Using Consumer Devices to Monitor Acute Medical Conditions for Infants

Lilian de Greef

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

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Shwetak N. Patel, Chair
Richard Anderson
James Fogarty

Program Authorized to Offer Degree:

Paul G. Allen School of Computer Science and Engineering

UNIVERSITY OF WASHINGTON

Abstract

Using Consumer Devices to Monitor Acute Medical Conditions for Infants

by Lilian de Greef

Chair of the Supervisory Committee:

Shwetak N. Patel

Paul G. Allen School of Computer Science and Engineering

Acute medical conditions need immediate attention, but early detection can require professional experience and specialized equipment that are unavailable at home. Consequently, babies with such conditions risk suffering damage from late interventions. We can leverage the world's increasingly ubiquitous devices to improve the accessibility of health care outside the hospital through machine learning and integrating a human-centered approach at every step of the process. This dissertation examines this approach through three projects: a smartphone-based system to screen newborns for dangerous levels of jaundice, an exploration on how machine learning can help an existing system better monitor infants with single ventricle heart disease, and a reflection on the methods and insights from working in this space to inform future work.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	**A**rtificial **I**ntelligence
BCB	**B**ili**C**am-estimated **B**ilirubin
CCT	**C**orrelated **C**olor **Temperature**
CDC	**C**enters for **D**isease **C**ontrol and **P**revention
CHAMP	**C**ardiac **H**igh **A**cuity **M**onitoring **P**rogram
CHW	**C**ommunity **H**ealth **W**orker
CMH	Kansas City **C**hildren's **M**ercy **H**ospital
ECG	**E**lectro**c**ardio**g**ram
EMR	**E**lectronic **M**edical **R**ecords
EXIF	**E**xchangeable **I**mage **F**ile **F**ormat
FDA	U.S. **F**ood and **D**rug **A**dministration
HLHS	**H**ypoplastic **L**eft **H**eart **S**yndrome
IRB	**I**nstitutional **R**eview **B**oard
GPS	**G**lobal **P**ositioning **S**ystem
ICU	**I**ntensive **C**are **U**nit
KMC	**K**angaroo **M**other **C**are
mHealth	**m**obile **Health**
NPV	**N**egative **P**redictive **V**alue
PCA	**P**rincipal **C**omponent **A**nalysis
PCP	(recommendation to visit a) **P**rincipal **C**are **P**hysician
PPV	**P**ositive **P**redictive **Value**
PR curve	**P**recision-**R**ecall Curve
PEWS	**P**ediatric **E**arly **W**arning **S**igns
ROC curve	**R**eceiver **O**perating **C**haracteristic curve
SaMD	**S**oftware **a**s a **M**edical **D**evice
SIDS	**S**udden **I**nfant **D**eath **S**yndrome
TcB	**T**rans**c**utaneous **B**ilirubinometer
TSB	**T**otal **S**erum **B**ilirubin
UW	**U**niversity of **W**ashington

Chapter 1

Introduction

1.1 Overview and Motivation

Early detection of acute medical conditions in infants can require professional experience and specialized equipment. Because neither are available at home, early detection is challenging in these settings, putting infants at risk of suffering damage from late interventions. Early detection in home environments can be enabled through integrating machine learning with commodity hardware into in-home screening tools. Families and equipped professionals can use these tools to quickly help determine whether and how urgently they need to connect about an infant's medical condition.

There is a wide spectrum of rarity and complexity of such medical conditions. On one end, an example of a common condition is neonatal jaundice, or the yellowing of the skin in newborns. As many as 84% of all newborns develop visible jaundice during their first week of life [15], and 1 in 14 newborns in the U.S. receive treatment to prevent the life-long debilitating conditions or fatalities caused by severe jaundice, known as kernicterus. In most cases, treatment is relatively straight-forward: phototherapy, or bathing the affected newborns in a specific frequency and intensity of blue light, is usually all they need. Screening for this condition, however, either involves expensive, specialized equipment and the medical training to operate it or relies on human eyes to make a visual assessment – a method known to be unreliable, especially from those without professional experience [103].

An example on the other end of this spectrum is single ventricle heart disease, a high mortality congenital heart defect that affects up to an estimated 8 per 10,000 live births [91] and requires a complex treatment of three specialized heart surgeries spaced over multiple years. The advent of a remote monitoring system called CHAMP (Cardiac High Acuity Monitoring Program), in which parents digitally share data and videos of their child with a medical care team, has dramatically decreased mortality rates [117]. However, most patients still experience costly emergency room

trips and unplanned hospital readmissions – acute conditions people may be able to mitigate with the help of earlier detection. In the most recent study, more than half of the initial unplanned readmissions were found by a medical team reviewing CHAMP data instead of the caregiver, highlighting the importance of professional experience for screening these patients [19]. While earlier detection could help mitigate these risky, expensive, and distressing interventions, they currently depend on the timing between data entry and when professionals happen to review them. Hence, there is a need for expediting the medical team's attention when acute conditions arise – especially given how these acute conditions can dramatically worsen in a matter of hours.

I believe a new generation of tools can help improve chances of early detection for the acute medical conditions that require professional experience to detect early, which for the aforementioned reasons is challenging in outpatient settings. My goal is to chart a road-map for building tools that offer always-available instant evaluations to inform decisions on whether to connect families with medical professionals, using equipment that is feasible for home and other outpatient settings. I guide my research in this space through the lens of concrete medical examples, leading me to have two further goals: developing such a tool for early detection of neonatal jaundice, and drafting a blueprint for what such tools can be for identifying interstage complications for infants with single ventricle heart disease.

Our approach combines machine learning and commodity hardware to enable this task. Machine learning models serve as instantly-available sources of experience which, while they cannot replace the invaluable expertise of medical professionals, can help supplement the limited experience of outpatient caregivers. Carefully chosen commodity hardware serve as sensing platforms that are feasible to have and use at home or other outpatient settings, to measure relevant physiological signals. Using commodity devices that are already ubiquitous in these outpatient settings can further aid the dissemination of such tools. I also inform many aspects of my approach with principles from human-computer interaction (HCI), because a tool's usability is also a key component of its effectiveness.

1.2 Overview of Current Approaches

Early detection of acute infant medical conditions like kernicterus in newborns or decompensation in CHAMP patients requires immediately available professional expertise and the right equipment to measure the relevant physiological signals – a combination that is not feasible to reliably have at home. A number of different workarounds operate by addressing a subset of these components.

