content.apa.org/journals/rev/71/6/438.

[96] Kimberly G Harmon et al. "American Medical Society for Sports Medicine position statement: concussion in sport". In: *British Journal of Sports Medicine* 47.1 (2013),

pp. 15–26. ISSN: 0306-3674. DOI: 10.1136/bjsports-2012-091941. URL: http://bjsm.bmj.com/lookup/doi/10.1136/bjsports-2012-091941.

[97] Joel A Harrison and Patricia D Mullen. "A meta-analysis of studies of the health belief model with adults". In: *Health Education Research* 7.1 (1992), pp. 107–116. URL: https://academic.oup.com/her/article-abstract/7/1/107/687158.

[98] Jordan Hashemi et al. "Computer vision tools for low-cost and noninvasive measurement of autism-related behaviors in infants". In: *Autism research and treatment* (2014). URL: https://www.hindawi.com/journals/aurt/2014/935686/abs/.

[99] G. Heath. "The episclera, sclera and conjunctiva. An overview of relevant ocular anatomy." In: *Differential Diagnosis of Ocular Disease* 9.2 (2006), pp. 36–42. URL: https://scholar.google.com/scholar?q=G.+Heath+The+episclera{\%}2C+sclera+and+conjunctiva{\%}2C+Optometry+today+Different.+Diagnosis+Ocular+Dis.{\%}2C+9+{\%}282{\%}29+{\%}282006{\%}29{\%}2C+pp.+36âĂŞ42{\&}btnG={\&}hl=en{\&}as{_}sdt=0{\%}2C29.

[100] Eric B Hekler et al. "Mind the Theoretical Gap: Interpreting, Using, and Developing Behavioral Theory in HCI Research". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*. 2013, pp. 3307–3316. ISBN: 9781450318990. DOI: 10.1145/2470654.2466452. arXiv: 0402594v3 [arXiv:cond-mat]. URL: https://dl.acm.org/citation.cfm?id=2466452http://dl.acm.org/citation.cfm?doid=2470654.2466452.

[101] Sonja Herdener et al. "Is the PASCAL??-Tonometer suitable for measuring intraocular pressure in clinical routine? Long- and short-term reproducibility of dynamic contour tonometry". In: *European Journal of Ophthalmology* 18.1 (2008),

pp. 39–43. ISSN: 11206721. URL: http://medlib.yu.ac.kr/eur{_}j{_}oph/ejo{_}pdf/12380.pdf.

[102] Godfrey Martin Hochbaum. *Public participation in medical screening programs: A socio-psychological study*. US Department of Health, Education, and Welfare, Public Health Service, Bureau of State Services, Division of Special Health Services, Tuberculosis Program, 1958.

[103] Sture Holm. "A Simple Sequentially Rejective Multiple Test Procedure". In: *Scandinavian Journal of Statistics* 6 (1979), pp. 65–70. ISSN: 03036898. DOI: 10.2307/4615733. arXiv: arXiv:1011.1669v3. URL: https://www.jstor.org/stable/4615733?casa{_}token=gMpAUp0IqiQAAAAA:W3XGMSr61fL2TntR{_}KGALtB11ciQ0y3xKxmoGVdipZYYvjiv5H-VsBm{_}Av55Jxeh-i{_}6wPxoSYX9M9mAJmEaPh-RHp3vVOkmVBSaIjo{_}8DEgfYteIjRo.

[104] Antoni Homs-Corbera et al. "Time-frequency detection and analysis of wheezes during forced exhalation". In: *IEEE Transactions on Biomedical Engineering* 51.1 (2004), pp. 182–186. URL: https://ieeexplore.ieee.org/abstract/document/1254008/.

[105] Daire Hooper, Joseph Coughlan, and Michael Mullen. "Structural equation modelling: Guidelines for determining model fit". In: *Articles* (2008), p. 2. URL: http://arrow.dit.ie/cgi/viewcontent.cgi?article=1001{\&}context=buschmanart.

[106] Harry Horwich and Goodwin M Breinin. "Phasic variations in tonography". In: *AMA archives of ophthalmology* 51.5 (1954), pp. 687–694. URL: http://archopht.jamanetwork.com/article.aspx?articleid=623954.

[107] Robert Hurling et al. "Using internet and mobile phone technology to deliver an automated physical activity program: Randomized controlled trial". In: *Journal of Medical Internet Research* 9.2 (2007). ISSN: 14388871. DOI: 10.2196/jmir.9.2.e7.

URL: https://www.ncbi.nlm.nih.gov/pmc/articles/pmc1874722/ https://www.scopus.com/inward/record.uri?eid=2-s2.0-34249682151{\&}doi=10.2196{\%}2Fjmir.9.2.e7{\&}partnerID=40{\&}md5=3749158623bf2698f0341d53c6939a4b.

[108] Robert Istepanian, Swarny Laxminarayan, and Constantinos S Pattichis. *M-health*. Springer, 2006. URL: http://link.springer.com/content/pdf/10.1007/b137697.pdf.

[109] N.K. Janz and M.H. Becker. "The health belief model: A decade later". In: *Health Education Quarterly* 11.1 (1984), pp. 1 –47. ISSN: 1090-1981. DOI: 10.1177/109019818401100101.

[110] Karl G Jöreskog. "A general approach to confirmatory maximum likelihood factor analysis". In: *Psychometrika* 34.2 (1969), pp. 183–202. ISSN: 00333123. DOI: 10.1007/BF02289343. URL: http://doi.wiley.com/10.1002/j.2333-8504.1967.tb00991.x.

[111] Raed A. Joundi et al. "Rapid tremor frequency assessment with the iPhone accelerometer". In: *Parkinsonism and Related Disorders* 17.4 (2011), pp. 288–290. ISSN: 13538020. DOI: 10.1016/j.parkreldis.2011.01.001. URL: http://www.sciencedirect.com/science/article/pii/S1353802011000022.

[112] Linda J Juretschke. "Kernicterus: still a concern." In: *Neonatal network* 24.2 (2005), pp. 7–19. ISSN: 0730-0832. URL: http://www.ncbi.nlm.nih.gov/pubmed/15835475.

[113] Daniel Kahneman and Amos Tversky. "Prospect Theory: An Analysis of Decision under Risk". In: *Econometrica* 47.2 (1979), p. 263. ISSN: 00129682. DOI: 10.2307/1914185. URL: http://www.worldscientific.com/doi/abs/10.1142/9789814417358{_}0006https://www.jstor.org/stable/1914185?origin=crossref.

[114] Hartmut E Kanngiesser, Christoph Kniestedt, and Yves C a Robert. "Dynamic contour tonometry: presentation of a new tonometer." In: *Journal of glaucoma* 14.5 (2005), pp. 344–50. ISSN: 1057-0829. DOI: 10.1097/01.ijg.0000176936.16015.4e. URL: http://journals.lww.com/glaucomajournal/Abstract/2005/10000/Dynamic{_}Contour{_}Tonometry{_}{_}Presentation{_}of{_}a{_}New.4.aspxhttp://www.ncbi.nlm.nih.gov/pubmed/16148581.

[115] Ravi Karkar et al. "Beacon: Designing a Portable Device for Self-Administering a Measure of Critical Flicker Frequency". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.3 (2018). DOI: 10.1145/3264927.