One approach is to schedule additional hospital visits, where the expertise and equipment are both available. In the case of neonatal jaundice for example, some hospitals implement the procedure of scheduling families to

bring newborns back for accurate jaundice assessment with blood tests or other specialized equipment. Similarly, interstage patients with single ventricle heart disease have scheduled appointments at their hospital for check-ups and screenings from specialized doctors with hospital equipment like echocardiograms and CT scanners. The difficulty of this approach is timing; acute conditions can arise outside these scheduled visits and it is not feasible for families and doctors to do these visits multiple times per day, every day, for every patient to cover all their bases. This infeasibility is especially noteworthy for families who live far away from the hospital. Hence, while hospital visits offer the means to make accurate diagnoses, they do not offer the availability or immediacy for all cases of early detection.

Similar to hospital visits, another approach is to schedule visits in outpatient settings that are more accessible to families. This approach is protocol in Seattle for neonatal jaundice; all newborns are scheduled doctor's visits – either at hospitals or outpatient physician offices – for screening at the age that jaundice usually peaks. A different version of this approach, in which visiting nurses visit homes to check on newborns, is also common in parts of the world like the U.K. While the convenience of these approaches for families can improve accessibility to professional experience, they still have the same timing constraints as scheduled hospital visits. Additionally, hospital equipment is also less available in these settings, making it more difficult to accurately diagnose a condition – instead, the medical professionals might screen these infants for whether to proceed with further testing at the hospital. These more approximate screening methods may not be reliable enough to catch all early cases. For instance, newborns are most commonly visually assessed for severe jaundice even though it is known to be unreliable, as the cost of the more accurate medical equipment can be prohibitively high for outpatient settings.

Where warranted, remote monitoring is a strategy that offers families much more frequent access to professional expertise. In this case, caregivers regularly log digital information about their infants for medical experts to review remotely. This strategy is the current state-of-the-art for patients with single ventricle heart disease, through the aforementioned CHAMP system, which has dramatically improved survival rates. A limitation of such a remote monitoring system is that the availability and timing of the professionals' attention dictates how early they can identify signs of decompensation (i.e., worsening of a patient's condition). Their expertise, while significantly more available than one or two scheduled visits, is not necessarily immediately available for every individual data entry. Hours can elapse before they examine data that indicate early signs of a medical complication, possibly making it too late for an early, less severe intervention for highly acute conditions.

In all of these cases, caregivers are usually given instructions on what warning signs their infants may exhibit and asked to reach out to professionals if they notice them. Doctors prepare new parents for jaundice by asking them to call

in if their child looks suspiciously yellow. Medical care teams for single ventricle patients teach caregivers about red flags, such as specific weight gain rates, oxygen levels, and increased work of breathing to watch out for, and instruct caregivers to page them if any of the red flags become apparent. While attentive caregivers can offer the continuous and immediate attention necessary for early detection, they rarely have the professional knowledge and experience to notice until the condition becomes more obvious, at which point time has passed since the earliest intervention was possible.

1.3 Thesis Statement and Research Contributions

In light of the aforementioned goals and approach, my thesis statement is as follows:

> **The integration of machine learning and human-centered design can forge consumer devices into tools that can help monitor infants for medical conditions in outpatient settings.**

Supporting this statement involves demonstrating its applicability to concrete examples. I do so through exploring applications for two medical conditions that span opposite ends of the spectrums of rarity, complexity, and the presence of preexisting systems to integrate with: neonatal jaundice and single ventricle heart disease. Hence, this thesis addresses the following research aims:

1. *Investigate the potential of using smartphones and machine learning to screen neonatal jaundice.*

   Jaundice is a phenomenon that results in a distinct reflection of visible light, which can be captured and measured by light sensors. This characteristic is the operating principle behind the transcutaneous bilirubinometer (TcB), an FDA-approved medical device for measuring neonatal jaundice that involves directly touching optical sensors against the newborn's skin. As smartphones have cameras, they also have optical sensors to capture this phenomenon. Hence, I investigated the hypothesis of whether a tool can be fashioned out of smartphones and machine learning to noninvasively estimate bilirubin levels in neonates within the same bounds of accuracy as the TcB. As an outcome, I contribute BiliCam: a system that pairs smartphones with a combination of HCI, computer vision, and machine learning to noninvasively estimate neonatal bilirubin levels. It involves computer vision techniques to automatically segment images for BiliCam, linear regression models that use these images to estimate neonatal bilirubin levels, HCI and computer vision techniques to ensure a baseline level of quality of image data. I also contribute an analysis of how constraints on the distribution of self-reported race and lighting conditions in photos can affect the accuracy or generalizability of a color-based system like BiliCam.

2. *Investigate opportunities for machine learning to improve early detection in CHAMP.*

   Recent studies demonstrate that specialized medical professionals reviewing CHAMP data often detect signs of acute patient decompensation earlier than their caregivers would. Because of the importance of early detection and the frequency of delay between caregivers recording this data and professionals reviewing it, there is an unmet need to immediately help triage and prioritize the medical team's attention with each data entry. For this reason, I worked with the CHAMP medical team at Children's Mercy Hospital approached to investigate

this hypothesis: whether machine learning can be integrated into CHAMP, a system that uses off-the-shelf measurement devices and a tablet PC, to assist professionals with the detection or triage of medical complications in interstage single ventricle heart patients at home. Doing so involved applying HCI techniques to unpack and drive to the heart of the current pain points. I contribute concrete, actionable goals and applications for machine learning in the preexisting CHAMP system to increase early detection of medical complications for interstage patients living at home. To further its impact, I also contribute an example machine learning method that could provide an automatic partial triage to improve the distribution of mental resources for CHAMP patients. This contribution includes building infrastructure to navigate the dataset, a pipeline to wrangle data into machine learning inputs, an example model, and an outline for an example use case for how to apply it. This example model leverages the techniques of boosted random forests, trained on retrospective medical data for the classification of patient health status. From the process of analyzing and ranking machine learning features, I also contribute the first ever empirical analysis and evaluation of CHAMP's current automatic classificaton system. This analysis offers benchmarks of current practice for future systems to compare against, as well as empirical validation for some of the prior medical intuition originally used to develop CHAMP.

3. *Insights for the development of similar tools.*

   The high-level contribution of this thesis are insights for when and how to develop tools that combine machine learning and commodity hardware to screen acute medical conditions in outpatient settings. The experiences that come from developing BiliCam and working with CHAMP offer insights that people could use for guidance when developing similar tools. Successful aspects of my process can serve as a framework for future processes, and my lessons learned can serve as insights to help people avoid repeating the same mistakes.