[116] Ravi Karkar et al. "TummyTrials: A Feasibility Study of Using Self-Experimentation to Detect Individualized Food Triggers". In: *Proc. CHI '17*. 2017, pp. 6850–6863. DOI: 10.1145/3025453.3025480. URL: https://dl.acm.org/citation.cfm?id=3025480.

[117] Moritz Kassner, William Patera, and Andreas Bulling. "Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction". In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct Publication - UbiComp '14 Adjunct*. New York, New York, USA: ACM Press, 2014, pp. 1151–1160. ISBN: 9781450330473. DOI: 10.1145/2638728.2641695. URL: http://dl.acm.org/citation.cfm?doid=2638728.2641695.

[118] Matthew Kay, Shwetak N. Patel, and Julie A. Kientz. "How Good is 85%?: A Survey Tool to Connect Classifier Evaluation to Acceptability of Accuracy". In: *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2015)*. 2015, pp. 347–356. ISBN: 9781450331456. DOI: bqd5. URL: http://dl.acm.org/citation.cfm?id=2702603.

[119] Matthew Kay et al. "There's no such thing as gaining a pound". In: *Proc. UbiComp '13*. 2013, pp. 401–410. ISBN: 9781450317702. DOI: 10.1145/2493432.2493456. URL: https://dl.acm.org/citation.cfm?id=2493456.

[120] Frank Kee et al. "Judgment analysis of prioritization decisions within a dialysis program in one United Kingdom region". In: *Medical Decision Making* 22.2 (2002), pp. 140–151. ISSN: 0272989X. DOI: 10.1177/0272989X0202200211. URL: http://journals.sagepub.com/doi/10.1177/0272989X0202200211.

[121] L W Keijsers, Martin W I M Horstink, and Stan C A M Gielen. "Ambulatory Motor Assessment in Parkinson ' s Disease". In: *Disorders* 21.1 (2006), pp. 34–44. DOI: 10.1002/mds.20633. URL: http://onlinelibrary.wiley.com/doi/10.1002/mds.20633/full.

[122] Noël L W Keijsers, Martin W I M Horstink, and Stan C a M Gielen. "Automatic assessment of levodopa-induced dyskinesias in daily life by neural networks." In: *Movement disorders : official journal of the Movement Disorder Society* 18.1 (2003), pp. 70–80. ISSN: 0885-3185. DOI: 10.1002/mds.10310. URL: http://onlinelibrary.wiley.com/doi/10.1002/mds.10310/full.

[123] Predrag Klasnja, Sunny Consolvo, and Wanda Pratt. "How to evaluate technologies for health behavior change in HCI research". In: *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*. 2011, p. 3063. ISBN: 9781450302289. DOI: 10.1145/1978942.1979396. URL: http://dl.acm.org/citation.cfm?id=1979396http://dl.acm.org/citation.cfm?doid=1978942.1979396.

[124] Rex B Kline. *Principles and practice of structural equation modeling*. Guilford Publications, 2005.

[125] Jeff Klingner, Rakshit Kumar, and Pat Hanrahan. "Measuring the task-evoked pupillary response with a remote eye tracker". In: *Proceedings of the 2008 symposium on Eye tracking research & applications - ETRA '08* 1.212 (2008), p. 69. DOI: 10.

1145/1344471.1344489. URL: http://dl.acm.org/citation.cfm?id=1344489http://portal.acm.org/citation.cfm?doid=1344471.1344489.

[126] Gudrun J Klinker, Steven A Shafer, and Takeo Kanade. "A physical approach to color image understanding". In: *International Journal of Computer Vision* 4.1 (1990), pp. 7–38. DOI: 10.1007/BF00137441. URL: http://link.springer.com/article/10.1007/BF00137441.

[127] LF Kozachenko and NN Leonenko. "Sample estimate of the entropy of a random vector". In: *Problemy Peredachi Informatsii* 23.2 (1987), pp. 9–16. ISSN: 0555-2923. URL: http://www.mathnet.ru/eng/ppi797.

[128] William H. Kruskal and W. Allen Wallis. "Use of ranks in one-criteron analysis of variance". In: *Journal of the American Statistical Association* 47.260 (1952), pp. 583–621. ISSN: 01621459. DOI: 10.2307/2280779. URL: https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1952.10483441http://www.jstor.org/stable/2280779?origin=crossref.

[129] Setor Kunutsor et al. "Using mobile phones to improve clinic attendance amongst an antiretroviral treatment cohort in rural Uganda: a cross-sectional and prospective study." In: *AIDS and behavior* 14.6 (2010), pp. 1347–1352. ISSN: 1573-3254 (Electronic). DOI: 10.1007/s10461-010-9780-2. URL: http://link.springer.com/10.1007/s10461-010-9780-2.

[130] Eric C Larson et al. "Accurate and privacy preserving cough sensing using a low-cost microphone". In: *Proc. UbiComp '11*. 2011, p. 375. ISBN: 9781450306300. DOI: 10.1145/2030112.2030163. URL: http://dl.acm.org/citation.cfm?id=2030163http://dl.acm.org/citation.cfm?doid=2030112.2030163.

[131] Eric C. Larson et al. "SpiroSmart: Using a Microphone to Measure Lung Function on a Mobile Phone". In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*. 2012, p. 280. ISBN: 9781450312240. DOI: 10.1145/2370216.2370261. URL: http://dl.acm.org/citation.cfm?id=2370261http://dl.acm.org/citation.cfm?id=2370261{\%}5Cnhttp://dl.acm.org/citation.cfm?doid=2370216.2370261.

[132] Merlin D Larson and Matthias Behrends. "Portable Infrared Pupillometry". In: *Anesthesia & Analgesia* 120.6 (2015), pp. 1242–1253. ISSN: 0003-2999. DOI: 10.1213/ANE.0000000000000314. URL: http://journals.lww.com/anesthesia-analgesia/Abstract/2015/06000/Portable{_}Infrared{_}Pupillometry{_}{_}{_}A{_}Review.14.aspxhttp://ovidsp.ovid.com/ovidweb.cgi?T=JS{\&}PAGE=fulltext{\&}D=ovft{\&}MODE=ovid{\&}NEWS=N{\&}SEARCH=0003-2999.is+and+120.vo+and+6.ip+and+1242.pg{\&}NEWS=n{\%}5Cn.

[133] Merlin D Larson and Isobel Muhiudeen. "Pupillometric analysis of the 'absent light reflex'." In: *Archives of neurology* 52.4 (1995), pp. 369–72. ISSN: 0003-9942. DOI: 10.1097/00000542-199409001-00312. URL: http://archinte.jamanetwork.com/article.aspx?articleid=593394http://www.ncbi.nlm.nih.gov/pubmed/7710372.

[134] Everett Lawson et al. "Computational retinal imaging via binocular coupling and indirect illumination". In: *Proc. SIGGRAPH '12*. 2012, p. 51. URL: http://dl.acm.org/citation.cfm?id=2342961.

[135] Everett Lawson et al. "Computational retinal imaging via binocular coupling and indirect illumination". In: *Proc. SIGGRAPH '12*. 2012, p. 51. URL: http://dl.acm.org/citation.cfm?id=2342961.

[136] Somsak Leartveravat. "Transcutaneous bilirubin measurement in full term neonate by digital camera". In: *Medical Journal of Srisaket Surin Buriram Hospitals* 24.1 (2009), pp. 105–118.

[137] Hopin Lee et al. "Smartphone and tablet apps for concussion road warriors (team clinicians): a systematic review for practical users." In: *British journal of sports medicine* 49.8 (2014), pp. 1–2. ISSN: 1473-0480. DOI: 10.1136/bjsports-2013-092930. URL: http://www.ncbi.nlm.nih.gov/pubmed/24668048.

[138] Richard T. Lester et al. "Effects of a mobile phone short message service on antiretroviral treatment adherence in Kenya (WelTel Kenya1): A randomised trial". In: *The Lancet* 376.9755 (2010), pp. 1838–1845. ISSN: 01406736. DOI: 10.1016/S0140-6736(10)61997-6. URL: http://www.sciencedirect.com/science/article/pii/S0140673610619976.

[139] Terence S Leung et al. "Screening neonatal jaundice based on the sclera color of the eye using digital photography". In: *Biomedical optics express* 6.11 (2015), pp. 4529–4538. URL: https://www.osapublishing.org/abstract.cfm?uri=boe-6-11-4529.

[140] Stephen J Lewis and Ken W Heaton. "Stool form scale as a useful guide to intestinal transit time". In: *Scandinavian Journal of Gastroenterology* 32.9 (1997), pp. 920–924. ISSN: 00365521. DOI: 10.3109/00365529709011203. URL: http://www.tandfonline.com/doi/full/10.3109/00365529709011203.

[141] Dongheng Li, David Winfield, and Derrick J Parkhurst. "Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*. Vol. 3. IEEE, 2005, pp. 79–79. ISBN: 0-7695-2372-2. DOI: 10.1109/CVPR.2005.531. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1565386.

[142] Rainer Lienhart and Jochen Maydt. "An extended set of Haar-like features for rapid object detection". In: *Proceedings. International Conference on Image Processing*. Vol. 1. 2002, pp. 0–3. ISBN: 0-7803-7622-6. DOI: 10.1109/ICIP.2002.1038171. arXiv: 9209032v1 [hep-th]. URL: http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=1038171.

[143] Brian Y Lim and Anind K Dey. "Assessing demand for intelligibility in context-aware applications". In: *Proc. UbiComp '09*. 2009, pp. 195–204. ISBN: 9781605584317. DOI: 10.1145/1620545.1620576. URL: http://dl.acm.org/citation.cfm?id=1620576http://dl.acm.org/citation.cfm?id=1620545.1620576{\%}5Cnhttp://dl.acm.org/citation.cfm?id=1620576.

[144] Brian Y Lim and Anind K Dey. "Investigating intelligibility for uncertain context-aware applications". In: *Proc. UbiComp '11*. New York, New York, USA: ACM Press, 2011, pp. 415–424. ISBN: 9781450306300. DOI: 10.1145/2030112.2030168. URL: http://dl.acm.org/citation.cfm?doid=2030112.2030168.

[145] Joiiiv R Lindsay. "The significance of a positional nystagmus in otoneurological diagnosis". In: *The Laryngoscope* 55.10 (1945), pp. 527–551. ISSN: 0023-852X. URL: http://onlinelibrary.wiley.com/doi/10.1288/00005537-194510000-00001/full.

[146] Yike Liu. "Noise reduction by vector median filtering". In: *Geophysics* 78.3 (2013), pp. V79–V87. URL: http://library.seg.org/doi/abs/10.1190/geo2012-0232.1.

[147] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 07-12-June. 2015, pp. 3431–3440. ISBN: 9781467369640. DOI: 10.1109/CVPR.2015.7298965. arXiv: 1411.4038. URL: http://arxiv.org/abs/1411.4038.

[148] Dan L Longo et al. *Harrison's Principles of Internal Medicine*. 18th. 2006. ISBN: 007174889X. URL: http://accessmedicine.mhmedical.com/content.aspx?bookid=331{\&}sectionid=40726762.

[149] Andrew Z Luo et al. "Automatic characterization of user errors in spirometry". In: *Proc. EMBC '17*. 2017, pp. 4239–4242. ISBN: 9781509028092. DOI: 10.1109/EMBC.2017.8037792. URL: http://ieeexplore.ieee.org/abstract/document/8037792/.

[150] Alexei Maklakoff. "L'ophthalmotonometrie". In: *Archives of opthalmology* 4.159 (1885).

[151] H. B. Mann and D. R. Whitney. "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other". In: *The Annals of Mathematical Statistics* 18.1 (1947), pp. 50–60. ISSN: 0003-4851. DOI: 10.1214/aoms/1177730491. URL: http://www.jstor.org/stable/2236101http://projecteuclid.org/euclid.aoms/1177730491.

[152] Kaweh Mansouri, Tarek Shaarawy, and D. Bertrand. "Continuous intraocular pressure monitoring with a wireless ocular telemetry sensor: initial clinical experience in patients with open angle glaucoma." In: *The British journal of ophthalmology* 95.5 (2011), pp. 627–9. ISSN: 1468-2079. DOI: 10.1136/bjo.2010.192922. URL: http://www.ncbi.nlm.nih.gov/pubmed/21216796.

[153] Marco Marcon, Eliana Frigerio, and Stefano Tubaro. "Sclera segmentation for gaze estimation and iris localization in unconstrained images". In: *CompIMAGE*. 2012, pp. 25–29. URL: https://books.google.com/books?hl=en{\&}lr={\&}id=Z5ph3owFbMEC{\&}oi=fnd{\&}pg=PA25{\&}dq=frigerio,+marcon,+sclera{\&}ots=Zqk83ybbIl{\&}sig=32EWe6tVF0RznLVeEd31DJX4sCI.

[154] Alex Mariakakis et al. "A Smartphone-based System for Assessing Intraocular Pressure". In: *Proc. EMBC '16*. 2016.

[155] Alex Mariakakis et al. "BiliScreen: smartphone-based scleral jaundice monitoring for liver and pancreatic disorders". In: *Proceedings of the 2017 ACM Interactive, Mobile, Wearable, Ubiquitous Technologies*. ACM. 2017.

[156] Alex Mariakakis et al. "PupilScreen: Using Smartphones to Assess Traumatic Brain Injury". In: *Proc. IMWUT '17*. Vol. 1. 3. 2017, 81:1–81:27. DOI: 10.1145/3131896.

[157] F Martínez-Ricarte et al. "Infrared pupillometry. Basic principles and their application in the non-invasive monitoring of neurocritical patients." In: *Neurología (Barcelona, Spain)* 28.1 (2013), pp. 41–51. ISSN: 21735808. DOI: 10.1016/j.nrleng.2010.07.001. URL: http://www.sciencedirect.com/science/article/pii/S2173580813000023http://www.ncbi.nlm.nih.gov/pubmed/21163229.

[158] Jun Maruta et al. "A unified science of concussion". In: *Annals of the New York Academy of Sciences* 1208.1 (2010), pp. 58–66. ISSN: 00778923. DOI: 10.1111/j.1749-6632.2010.05695.x. URL: http://doi.wiley.com/10.1111/j.1749-6632.2010.05695.x.