1.4 Document Overview

Descriptions of related work are split over three different parts of this document. In chapter 2 is a general overview of literature and commercial products pertinent to this thesis. Overviews of related works that are highly specific to BiliCam and CHAMP, such as their medical background and alternative approaches, are located at the start of their respective chapters (section 3.1 and section 4.1, respectively).

BiliCam is described in depth in chapter 3. The chapter begins with the relevant medical background necessary for understanding the space and a review of what other approaches have been explored. As BiliCam is a large project, a

project overview precedes all other sections in this chapter. Work on BiliCam involved multiple iterations on the entire project, including study design, app design, and data analysis. To illustrate the specific findings that informed these iterations, the chapter individually describes each study in chronological order in their own respective sections. An overview of all three studies and the common thread that ties them all together precedes these three sections. The chapter closes with a discussion on findings, limitations, future work, and impact.

Work on uncovering AI opportunities for CHAMP in depth is described in chapter 4. Like BiliCam's chapter, this chapter begins with the relevant medical background necessary for understanding the space and a review of what approaches have been explored so far. The section after describes further concepts that are necessary to understand work with the CHAMP system: the available materials to work with. The chapter then describes my work and contributions: the results of a design ethnography to uncover what specific applications of machine learning can benefit CHAMP, and an example machine learning method to explicitly demonstrate how it can be possible. It closes with a discussion on findings, interpretations of the example machine learning method's results, limitations, and future work.

Reflections and insights for working in this space is discussed in chapter 5. The chapter is broken down into two parts: recommendations for how to decide whether to pursue a particular direction or apply machine learning, and general advice for working in this space based on the experiences in working on BiliCam and CHAMP.

Finally, chapter 6 concludes this dissertation. While this chapter reiterates some of the main contributions of this thesis, refer to the discussion sections (section 3.7 and section 4.5) for more on contributions from BiliCam and work on CHAMP, respectively.

Chapter 2

Related Literature

This chapter provides an overview of the literature and commercial products related to my work. Note that I cover literature that are uniquely specific to BiliCam in section 3.1 and CHAMP in section 4.1 instead. They cover the medical background necessary to understand these projects and alternative monitoring approaches for these medical conditions.

2.1 Risk Prediction in Hospital Settings

While my work focuses on outpatient settings, there are a number of similarities and insights to draw from work that focuses in inpatient settings. Inpatient settings are ripe for machine learning, especially because they can enable the collection of large quantities and types of data (a challenge for outpatient settings which, as a result, have fewer examples). Like my efforts for CHAMP, much of the work in these inpatient settings draw on the ability and intention to combine multiple factors and physiological measurements for predictive health monitoring.

2.1.1 Big Data and Electronic Medical Records

The rise of computation has enabled hospitals to increasingly store patient data in electronic medical records (EMR). Combined with the rapid development of more sophisticated analytic tools and computational power, they open many opportunities to improve patient care with machine learning and big data. It offers the power to expand the capacity to generate new medical knowledge, help disseminate knowledge, personalize some medicine initiatives, and offer ways to deliver information directly to patients [90]. More directly related to my work, these advances in clinical analytics are also introducing new methods to screen and monitor patients. A number of efforts are harnessing big data to predict

a patient's medical risks, which is especially helpful in reducing costs for high-cost patients, hospital readmissions, triage, decompensation (when a patient's condition worsens), and adverse medical events [11].

Hospital readmissions have a track record of generating high expenses [60], and while they remain prevalent and costly, they are also largely preventable [63]. Although executing interventions are part of addressing this issue, calls to reduce hospital readmissions describe how an important part of the process is also predicting or identifying where these readmissions occur [33] – a suitable task for machine learning. Several efforts in the cardiology, including my work for patients with single ventricle heart disease, sit in this space.

One approach is to analyze EMR for binary predictions on whether patients with congestive heart failure will have a hospital readmission within 30 days of discharge. Unlike my work, these models have data from millions of patients to draw from and the ability to apply big-data techniques [146]. While they do not necessarily base their models on sensor data, their promising results speak to the effectiveness of incorporating patient background information like demographics [4] – a concept I integrate in my work on CHAMP. Follow-up work has shown promising ways to translate these models to proactive reductions in hospital readmissions. One such application involves using these models to allocate resources for patients; those with a predicted high risk were allotted the scarce resources for intensive evidence-based interventions, dropping overall readmission rates in a prospective study with over a thousand heart failure patients [5].

Another approach to improving health care with EMR and machine learning involves taking continuous patient measurements. Through integrating real-time physiological measurements with other EMR, a model can repeatedly update a patient's "risk score." One such approach, dubbed the "Rothman Index", aims to predict one-year post-discharge mortality independent of the patient's diagnosis [44, 104, 105]. While it has less to do with screening specific acute medical conditions, this approach comes closer to my work in that it leverages sensor data for physical indicators of the patients' health.

2.1.2 Real-Rime Risk Scores for Infants

Hospitals have long used risk scores to identify infant decompensation; the Apgar score for inpatient newborn triage has withstood the test of time [45] since its introduction in 1953 [7] that revolutionized the management of newborn resuscitation. A similar inpatient risk score that encompasses older children is the Pediatric Early Warning Signs, or PEWS, which involves only a few relatively quick measures to predict whether a patient needs more attention [87]. Pediatric cardiovascular patients have also benefited from a modified version of PEWS, called C-CHEWS (Cardiac

Children's Hospital Early Warning Score) [84]. All of these scores benefit from the speed of using metrics that can be visually assessed in a brief moment (e.g., using pallor as a proxy for blood profusion). However, these same metrics also introduce the drawbacks of biases from their unavoidable subjectivity [3, 7].

Sensor data and the consistency of computational evaluations offer an alternative approach that has the potential to improve upon these existing scoring systems. Researchers at Stanford investigated using a non-parametric Bayesian model that uses continuous sensor data, such as heart rate, to compute morbidity risk scores for premature infants. So far, studies reveal promising results that outperform the Apgar score [113, 114], supporting the potential of using sensors and machine learning to improve health monitoring.

In addition to these scoring systems for general health, sensors and computation are also useful for monitoring specific medical conditions. One example is central apnea, a common but serious clinical problem for premature infants in which they stop breathing for too long and need immediate medical attention. In a neonatal intensive care unit (NICU), signal processing and modeling techniques improve apnea detection through the use of electrodes that simultaneously monitor the patient's electrocardiogram and chest impedance [70]. Another example is early-onset sepsis in newborns. Through a combination of modeling techniques and samples from over 600,000 patients, researchers were able to develop a risk stratification for early-onset sepsis based on vital signs during the first 24 hours of life as well as maternal factors, demographics, and specific clinical milestones (such as seizures or apnea) [43]. Much like the techniques discussed earlier (e.g., [4]), this work demonstrates how background information about the patient can improve specific risk prediction models – an approach I use in my work on CHAMP.