[159] Jun Maruta et al. "EYE-TRAC: monitoring attention and utility for mTBI". In: *Sensing Technologies for Global Health, Military Medicine, Disaster Response, and Environmental Monitoring Ii: And Biometric Technology for Human Identification Ix* 8371 (2012), p. 11. ISSN: 0277786X. DOI: 10.1117/12.927790. URL: http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=1353923.

[160] Alex Marvez. *Players could try to beat concussion tests | FOX Sports*. 2011. URL: http://www.foxsports.com/nfl/story/NFL-players-could-try-to-beat-concussion-tests-042111.

[161] Masimo. *Pulse CO-Oximeter - Pronto*. URL: http://www.masimo.com/home/rainbow-pulse-co-oximetry/rainbow-monitors/pronto/ (visited on 02/28/2018).

[162] Paul McCrory et al. "Consensus Statement on Concussion in Sport: The 4th International Conference on Concussion in Sport Held in Zurick, November 2012". In: *British Journal of Sports Medicine* 47.2 (2013), pp. 250–258. ISSN: 1062-6050. DOI: 10.1136/bjsports-2013-092313. URL: http://www.ncbi.nlm.nih.gov/pubmed/23479479.

[163] G V McDonnell and T F G Esmonde. "A homesick student". In: *Postgraduate Medical Journal* 75.884 (1999), pp. 375–378. ISSN: 0032-5473. DOI: 10.1136/pgmj.75.884.375. URL: http://pmj.bmj.com/cgi/doi/10.1136/pgmj.75.884.375.

[164] Matthew S McGlone and Ann B Reed. "Anchoring in the interpretation of probability expressions". In: *Journal of Pragmatics* 30.6 (1998), pp. 723–733. ISSN: 03782166. DOI: 10.1016/s0378-2166(98)00011-3. URL: https://www.sciencedirect.com/science/article/pii/S0378216698000113.

[165] Michele Meeker et al. "Pupil examination: validity and clinical utility of an automated pupillometer." In: *The Journal of neuroscience nursing : journal of the American Association of Neuroscience Nurses* 37.1 (2005), pp. 34–40. ISSN: 0888-0395. URL: http://journals.lww.com/jnnonline/abstract/2005/02000/pupil{_}examination{_}{_}validity{_}and{_}clinical{_}utility.6.aspxhttp://www.ncbi.nlm.nih.gov/pubmed/15794443.

[166] Daryush D. Mehta et al. "Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform". In: *IEEE Transactions on Biomedical Engineering* 59.12 PART2 (2012), pp. 3090–3096. ISSN: 00189294. DOI: 10.1109/TBME.2012.2207896. arXiv: NIHMS150003. URL: http://ieeexplore.ieee.org/document/6257444/.

[167] Marc Mitchell et al. "Improving care-improving access: the use of electronic decision support with AIDS patients in South Africa". In: *International Journal of Healthcare Technology and Management* 10.3 (2009), pp. 156–168. ISSN: 1368-2156.

URL: http://www.inderscienceonline.com/doi/abs/10.1504/IJHTM.2009.025819.

[168] Reham Mohamed and Moustafa Youssef. "HeartSense: Ubiquitous Accurate Multi-Modal Fusion-based Heart Rate Estimation Using Smartphones". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1.3 (2017), 97:1–97:18. ISSN: 2474-9567. DOI: 10.1145/3132028. URL: https://dl.acm.org/citation.cfm?id=3132028http://doi.acm.org/10.1145/3132028.

[169] Erika A. Montanaro and Angela D. Bryan. "Comparing theory-based condom interventions: health belief model versus theory of planned behavior." In: *Health psychology : official journal of the Division of Health Psychology, American Psychological Association* 33.10 (2014), pp. 1251–1260. ISSN: 19307810. DOI: 10.1037/a0033969. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0033969.

[170] Tim Morris, Paul Blenkhorn, and Farhan Zaidi. "Blink detection for real-time eye tracking". In: *Journal of Network and Computer Applications* 25.2 (2002), pp. 129–143. ISSN: 10848045. DOI: 10.1016/S1084-8045(02)90130-X. URL: http://www.sciencedirect.com/science/article/pii/S108480450290130X.

[171] Rosemarie Scolaro Moser et al. "Neuropsychological evaluation in the diagnosis and management of sports-related concussion." In: *Archives of clinical neuropsychology : the official journal of the National Academy of Neuropsychologists* 22.8 (2007), pp. 909–16. ISSN: 0887-6177. DOI: 10.1016/j.acn.2007.09.004. URL: http://www.ncbi.nlm.nih.gov/pubmed/17988831.

[172] Simon Moss. *Fit indices for structural equation modeling*. 2015. URL: https://www.sicotests.com/psyarticle.asp?id=277 (visited on 03/26/2019).

[173] Amir Muaremi et al. "Assessing bipolar episodes using speech cues derived from phone calls". In: *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*. Vol. 100. 2014, pp. 103–114. ISBN:

9783319115634. DOI: 10.1007/978-3-319-11564-1_11. URL: http://link.springer.com/chapter/10.1007/978-3-319-11564-1{_}11.

[174] Onur Mudanyali et al. "Integrated rapid-diagnostic-test reader platform on a cellphone". In: *Lab on a Chip* 12.15 (2012), p. 2678. ISSN: 1473-0197. DOI: 10.1039/c2lc40235a. URL: http://xlink.rsc.org/?DOI=c2lc40235a.

[175] Franz Nachbar et al. "The ABCD rule of dermatoscopy". In: *Journal of the American Academy of Dermatology* 30.4 (1994), pp. 551–559. ISSN: 01909622. DOI: 10.1016/S0190-9622(94)70061-3. URL: http://linkinghub.elsevier.com/retrieve/pii/S0190962294700613http://www.sciencedirect.com/science/article/pii/S0190962294700613.

[176] Rachel Nall and Michael Charles. *Is it strep throat? Pictures and symptoms*. 2017. URL: https://www.medicalnewstoday.com/articles/312433.php (visited on 04/08/2019).

[177] John T Nathanson, James G Connolly, and Frank Yuk. "Concussion Incidence in Professional Football Position-Specific Analysis With Use of a Novel Metric". In: *Orthopaedic Journal of Sports Medicine* 4.1 (2016), p. 2325967115622621. DOI: 10.1177/2325967115622621. URL: http://ojs.sagepub.com/content/4/1/2325967115622621.short.

[178] National Center for Health Statistics (US) and National Center for Health Services Research. *Health, United States*. US Department of Health, Education, and Welfare, Public Health Service, Health Resources Administration, National Center for Health Statistics, 2010.

[179] Fidele Ngabo et al. "Designing and Implementing an Innovative SMS-based alert system (RapidSMS-MCH) to monitor pregnancy and reduce maternal and child deaths in Rwanda". In: *Pan African Medical Journal* 13.31 (2012), pp. 1–16. URL: http://www.panafrican-med-journal.com/content/article/13/31/full/.

[180] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. "Learning Deconvolution Network for Semantic Segmentation". In: 1 (2015). ISSN: 15505499. DOI: 10.1109/ICCV.2015.178. arXiv: 1505.04366. URL: http://arxiv.org/abs/1505.04366.

[181] Nokia. *Nokia BPM+ | Wireless Blood Pressure Monitor*. 2017. URL: https://health.nokia.com/us/en/blood-pressure-monitor (visited on 02/27/2018).

[182] Dai-Wai M Olson et al. "Interrater Reliability of Pupillary Assessments". In: *Neurocritical Care* 24.2 (2016), pp. 251–257. ISSN: 1541-6933. DOI: 10.1007/s12028-015-0182-1. URL: http://www.ncbi.nlm.nih.gov/pubmed/26381281http://link.springer.com/10.1007/s12028-015-0182-1.

[183] Daniel M. Oppenheimer, Tom Meyvis, and Nicolas Davidenko. "Instructional manipulation checks: Detecting satisficing to increase statistical power". In: *Journal of Experimental Social Psychology* 45.4 (2009), pp. 867–872. ISSN: 00221031. DOI: 10.1016/j.jesp.2009.03.009. URL: http://www.sciencedirect.com/science/article/pii/S0022103109000766.

[184] Calvin KL Or and Ben-Tzion Karsh. "A systematic review of patient acceptance of consumer health information technology". In: *Journal of the American Medical Informatics Association* 16.4 (2009), pp. 550–560. URL: http://www.sciencedirect.com/science/article/pii/S1067502709000838http://jamia.oxfordjournals.org/content/16/4/550.short.

[185] Nobuyuki Otsu. "A threshold selection method from gray-level histograms". In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (1979), pp. 62–66. ISSN: 0018-9472. DOI: 10.1109/TSMC.1979.4310076. URL: http://ieeexplore.ieee.org/document/4310076/http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4310076.

[186] Christopher G Owen et al. "Diabetes and the tortuosity of vessels of the bulbar conjunctiva". In: *Ophthalmology* 115.6 (2008), e27–e32. URL: http://www.sciencedirect.com/science/article/pii/S0161642008001656.

[187] Christopher G Owen et al. "Vascular response of the bulbar conjunctiva to diabetes and elevated blood pressure". In: *Ophthalmology* 112.10 (2005), pp. 1801–1808. URL: http://www.sciencedirect.com/science/article/pii/S016164200500713X.

[188] Vitor F Pamplona, Manuel M Oliveira, and Gladimir V. G. Baranoski. "Photorealistic models for pupil light reflex and iridal pattern deformation". In: *ACM Transactions on Graphics* 28.4 (2009), pp. 1–12. ISSN: 07300301. DOI: 10.1145/1559755.1559763. arXiv: 1006.4903.

[189] Vitor F Pamplona et al. "Catra: cataract probe with a lightfield display and a snap-on eyepiece for mobile phones". In: *Proc. SIGGRAPH '11*. 2011, pp. 7–11. URL: http://doi.acm.org/10.1145/1964921.1964942.

[190] Vitor F Pamplona et al. "NETRA: interactive display for estimating refractive errors and focal range". In: *ACM transactions on graphics (TOG)* 29.4 (2010), p. 77.

[191] W C Panek et al. "Intraocular pressure measurement with the Tono-Pen through soft contact lenses". In: *American Journal of Ophthalmology* 109.1 (1990), pp. 62–65. URL: http://www.sciencedirect.com/science/article/pii/S0002939414755801http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve{\&}db=PubMed{\&}dopt=Citation{\&}list{_}uids=2297033.

[192] Gede Pardianto. "Intraocular pressure measure on normal eyes". In: *Mimbar Ilmiah Ofalmologi* (2005).

[193] Danny Pascale. *RGB coordinates of the Macbeth ColorChecker*. Tech. rep. 2006, pp. 1–16. URL: http://kronometric.org/phot/lighting/MacbethColorCheckerTables.pdf.

[194] Trevor Perrier et al. "Engaging Pregnant Women in Kenya with a Hybrid Computer-Human SMS Communication System". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. New York, New York, USA: ACM Press, 2015, pp. 1429–1438. ISBN: 978-1-4503-3145-6. DOI: 10.1145/2702123.2702124. URL: http://dl.acm.org/citation.cfm?doid=2702123.2702124http://doi.acm.org/10.1145/2702123.2702124.

[195] Stephen M Pizer et al. "Adaptive histogram equalization and its variations". In: *Computer Vision, Graphics, and Image Processing* 39.3 (1987), pp. 355–368. ISSN: 0734189X. DOI: 10.1016/S0734-189X(87)80186-X. URL: http://www.sciencedirect.com/science/article/pii/S0734189X8780186Xhttp://linkinghub.elsevier.com/retrieve/pii/S0734189X8780186X.

[196] Adolph Posner. "Modified conversion tables for the Maklakov tonometer". In: *Eye, ear, nose & throat monthly* 41 (1962), pp. 638–644.

[197] Adolph Posner. "The Applanometer, a Modified Maklakov Applanation Tonometer". In: *Eye, ear, nose & throat monthly* 44 (1965), pp. 77–80. URL: http://europepmc.org/abstract/med/14273037.

[198] Adolph Posner and Richard Inglima. "The tonomat applanation tonometer: a comparison with the Goldmann applanation tonometer and the applanometer". In: *Eye, ear, nose & throat monthly* 48.3 (1969), pp. 189–194. ISSN: 00145491. URL: http://europepmc.org/abstract/med/5777460.

[199] James O Prochaska and Carlo C DiClemente. *The transtheoretical approach: Crossing traditional boundaries of therapy*. Dow Jones-Irwin, 1984.

[200] Hugo Proenca et al. "The UBIRIS.v2: A Database of Visible Wavelength Iris Images Captured On-the-Move and At-a-Distance". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.8 (2010), pp. 1529–1535. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2009.66. URL: http://ieeexplore.ieee.org/document/4815254/.

[201] Harry A Quigley and Aimee T Broman. "The number of people with glaucoma worldwide in 2010 and 2020". In: *British journal of ophthalmology* 90.3 (2006), pp. 262–267. URL: http://bjo.bmj.com/content/90/3/262.short.

[202] I. A. Qureshi et al. "Variations in ocular pressure during menstrual cycle." In: *The Journal of the Pakistan Medical Association* 48.2 (1998), pp. 37–40. ISSN: 00309982. URL: http://europepmc.org/abstract/med/9610091.

[203] Sohail Rafiqi et al. "PupilWare: towards pervasive cognitive load measurement using commodity devices". In: *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments - PETRA '15*. New York, New York, USA: ACM Press, 2015, pp. 1–8. ISBN: 9781450334525. DOI: 10.1145/2769493.2769506. URL: http://dl.acm.org/citation.cfm?doid=2769493.2769506.

[204] Tauhidur Rahman et al. "BodyBeat: Eavesdropping on our Body Using a Wearable Microphone." In: *GetMobile*. Vol. 19. 1. 2015, pp. 14–17. ISBN: 9781450327930. DOI: 10.1145/2786984.2786989. URL: http://dl.acm.org/citation.cfm?id=2786989http://doi.acm.org/10.1145/2786984.2786989{\%}5Cnpapers2://publication/doi/10.1145/2786984.2786989.