Of the different strategies to compute risk scores for pediatric patients, the closest to my work on CHAMP is [109]. Here Rusin et al. also work on predicting deterioration for infants with parallel systemic and pulmonary circulations (i.e., their hearts do not separate oxygenated and deoxygenated blood, and pump this mixture to both the lungs and body in parallel, unlike my typical serial circulation which connects the lungs and body in series) – a category that generalizes from infants with single ventricle heart disease before their second surgery. This patient population presents many of the same challenges to their work as to ours. Their physiological abnormalities effect their baseline vital signs in a number of ways, often rendering existing metrics and monitoring technologies unsuitable. The rarity of this disease also guarantees a very small sample population (25 patients in their case) which stands in stark contrast to many of the aforementioned works with data from hundreds of thousands or even millions of patients. We both take a similar approach to address these challenge: rather than comparing infants who eventually experience decompensation against infants without, I use multiple samples from the same infants and base my models on the assumption that

"physiology immediately before deterioration is abnormal, and that physiology not in close proximity to deterioration events is stable" [109].

The inpatient setting of their work, however, prevents a number of their insights from being usable in my work. All of their participants were in the cardiac intensive care unit (ICU), which enabled the use of continuous bedside monitoring and medical equipment unavailable at home. Some of the highest predictors of decompensation that Rusin et al. found are only available with this type of monitoring: respiration rate variability, beat-to-beat heart rate variability, and several aspects of the ST segment from an electrocardiogram (ECG). Their continuous measurements offered them not only the fine resolution required for these kinds of features, but also the ability to analyze data changes on an hourly scale whereas CHAMP only offers a few data points per day at best. The same comparison can be said for the other aforementioned real-time risk score methods.

2.2 Infant Home Monitoring Devices

Compared to the landscape of infant health monitoring in the hospital, the available methods for home monitoring are very different. Many of the methods that incorporate technology tend to focus on general tracking, frequently more for a caretaker's personal use or reassurance than for specific medical conditions. For most of these cases, it is up to the caretaker to ascertain meaning from the data they collect, whereas my work focuses on incorporating machine learning to assist this process. It is also worth noting that the U.S. Food and Drug Administration (FDA) does not yet regulate these technologies. An examination of all the available FDA Reports to Congress (years 2008, 2009, and 2012-2016), which include reports on recently cleared or approved devices for pediatric use, reveal that the only devices approved for infants those years are laboratory tests and, for ages 2 or more, glucose monitoring devices [96]. To illustrate the landscape of infant home monitoring devices, this section instead describes products on the market and approaches documented in research studies.

2.2.1 Smartphone-Based Methods

The most available technology for infant home monitoring comes through smartphone app stores. There are many apps available in which caretakers can log a number of different aspects of their babies. For instance, with the Baby Connect app, parents, nannies, and day cares can collectively log a child's feedings, nursing, naps, diapers, milestones, mood, temperature, photos, Global Positioning System (GPS) location, and more [8]. Similarly, Trixie Tracker™– primarily marketed to log a baby's sleep schedule – also lets caretakers make detailed entries about breastfeeding, solids and

foods, pumping, their milk inventory, and medicine doses [136]. While these apps offer a very flexible design to help parents track almost anything to minute detail, interviews indicate that this approach can also be overwhelming and burdensome [57].

Some researchers at the University of California Irvine explored an alternate approach, in which they designed principles, prototyped, and ran a proof of concept study for a much more intentional preterm health tracking app [57]. In this instance, parents still logged information about their infants through a smartphone app, but they were trimmed down to only two kinds of data (weight and diapers) that have specific clinical significance for premature infants who are at higher risk for slow weight gain and concerning gastrointestinal infections. Another key difference is that this app is part of a larger system in which these parents receive coaching from both a virtual coach and medical professionals (readily accessible through a clinician portal) and researchers provided parents with a baby weight scale. My work on CHAMP shares this combination of targeting a specific at-risk infant group, incorporation of medical equipment for parent-entered measurements, and professional medical assistance or oversight. In comparison to both my work on CHAMP and BiliCam, [57] operates on a much smaller scale (both in terms of the quantity of measurements and the number of participants), depends on its users to find patterns, and focuses on general health instead of acute medical conditions that require immediate attention.

There are several works that investigate ways to use smartphones to screen neonatal jaundice [9, 38, 51, 73, 92, 119, 125, 132, 137], which I describe in section 3.1 as part of the background materials for BiliCam.

2.2.2 Wearable Devices

Another approach to using technology to monitor infants lies in the area of wearable devices that monitor temperature for hypothermia. The World Health Organization (WHO) recognizes hypothermia as a major cause of newborn illness and death in low resource settings [65]. Preventing and managing neonatal hypothermia does not require special equipment, as it can be done through Kangaroo Mother Care (KMC), a technique of prolonged, continuous skin-to-skin contact between mothers and their infants. Detecting that the child may be entering hypothermia and needs KMC, on the other hand, is an active area of research – something especially relevant in hospitals short on staff to monitor infants as well as at home. One approach lead to the development of ThermoSpot, a sticker made with a liquid crystal that changes colors at a temperature threshold and only costs 7 Indian Rupees. Several in-hospital studies suggest that the ThermoSpot has the potential to improve hypothermia-related outcomes and reduce the workload of over-stretched staff in low-resource hospitals [86, 99]. One study shows promise for ThermoSpot use in low-resource homes, through

the care of non-medically trained local volunteers in an Indian urban slum [52]. Another approach to monitoring neonatal hypothermia is through a BEMPU Bracelet worn around the newborn's wrist, which flashes a different color and sounds an alarm when it detects hypothermia [12]. This method showed promise in a prospective study that took place at a hospital step-down unit, which more closely resembles a home environment than the NICU [127].

A number of more expensive wearable devices targeted toward monitoring infants at home have also been cropping up in the market. Much likes the microphone- or camera-based baby monitors that parents can place near a crib, some of these wearable devices are built to monitor a baby's sleep or movement activity [85, 93, 123]. Some wearable devices also target more specific physiological signals, such as the Owlet's "smart sock" for infants, which measures heart rate and oxygen saturation and alerts parents if they drop below a threshold [93], and Snuza's breathing monitor, which clips onto a diaper and vibrates to arouse the baby if it doesn't detect breathing for 15 seconds and alerts parents for longer stretches of time [123]. Their websites tout multiple awards and hundreds of positive testimonials, suggesting that these products are very well-received. That being said, none of these products are medical devices – which their respective websites explicitly state in their FAQ's – and have no connection to specific medical conditions. They emphasize that the devices are "intended to provide peace of mind."

However, "the peace of mind for which parents buy expensive but unregulated wearable monitors, and on which their marketing depends, may be illusory" [62]. A number of critical articles point out how these devices point out that these devices can actually cause more unnecessary anxiety and harm [20, 28, 62]. There is no "publicly available evidence supporting the safety, accuracy, effectiveness, or role of these monitors in the care of well infants" to suggest that one can trust these devices to correctly perform their purported purpose. Even if they are 100% accurate with their physiological measurements, they can cause overdiagnosis (e.g., healthy infants can occasionally have their oxygen saturation drop below 80% without consequence), prompting unnecessary emergency room visits, laboratory tests, imaging studies, hospital admissions, anxiety, and a false assumption that the infant is at risk of dying [20].

The sales and marketing of these devices play on parental fears about Sudden Infant Death Syndrome (SIDS) [62], with a number of testimonials on their websites in which parents discuss how they use these products to help assuage their anxieties over SIDS. These products exist in a greater context and history of cardiorespiratory devices marketed toward preventing SIDS. Apnea, the temporary cessation of breathing, was first hypothesized to predict SIDS in 1977 when a few infants with documented apnea during hospitalizations died unexpectedly after discharge. Although evidence of infanticide for each of these patients later emerged, and that the connection between apnea and SIDS fails to have supporting evidence despite extensive independent research, home cardiorespiratory monitoring devices for

SIDS proliferated [97]. Among their recent recommendations for SIDS-related infant safety, the American Academy of Pediatrics explicitly states that "infant home cardiorespiratory monitors should not be used as a strategy to reduce the risk of SIDS" and that there is no data that other commercial devices designed to monitor infant vitals reduce the risk of SIDS [88]. Home cardiorespiratory monitoring can have its uses, but only in specific circumstances; they may be warranted for infants with high risk specific pulmonary conditions, dependent on a technology like a tracheostomy, or specific rare medical conditions that affect their regulation of breathing – and only after training caregivers on how to operate the monitor, observations, and infant resuscitation techniques [97]. To date, the FDA has not cleared or approved of *any* products to prevent or reduce the risk of SIDS [9].

2.3 Mobile Health Monitoring

With their growing ubiquity, on-board sensors, and computational power, smartphones are increasingly becoming a platform for medical and health applications. Prior work in this space inspired and motivated the design of BiliCam.

A number of mobile health-sensing systems augment the phone's capabilities with additional, custom hardware. A common example is exercise and physical activity monitoring systems that use sensors to track movement, such as the commercially available FitBit and Nike+. Work like UbiFit leverages such hardware to provide additional feedback and exercise incentives through a phone's background display [34]. In addition to physical activity, smartphones can measure other physiological signs like heart rate. Poh et al.. demonstrated a means of monitoring heart rate through a PPG attached to earbuds while playing music [100]. Wello, an upcoming specialized phone case embedded with sensors, promises to let people measure heart rate, temperature, blood pressure, pulse oximetry, and lung function from their phone. The space stretches beyond monitoring everyday health — phones are also becoming diagnostic tools. For instance, Franko et al.. built a method of screening for scoliosis using a smartphone and a custom plastic accessory [48]. Smartphones are also beginning to emulate standard medical tools and making them more accessible. For example, Mobisanté develops commercially available hardware that plugs into a phone to generate basic ultrasound images, making a more portable and affordable alternative to traditional ultrasound equipment.

A number of phone-based medical devices, like BiliCam, do not require external hardware and are purely software-based solutions on the existing platform. For instance, there are pulmonary-focused systems that harness the built-in microphone. Prior work uses the smartphones to measure lung function (spirometry) in order to detect and monitor chronic lung conditions, and has achieved results akin to a clinical spirometer [67]. Another project monitors audio signals to track the frequency and quality of coughs, helping patients monitor coughing episodes and objectively report

their coughing frequency to their doctors [68]. Similarly, Chen et al. used the microphone to continually monitor nasal conditions, such as sneezing and runny nose [30].

Like BiliCam, a number of recent explorations of health applications are vision based. Overall, the use of cameras for health sensing is becoming increasingly popular. For instance, some systems use a phone's camera to measure heart rate anywhere and anytime by tracking a person's finger for subtle flushes in the skin from blood flow [71, 72]. Researchers have also investigated assisting rehabilitative physical therapy using a depth camera [29] or with infrared cameras in a touch-screen table [22]. Smartphones have been shown to improve and automate point-of-care diagnostics, which require visually analyzing test results from blood or urine samples on specialized materials [37, 116]. Other camera-based systems directly evaluate physiological conditions. Pamplona et al. demonstrated a method to screen eyes for specific impairments using an instrumented smartphone camera [94]. Also examining the eye using a phone camera, Bourouis et al. developed a method of detecting retinal cancer [23]. Other active areas of research with computer vision include recognizing skin cancer [139] and tracking chronic foot ulcers from diabetes [141].

Chapter 3

BiliCam: Using Mobile Phones to Monitor Newborn Jaundice

BiliCam is a system I developed that uses smartphones to monitor newborn jaundice, which manifests as a yellow discoloration of the skin. Although a degree of jaundice is common in healthy newborns, early detection of extreme jaundice is essential to prevent permanent brain damage or death. Current detection techniques, however, require clinical tests with blood samples or other specialized equipment. Consequently, newborns often depend on visual assessments of their skin color at home, which is known to be unreliable. To this end, I present BiliCam, a low-cost system that uses smartphone cameras to assess newborn jaundice.

3.1 Background and Significance

3.1.1 Neonatal Jaundice

Jaundice is defined as the yellow discoloration of the skin caused by excess bilirubin, a chemical byproduct of recycling old blood cells. Bilirubin is a natural product of the breakdown of expired red blood cells, which the liver further metabolizes for excretion. The accumulation of excess bilirubin results in the yellow discoloration of the skin known as jaundice. Newborns tend to metabolize bilirubin slower (as their livers may not function at full capacity yet), have blood cells with shorter lifespans, and have higher concentrations of red blood cells than adults. Consequently, jaundice is one of the most common physiological conditions in newborns; up to 84% of them develop jaundice during their first

week of life [15]. A moderate level of bilirubin is normal in healthy newborns. This temporary excess of bilirubin is usually harmless.