[205] R. A. Ramlee and S. Ranjit. "Using iris recognition algorithm, detecting cholesterol presence". In: *Proceedings - 2009 International Conference on Information Management and Engineering, ICIME 2009*. IEEE, 2009, pp. 714–717. ISBN: 9780769535951. DOI: 10.1109/ICIME.2009.61. URL: http://ieeexplore.ieee.org/document/5077127/.

[206] David B Rein et al. "The economic burden of major adult visual disorders in the United States". In: *Archives of opthalmology* 124.12 (2006), pp. 1754–1760. URL: http://archsurg.jamanetwork.com/data/Journals/OPHTH/9977/ese60005{_}1754{_}1760.pdf.

[207] Corwin N Rhyan. *Travel and Wait Times are Longest for Health Care Services and Result in an Annual Opportunity Cost of $89 Billion*. Tech. rep. 2019, pp. 1–6. URL: https://altarum.org/sites/default/files/uploaded-publication-files/Altarum{_}Travel-and-Wait-Times-for-Health-Care-Services{_}Feb-22.pdf.

[208] Yvonne Rogers et al. "Why It's Worth the Hassle - The Value of In-Situ Studies When Designing Ubicomp". In: *UbiComp'07*. 2007, pp. 336–353. ISBN: 9783540748526. DOI: 10.1007/978-3-540-74853-3. URL: http://link.springer.com/content/pdf/10.1007/978-3-540-74853-3.pdf{\#}page=353http://www.springerlink.com/index/10.1007/978-3-540-74853-3.

[209] John Rooksby et al. "Personal tracking as lived informatics". In: *Proc. CHI 2014*. 2014, pp. 1163–1172. DOI: 10.1145/2556288.2557039. URL: https://dl.acm.org/citation.cfm?id=2557039.

[210] Irwin M. Rosenstock, Victor J. Strecher, and Marshall H. Becker. "Social Learning Theory and the Health Belief Model". In: *Health Education Quarterly* 15.2 (1988), pp. 175–183. ISSN: 0195-8402. DOI: 10.1177/109019818801500203. URL: http://journals.sagepub.com/doi/10.1177/109019818801500203.

[211] Yves Rosseel. "lavaan: An R Package for Structural Equation Modeling". In: *Journal of Statistical Software* 48.2 (2012), pp. 1–36.

[212] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "Grabcut: Interactive foreground extraction using iterated graph cuts". In: *ACM Transactions on Graphics*

(TOG) '04 23.3 (2004), pp. 309–314. URL: http://dl.acm.org/citation.cfm?id=1015720.

[213] Alessandro Rubini, Andrea Parmagnani, and Michela Bondì. "Daily variations in lung volume measurements in young healthy adults". In: *Biological Rhythm Research* 42.3 (2011), pp. 261–265. ISSN: 0929-1016. DOI: 10.1080/09291016.2010.505456. URL: http://www.tandfonline.com/doi/abs/10.1080/09291016.2010.505456http://dx.doi.org/10.1080/09291016.2010.505456.

[214] Mario A Ruiz, Sammy Saab, and Leland S Rickman. "The clinical detection of scleral icterus: observations of multiple examiners." In: *Military medicine* 162.8 (1997), pp. 560–563. URL: http://europepmc.org/abstract/med/9271910.

[215] David P Ryan, Theodore S Hong, and Nabeel Bardeesy. "Pancreatic Adenocarcinoma". In: *New England Journal of Medicine* 371.11 (2014), pp. 1039–1049. ISSN: 0028-4793. DOI: 10.1056/NEJMra1404198. URL: http://www.nejm.org/doi/10.1056/NEJMra1404198.

[216] Mandy Ryan. "Using conjoint analysis to take account of patient preferences and go beyond health outcomes: An application to in vitro fertilization". In: *Social Science and Medicine* 48.4 (1999), pp. 535–546. ISSN: 02779536. DOI: 10.1016/S0277-9536(98)00374-8. URL: https://www.sciencedirect.com/science/article/pii/S0277953698003748?via{\%}3Dihub.

[217] Suchi Saria, Daphne Koller, and Anna Penn. "Learning individual and population level traits from clinical temporal data". In: *Proc. Neural Information Processing Systems (NIPS), Predictive Models in Personalized Medicine Workshop*. 2010, pp. 1–9. DOI: 10.1016/j.jhevol.2007.05.008. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.232.390{\&}rep=rep1{\&}type=pdf.

[218] H Schiøtz. "Ein neues TonometerArch f Augenh". In: *Arch Augenh* (). URL: https://scholar.google.com/scholar?q=Ein+neues+TonometerArch+f+Augenh{\%}2C+schiotz{\&}btnG={\&}hl=en{\&}as{_}sdt=0{\%}2C48.

[219] Gunnar Schmidtmann et al. "Intraocular pressure fluctuations in professional brass and woodwind musicians during common playing conditions". In: *Graefe's Archive for Clinical and Experimental Ophthalmology* 249.6 (2011), pp. 895–901. ISSN: 0721-832X. DOI: 10.1007/s00417-010-1600-x. URL: http://link.springer.com/10.1007/s00417-010-1600-x.

[220] Hema Seshan and M Shwetha. "Gingival inflammation assessment: Image analysis". In: *Journal of Indian Society of Periodontology* 16.2 (2012), p. 231. ISSN: 0972-124X. DOI: 10.4103/0972-124x.99267. URL: http://www.ncbi.nlm.nih.gov/pubmed/23055590http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3459504.

[221] Gaurav Sharma, Wencheng Wu, and Edul N. Dalal. "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations". In: *Color Research and Application* 30.1 (2005), pp. 21–30. ISSN: 03612317. DOI: 10.1002/col.20070. URL: http://doi.wiley.com/10.1002/col.20070.

[222] Li Shen, Joshua A Hagen, and Ian Papautsky. "Point-of-care colorimetric detection with a smartphone". In: *Lab on a Chip* 12.21 (2012), pp. 4240–4243. DOI: 10.1039/c2lc40741h. URL: http://pubs.rsc.org/is/content/articlehtml/2012/lc/c2lc40741h.

[223] Vimal K Shrivastava et al. "Computer-aided diagnosis of psoriasis skin images with HOS, texture and color features: A first comparative study of its kind". In: *Computer Methods and Programs in Biomedicine* 126 (2016), pp. 98–109. ISSN:

0169-2607. DOI: 10.1016/J.CMPB.2015.11.013. URL: https://www.sciencedirect.com/science/article/pii/S0169260715300699.

[224] DO Sillence, Alison Senn, and DM Danks. "Genetic heterogeneity in osteogenesis imperfecta." In: *Journal of medical genetics* 16.2 (1979), pp. 101–116. URL: http://jmg.bmj.com/content/16/2/101.short.

[225] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: (2014). arXiv: 1409.1556. URL: http://arxiv.org/abs/1409.1556.

[226] J Lawton Smith and John O Susac. "Unilateral arcus senilis: sign of occlusive disease of the carotid artery". In: *JAMA* 226.6 (1973), p. 676. URL: http://jama.jamanetwork.com/article.aspx?articleid=351421.

[227] Sue Stevens, Clare Gilbert, and Nick Astbury Astbury. "How to measure intraocular pressure: applanation tonometry". In: *Community Eye Health* 20.64 (2007), p. 74. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2206330/.