However, if not treated, highly elevated concentrations of bilirubin in newborns are neurotoxic, and can be fatal or cause devastating and irreversible brain damage. This potentially lethal condition, called kernicterus, can cause deafness or hearing loss, cerebral palsy, and profound developmental delay. Fortunately, kernicterus is avoidable through early detection and treatment. High levels of bilirubin can be controlled through phototherapy, a process that involves bathing the affected newborn in specific wavelengths of blue light that convert bilirubin into a harmless, excretable form. For extremely high levels, excess bilirubin must be removed through exchange blood transfusions [103].

3.1.2 Screening Methods

Accurate medical tests to assess this condition require a blood draw or the use of a specialized measuring device, making them impractical outside of medical settings. However, bilirubin levels typically peak well after most infants are discharged from the hospital. Consequently, visual assessment is the most common method to monitor jaundice in a family's home, where clinical technology is unavailable, as well as at most outpatient clinics, where administering a blood test is logistically difficult. While parents and clinicians are usually able to visually identify the presence of jaundice, numerous studies show that even experienced healthcare providers cannot accurately estimate the severity of jaundice [103]. The importance of monitoring newborn jaundice at home under these conditions creates the need for an accessible screening system such as BiliCam.

3.1.2.1 Current Methods

To determine whether a newborn should receive phototherapy or an exchange blood transfusion, doctors or nurses reference specialized graphs with the newborn's age, number of weeks of gestation, and bilirubin level [103]. One such graph is the Bhutani Nomogram [14], like the one shown in Figure 3.1, which came from an extensive study by Bhutani et al.. The nomogram provides a means to assess a newborn's risk based on the percent of newborns in the study with given bilirubin levels and ages. High- intermediate risk is considered above the 75th percentile, and high risk above the 95th percentile. Bilirubin levels are commonly expressed in milligrams per deciliter (mg/dL) or micromoles of bilirubin per liter (μmol/L) [103].

Clinicians measure the blood concentration of bilirubin on a continuous scale with either a TSB or TcB. A total serum bilirubin (TSB) test directly measures the bilirubin from a blood sample. Although invasive, the TSB is the most

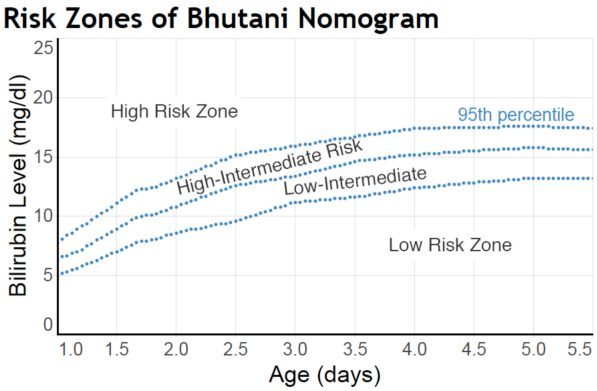


FIGURE 3.1: A Bhutani nomogram used to asses a newborn's risk based on bilirubin level and age, generated with risk zone boundaries from BiliTool™ [18].

accurate way to measure bilirubin and serves as the medical gold standard. A transcutaneous bilirubinometer (TcB) is a specialized meter for a non-invasive alternative that indirectly measures bilirubin levels. Health practitioners touch the end of this device to the newborn's forehead or sternum. It emits specific wavelengths of light and measures the resultant reflectance and absorbency of the skin to infer bilirubin levels. TcBs are considered by the medical community to be unreliable above 14.5 mg/dL of bilirubin, thus high measures from TcBs must be followed by a TSB [103]. In this way, TcBs are used as screening tools. A TcB also costs several thousands of dollars and requires frequent calibration. Although large nurseries usually have more than one TcB, I found that TcBs are not common in primary care clinician offices and depend on the expected number of newborn patients. Hence, this screening tool is not available in all clinics due to its cost.

Visual assessments are common in outpatient settings, such as physician offices and a family's home, where the aforementioned technology is unavailable. While both parents and clinicians are usually able to identify the presence of jaundice, there is ample evidence from many studies that even experienced healthcare providers cannot accurately estimate the severity of jaundice through visual assessment. In studies comparing visual assessment of jaundice with TSB levels, correlation coefficients are generally in the 0.35 – 0.75 range with poor inter-observer agreement [103]. More concerning is evidence indicating that clinicians frequently *underestimate* the severity of jaundice when using this method. Visual assessments have even proven unreliable with the aid of reference colors, such as with an icterometer [103]. Icterometers are specialized Plexiglas rulers marked with different tones of yellow to reference when pressed against a newborn's skin [2]. Clinical guidelines explicitly advocate against using icterometers [103].

3.1.2.2 Approaches in Research

A number of studies investigate alternative, non-optical methods to predict high bilirubin levels, including measuring blood bilirubin or antiglobulin in umbilical cords and end-tidal carbon monozide measurement (ETCOc). They have so far proven either unsuccessful or unreliable [103].

Other researchers have taken similar approaches to BiliCam by investigating the use of cameras to estimate jaundice. Many of them cite my first publication of BiliCam [51]. Most of them use proxies for newborn skin for sources of data. [38] developed a custom simulation of the skin, incorporating different equations to represent aspects like the epidermis and dermis' light absorption properties, into which they injected different concentrations of bilirubin. Others took more approximate approaches, such as [119] which used strips of paper soaked with different concentrations of bilirubin and [137] in which the author of this master's thesis took photos of their own skin, using bruises and carrot juice to simulate neonatal jaundice. As newborn skin has different optical properties than adult skin or pieces of paper, my approach differs from these related works by using photos of actual newborn skin.

Some works do incorporate newborns skin in their data collection. In [69], researchers used Photoshop to find a correlation between jaundice measurements and pixels from hand-held digital cameras. Researchers in [89] used Excel to look for trends from photos of 11 Caucasian babies, which they were unable to find from normal photographs (possibly because they did not use any sources of white balancing or color calibration) but succeeded in finding a trend when they pressed a dermatoscope against the skin.

Instead of examining the skin, which is subject to variations in baseline skin tones, some researchers looked into using photographs of the newborn's conjuctiva and sclera (i.e., the whites of their eyes) [73, 92]. Similarly, [83] uses this approach to detect jaundice in adults. Although the skin introduces the challenges of different baseline tones, I decided to focus on the skin instead of the eyes for usability reasons; newborns often have their eyes closed, and I anticipate that forcing their eyes open and holding a camera with a bright LED pointed at their eyes would not fare as well as it would with adults.