[228] Xiao Sun et al. "SymDetector: Detecting Sound-Related Respiratory Symptoms Using Smartphones". In: *Proc. UbiComp '15*. 2015, pp. 97–108. ISBN: 9781450335744. DOI: 10.1145/2750858.2805826. URL: https://dl.acm.org/citation.cfm?id=2805826http://dl.acm.org/citation.cfm?id=2805826.

[229] Lech Swirski, Andreas Bulling, and Neil Dodgson. "Robust real-time pupil tracking in highly off-axis images". In: *Etra* (2012), pp. 1–4. DOI: 10.1145/2168556.2168585. URL: http://dl.acm.org/citation.cfm?doid=2168556.2168585http://www.cl.cam.ac.uk/research/rainbow/projects/pupiltracking/files/Swirski,Bulling,Dodgson-2012-Robustreal-timepupiltrackinginhighlyoff-axisimages.pdf.

[230] David Taylor et al. "A Review of the use of the Health Belief Model (HBM), the Theory of Reasoned Action (TRA), the Theory of Planned Behaviour (TPB) and the Trans-Theoretical". In: *London, UK: National Institute for Health and Clinical Excellence* June (2006), pp. 1–215. URL: http://www.academia.edu/download/33424122/Behaviour{_}Change-Taylor{_}et{_}al-models{_}review{_}tables{_}appendices.pdfhttps://www.nice.org.uk/guidance/ph6/documents/behaviour-change-taylor-et-al-models-review2.

[231] James A Taylor et al. "Use of a Smartphone App To Assess Neonatal Jaundice". In: *Pediatrics* 140.3 (2017), e20170312. ISSN: 0031-4005. DOI: 10.1542/peds.2017-0312. URL: http://www.ncbi.nlm.nih.gov/pubmed/28842403http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5574723http://pediatrics.aappublications.org/lookup/doi/10.1542/peds.2017-0312.

[232] William R Taylor et al. "Quantitative pupillometry, a new technology: normative data and preliminary observations in patients with acute head injury. Technical note." In: *Journal of neurosurgery* 98.1 (2003), pp. 205–213. ISSN: 0022-3085. DOI: 10.3171/jns.2003.98.1.0205. URL: http://thejns.org/doi/abs/10.3171/jns.2003.98.1.0205.

[233] N Theofilopoulos et al. "Effects of reboxetine and desipramine pupillary light reflex". In: *British journal of clinical pharmacology* 39.3 (1995), pp. 251–255. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2125.1995.tb04444.x/abstract.

[234] Preethi Thiagarajan and Kenneth J Ciuffreda. "Pupillary responses to light in chronic non-blast-induced mTBI". In: *Brain Injury* 29.12 (2015), pp. 1420–1425. DOI: 10.3109/02699052.2015.1045029. URL: http://www.ncbi.nlm.nih.gov/pubmed/26182230.

[235] Fabian Timm and Erhardt Barth. "Accurate Eye Centre Localisation by Means of Gradients". In: *VISAPP*. 2011, pp. 125–130. URL: http://cjee.lakeheadu.ca/public/journals/22/TiBa11b.pdf.

[236] Lyndal J Trevena et al. "Presenting quantitative information about decision outcomes: a risk communication primer for patient decision aid developers". In: *BMC Medical Informatics and Decision Making* 13.Suppl 2 (2013), S7. ISSN: 1472-6947. DOI: 10.1186/1472-6947-13-S2-S7. URL: http://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-13-S2-S7.

[237] James Q Truong and Kenneth J Ciuffreda. "Comparison of pupillary dynamics to light in the mild traumatic brain injury (mTBI) and normal populations". In: *Brain Injury* 30.11 (2016), pp. 1378–1389. ISSN: 0269-9052. DOI: 10.1080/02699052.2016.1195922. URL: http://www.ncbi.nlm.nih.gov/pubmed/27541745https://www.tandfonline.com/doi/full/10.1080/02699052.2016.1195922.

[238] James Q Truong and Kenneth J Ciuffreda. "Objective Pupillary Correlates of Photosensitivity in the Normal and Mild Traumatic Brain Injury Populations". In: *Military Medicine* 181.10 (2016), pp. 1382–1390. ISSN: 0026-4075. DOI: 10.7205/MILMED-D-15-00587. URL: http://www.ncbi.nlm.nih.gov/pubmed/27753579http://publications.amsus.org/doi/10.7205/MILMED-D-15-00587.

[239] James Q Truong and Kenneth J Ciuffreda. "Quantifying pupillary asymmetry through objective binocular pupillometry in the normal and mild traumatic brain injury (mTBI) populations". In: *Brain Injury* 30.11 (2016), pp. 1372–1377. ISSN: 0269-9052. DOI: 10.1080/02699052.2016.1192220. URL: http://www.ncbi.nlm.nih.gov/pubmed/27712127https://www.tandfonline.com/doi/full/10.1080/02699052.2016.1192220.

[240] James Q Truong, Nabin R Joshi, and Kenneth J Ciuffreda. "Influence of refractive error on pupillary dynamics in the normal and mild traumatic brain injury (mTBI) populations". In: *Journal of Optometry* (2017). ISSN: 18884296. DOI: 10.1016/j.optom.2016.12.005. URL: http://www.ncbi.nlm.nih.gov/pubmed/28262507http://linkinghub.elsevier.com/retrieve/pii/S1888429617300031.

[241] Khai N. Truong, Gillian R. Hayes, and Gregory Abowd. "Storyboarding: an empirical determination of best practices and effective guidelines". In: *Proceedings of the 6th ACM conference on Designing Interactive systems - DIS '06*. 2006, pp. 12–21. ISBN: 1595933670. DOI: 10.1145/1142405.1142410. URL: https://dl.acm.org/citation.cfm?id=1142410http://portal.acm.org/citation.cfm?doid=1142405.1142410.

[242] Markos G. Tsipouras et al. "On automated assessment of Levodopa-induced dyskinesia in Parkinson's disease". In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. 2011, pp. 2679–2682. ISBN: 9781424441211. DOI: 10.1109/IEMBS.2011.6090736. URL: http://ieeexplore.ieee.org/abstract/document/6090736/.

[243] Zenith USA. *Smartphone penetration to reach 66% in 2018*. 2018. URL: https://www.zenithusa.com/smartphone-penetration-reach-66-2018/ (visited on 04/29/2019).

[244] Gonzalo M. Vazquez-Prokopec et al. "Using GPS Technology to Quantify Human Mobility, Dynamic Contacts and Infectious Disease Dynamics in a Resource-Poor Urban Environment". In: *PLoS ONE* 8.4 (2013). Ed. by Vittoria Colizza, e58802. ISSN: 19326203. DOI: 10.1371/journal.pone.0058802. URL: http://dx.plos.org/10.1371/journal.pone.0058802.

[245] Viswanath Venkatesh and Fred D. Davis. "Theoretical extension of the technology acceptance model: four longitudinal field studies." In: *Management science* 46.2

(2000), pp. 186–204. URL: http://pubsonline.informs.org/doi/abs/10.1287/mnsc.46.2.186.11926.

[246] Audrey Vincent et al. "Pancreatic cancer". In: *The Lancet* 378.9791 (2011), pp. 607–620. ISSN: 01406736. DOI: 10.1016/S0140-6736(10)62307-0. URL: http://www.ncbi.nlm.nih.gov/pubmed/21620466http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3062508http://linkinghub.elsevier.com/retrieve/pii/S0140673610623070.

[247] Paul Viola and Michael J Jones. "Rapid object detection using a boosted cascade of simple features". In: *Computer Vision and Pattern Recognition (CVPR)*. Vol. 1. 2001, pp. 511–518. ISBN: 0-7695-1272-0. DOI: 10.1109/CVPR.2001.990517. arXiv: arXiv:1011.1669v3. URL: http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=990517.

[248] Tarun Wadhawan et al. "Implementation of the 7-point checklist for melanoma detection on smart handheld devices". In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. IEEE, 2011, pp. 3180–3183. ISBN: 9781424441211. DOI: 10.1109/IEMBS.2011.6090866. URL: http://ieeexplore.ieee.org/document/6090866/.

[249] Abram L. Wagner et al. "Perceptions of measles, pneumonia, and meningitis vaccines among caregivers in Shanghai, China, and the health belief model: A cross-sectional study". In: *BMC Pediatrics* 17.1 (2017), p. 143. ISSN: 14712431. DOI: 10.1186/s12887-017-0900-2. URL: http://bmcpediatr.biomedcentral.com/articles/10.1186/s12887-017-0900-2.

[250] H Kenneth Walker, W Dallas Hall, and J Willis Hurst. *Clinical Methods*. Butterworths, 1990, pp. 1–76. ISBN: 040990077X. URL: http://www.ncbi.nlm.nih.gov/pubmed/21250045.

[251] Edward J Wang et al. "Noninvasive hemoglobin measurement using unmodified smartphone camera and white flash". In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. IEEE, 2017, pp. 2333–2336. ISBN: 9781509028092. DOI: 10.1109/EMBC.2017.8037323. URL: http://ieeexplore.ieee.org/document/8037323/.

[252] Edward Jay Wang et al. "HemaApp: Noninvasive Blood Screening of Hemoglobin Using Smartphone Cameras". In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '16*. 2016, pp. 593–604. ISBN: 9781450344616. DOI: 10.1145/2971648.2971653. URL: http://dl.acm.org/citation.cfm?id=2971653http://dl.acm.org/citation.cfm?doid=2971648.2971653.

[253] Edward Jay Wang et al. "Seismo: Blood Pressure Monitoring using Built-in Smartphone Accelerometer and Camera". In: *CHI '18*. 2018, p. 425. ISBN: 9781450356206. DOI: 10.1145/3173574.3173999. URL: https://dl.acm.org/citation.cfm?id=3173999.

[254] Amy Wesolowski et al. "Quantifying the Impact of Human Mobility on Malaria". In: *Science* 338.6104 (2012), pp. 7–9.

[255] Ryen W. White and Eric Horvitz. "Cyberchondria: studies of the escalation of medical concerns in web search". In: *ACM Transactions on Information Systems* 27 (2009), pp. 1–37. ISSN: 10468188. DOI: 10.1145/1629096.1629101. URL: https://dl.acm.org/citation.cfm?id=1629101http://dl.acm.org/citation.cfm?id=1629096.1629101.

[256] Russell Wiesner et al. "Model for end-stage liver disease (MELD) and allocation of donor livers". In: *Gastroenterology* 124.1 (2003), pp. 91–96. ISSN: 00165085. DOI: 10.1053/gast.2003.50016. URL: http://www.ncbi.nlm.nih.gov/pubmed/12512033http://linkinghub.elsevier.com/retrieve/pii/S0016508503500221.

[257] Mark E Williams. *Geriatric physical diagnosis: a guide to observation and assessment*. McFarland Inc, 2009, p. 96. ISBN: 9780786451609. URL: https://books.google.com/books?id=FX7{\%}5C{_}PesP6eMC.

[258] Stephen Wolf. "Color correction matrix for digital still and video imaging systems". In: (2003), pp. 1–40. URL: http://www.its.bldrdoc.gov/publications/04-406.aspx{\%}5Cnhttp://scholar.google.com/scholar?hl=en{\&}btnG=Search{\&}q=intitle:Color+Correction+Matrix+for+Digital+Still+and+Video+Imaging+Systems{\#}0.

[259] Roger CW Wolfs et al. "Distribution of central corneal thickness and its association with intraocular pressure: The Rotterdam Study". In: *American journal of ophthalmology* 123.6 (1997), pp. 767–772. DOI: 10.1016/S0002-9394(14)71125-0. URL: http://dx.doi.org/10.1016/S0002-9394(14)71125-0.

[260] E Wood et al. *Rendering of Eyes for Eye-Shape Registration and Gaze Estimation*. 2015. DOI: 10.1109/ICCV.2015.428. arXiv: 1505.05916. URL: http://www.cv-foundation.org/openaccess/content{_}iccv{_}2015/html/Wood{_}Rendering{_}of{_}Eyes{_}ICCV{_}2015{_}paper.html.

[261] Erroll Wood and Andreas Bulling. "Eyetab: Model-based gaze estimation on unmodified tablet computers". In: *Etra* (2014), pp. 3–6. DOI: 10.1145/2578153.2578185. URL: http://dl.acm.org/citation.cfm?doid=2578153.2578185http://dl.acm.org/citation.cfm?id=2578185.

[262] D.F. Woodhouse. "Five g ME applanation tonometry Pt to P0 conversion nomogram and table". In: *Experimental Eye Research* 15.4 (1973), pp. 509–512. ISSN: 00144835. DOI: 10.1016/0014-4835(73)90143-7. URL: http://www.sciencedirect.com/science/article/pii/0014483573901437.

[263] World Health Organization. *WHO | Health workforce*. 2017. URL: http://gamapserver.who.int/gho/interactive{_}charts/health{_}workforce/PhysiciansDensity{_}Total/atlas.html (visited on 02/23/2017).

[264] Sahar F Zafar and Jose I Suarez. *Automated pupillometer for monitoring the critically ill patient: A critical appraisal*. 2014. DOI: 10.1016/j.jcrc.2014.01.012. URL: http://www.sciencedirect.com/science/article/pii/S0883944114000409.

[265] Weidan Zhao et al. "Inter-device reliability of the NPi-100 pupillometer". In: *Journal of Clinical Neuroscience* 33 (2016), pp. 79–82. ISSN: 15322653. DOI: 10.1016/j.jocn.2016.01.039.

[266] Zhi Zhou et al. "A comprehensive approach for sclera image quality measure". In: *International Journal of Biometrics* 5.2 (2013), pp. 181–198. ISSN: 17558301. DOI: 10.1504/IJBM.2013.052972. URL: http://www.inderscience.com/link.php?id=52972http://dx.doi.org/10.1504/IJBM.2013.052972.

[267] Zhi Zhou et al. "A New Human Identification Method: Sclera Recognition". In: *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 42.3 (2012), pp. 571–583. ISSN: 1083-4427. DOI: 10.1109/TSMCA.2011.2170416. URL: http://ieeexplore.ieee.org/document/6065764/.