A number of approaches to bilirubin estimation involve building custom hardware. [125] developed an external custom attachment with a plastic window and light occlusion to measure jaundice with a smartphone, as follow-up work from [95]. [38] also developed custom hardware for their phone-based TcB prototype. Some researchers did not use smartphones, such as the Arduino-based prototype from [32] and the hand-held rechargeable battery reflectance reader from [145]. While custom hardware offers more control to the physical components of the overarching system, I decided to focus on using uninstrumented commodity smartphones and a piece of paper because of their ubiquity and

availability in outpatient settings.

3.2 Project Overview

The importance of monitoring newborn jaundice outside the hospital creates the need for an accessible screening system. To this end, I spent multiple years developing and investigating BiliCam, a smartphone-based method to assess newborn jaundice.

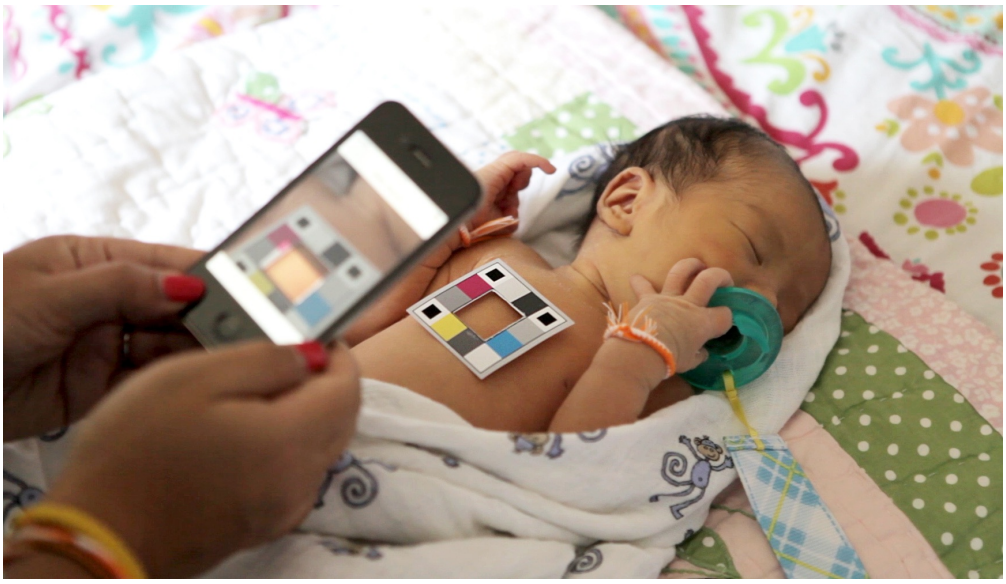


FIGURE 3.2: A photo illustrating how parents or medical practitioners could use BiliCam to monitor a newborn's jaundice with their smartphone.

3.2.1 Concept

As demonstrated by their recent popularity for health sensing in the UbiComp community [1], smartphones offer distinct advantages as a medical platform in terms of cost, accessibility, and computational and sensing capabilities. The programmability and Internet connectivity of these devices allow algorithms to adapt much more effectively. Most importantly, their ubiquity enables a multitude of families with newborns to use their phones as medical devices, helping many of them avoid the cost, anxiety, and hassle of extra hospital visits.

By leveraging these inherent advantages of smartphones, BiliCam mitigates the risks in visually assessing jaundice. BiliCam uses the phone's built-in camera to photograph a newborn. After confirming that the images are usable, the system uploads the relevant portions to a server, which analyzes the newborn's skin to estimate the bilirubin level. It then communicates the results back to the user and recommends a course of action. Each photograph includes a custom, low-cost color calibration card to help BiliCam adjust for different lighting conditions and apply color corrections. Other than the smartphone and the color calibration card, this non-invasive solution requires no additional hardware.

BiliCam's underlying machine learning model is a regression on bilirubin concentration instead of a classification on newborn risk for two reasons. On reason is that the threshold jaundice level for high risk differs with a newborn's age, as illustrated by the Bhutani Nomogram in Figure 3.1. Hence, the skin color alone is not enough information to determine risk, and incorporating age as a feature for classification models adds unecessary complication and could muddy the already clear definition of how age connects to risk level. Another reason is that by mapping predictions to real-world TSB values, a regression enables a direct comparison to TcB for evaluation. The correlation between bilirubin estimates and TSB is a standard metric in medical literature, so a regression's ability to produce this metric enables the head-to-head comparison for the medical community and the US Food and Drug Administration (FDA).

3.2.2 Team

The BiliCam team encompasses many people; the team changed and grew over multiple years to include many data collectors, experts in HCI and design, and experts in commercialization in addition to the core research team. The core research team comprises of two experienced pediatricians from Seattle Children's Hospital (James A. Taylor, MD and James W. Stout, MD) and four members of the University of Washington's Ubiquitous Computing Lab (Lilian de Greef, Mayank Goel, Eric C. Larson, and Shwetak N. Patel, PhD).

3.2.3 Studies

My collaborators and I conducted three clinical studies to develop and validate BiliCam: a pilot study with 40 newborns, a formal local study with 100 newborns, and a nation-wide study with 600 newborns. In these studies, we collected BiliCam photographs of newborn participants and medical measurements of their bilirubin levels. The rest of this chapter describes these studies and their results in detail.

3.3 Overview of Study Design

Our studies have three main goals: collect training data with which I can develop a machine learning model, collect test data to evaluate the machine learning model, and inform and test BiliCam's data capture methods. I separate the training and test data via cross-validation.

I collected BiliCam data and the medical ground truth from real, live newborns instead of phantom skins (i.e., fake, substitute skin) because I wanted to test BiliCam on as close to its usage scenario as possible. Using data from real newborn skin would exclude the possibility of misrepresenting any qualities of skin in phantoms. Furthermore, the human factors involved in collecting data from live newborns introduce additional challenges that need vetting before BiliCam could be considered feasible.

As this work focuses on conducting a proof of concept for BiliCam, I constrain as many factors as possible (e.g., the type of phone, the paper and ink for reference color cards). That way, if the proof of concept fails, I know I can exclude the controlled variables as possible reasons.

3.3.1 Collected Data

A human-centered approach informed the choice of data to collect. This approach involved applying the HCI techniques of semi-structured interviews, focus groups, and contextual inquiry with medical experts. Regular check-ins between the technical and medical experts in the team and thoroughly reviewing the relevant medical literature also contribute to these decisions.

I assigned a unique study ID to each participant to match the participant's medical results and image data while maintaining confidentiality. All data was collected in clinics through dedicated data collectors (often nurses) who I hired for the study.

3.3.1.1 Medical Data

One important insight from working closely with medical experts is the importance of measuring absolute bilirubin levels in a newborn. Without the proper medical background, people may decide to develop a study and system for detecting whether a newborn has jaundice. However, the presence of jaundice in newborns has little medical significance, as most jaundiced newborns are also healthy. Determining whether a newborn's jaundice has reached a dangerous level depends on more than just the yellowness of their skin as well. Their age is an important factor, as illustrated in the Bhutani Nomogram (Figure 3.1). For example, a newborn with a bilirubin level of 10 mg/dL may be at

low risk if they were 3 days old, they would be at high risk if they were 1 day old. Hence, a model trained to classifying skin for whether a child received treatment would make for an ineffective screening tool. I recognize that BiliCam must instead use a regression model on the exact blood concentrations of bilirubin.

The medical data I collected include a TSB blood sample and a TcB. In general, the TSB provided ground-truth data and the TcB as a source of comparison. Our pilot study collected whichever measurements the hospital was already administering, so some of its participants had only TcB. I paid for and required TSBs for all participants in the other studies, and made TcB optional in the formal nation-wide study to enable multi-clinic logistics. I also noted which blood samples were effected by hemolysis, a condition that affects the accuracy of TSB readings [26]. The TcB measures came from a Philips BiliCheck or Draeger Jaundice meter JM-103. All medical measurements (TSB and TcB) were taken within 2 hours of capturing images for BiliCam.

I also collected the newborn's age to enable classification using the Bhutani Nomogram as well as reported race so I can check and report the study population's demographics.

3.3.1.2 BiliCam Data

For each participant, I collected at least one set of multiple images as the BiliCam data. I encouraged the data collectors to collect at least two sets of images for redundancy in case there are any problems with the first set.

Every set of images includes pairs of photos at the same fixed distance from the newborn – one photo with the flash LED turned on, another with the LED turned off. In the nation-wide study, I introduced additional distances.

Each photograph includes a custom, low-cost color calibration card to help BiliCam adjust for different lighting conditions and apply color corrections. While the card design differs between the local studies and the nation-wide study, they are all approximately (if not precisely) business-card sized and include black, white, grey, cyan, magenta, and yellow color patches. I included cyan, magenta, and yellow because I can print them such that only the respective inks were used on each patch; the colors did not contaminate each other. Meanwhile, white and grey are common choices for white-balancing in photography, and including both white and black also enables me to check for clipping from over- or under-exposure.

I segmented each image to extract the pixel values of the sternum and color patches on the color calibration card. In the local studies, I also segmented the forehead from images. The sternum and forehead are the primary locations of interest for skin samples for several reasons. Medical practices standardize TcBs, which are also light- based, to take readings from these two locations. Both the forehead and sternum also offer prominent, flat regions of skin on which I

expect even lighting. I expect the whites of a newborn's eyes to be more consistent across skin tones and potentially a better location of interest for a visual system. However, the eyes are closed (e.g., while sleeping or crying) the majority of the time. Even when open, the whites are hard to discern, given their small size compared to the iris.

To focus on whether the *concept* behind BiliCam is feasible, I controlled as many variables as possible. That way, if the results of the work fails to support the efficacy of BiliCam, I know it's not due to the variables I controlled. For each study, I restricted data collection to come from only one type of phone (e.g., the iPhone 4S in the pilot and formal local study, and the iPhone 5S for the nation-wide study). I also designed the data collection app to standardize factors like white balancing and constrain the position of the phone relative to the baby and card (see App Features in subsection 3.3.2). Through both talking to the data collectors and instructions on the BiliCam data collection app, I also standardized the placement of the card on the baby.

To make the most out of the data collection process, I collected multiple kinds of data from each participant (while keeping the data collection demands within reason for the data collectors). The idea here is that it's easier to discard data I later decide is unnecessary than it is to redo a study because I failed to collect data I later realize is important (these studies take a remarkable amount of time, money, coordination, and effort!). In the pilot and formal local study, I collected still images with flash, still images without flash, and 10 second videos in case BiliCam needs to account for brief changes in skin color, such as subtle flushes from blood flow. I also designed the card to feature multiple color patches instead of just white. In the nation-wide study I focused on still images of only the sternum based on evidence from the local studies, but introduced taking photos at multiple distances from the participant. I also decided to take additional BiliCam data from participants when they were less than 24 hours old, during which they should have little to no bilirubin in their system in case these additional "baseline" photos were necessary for compensating differences in skin tone, serving as a reference point for how follow-up skin color changed as a result of bilirubin.

3.3.1.3 Participant Requirements

I captured BiliCam and medical data from newborns at the age in which they would have the widest range of bilirubin levels (i.e., the first week of life, if not between 2.5 to 5.5 days old, depending on the study). As mentioned at the end of subsection 3.3.1, I also captured "baseline" BiliCam data within the first 24 hours of life, during which the newborn's bilirubin is typically very low. I required these "baseline" photos in the formal local study, but made them optional in the nation-wide study because results from the local studies suggest they are not necessary and this logistical flexibility enabled me to collect significantly more data.

I limited enrollment to English-speaking parents of newborns who were born at more than 35 weeks of gestation (i.e., full term newborns). Participants who required phototherapy prior to the follow-up became ineligible, due to the effect of phototherapy on skin color, which is a known issue for the TcB [121].

3.3.2 Overview of Data Collection App

I designed, developed, and used custom data collection apps for the studies. This subsection contains an overview for the data collection app. Other sections in this chapter dive into more detail about the data collection app. In particular, subsection 3.4.1 and subsection 3.4.3 contains details on early app development in the pilot study. The app underwent major revisions for the subsequent studies, for which subsubsection 3.5.1.4 describes improvements for the formal local study and subsubsection 3.6.1.4 describes major changes for the nation-wide study.

3.3.2.1 Reasons to Develop a Custom App

There are several reasons that developing a custom data collection app was critical for conducting the studies rather than using the phone's native camera app.

Developing our own app enables me to control and constrain aspects of data capture. For instance, I can fix camera properties such as the white balance. By incorporating a viewfinder for users to align with the color calibration card, I can also constrain the distance, position, and angle of the phone relative to the newborn and card.

Also important to controlling variables in the data, the custom app enabled me to develop and enforce criteria for consistent BiliCam-specific data quality. I were able to introduce automatic image quality control within the data collection app for the formal BiliCam studies.

A custom app also enabled more control over security measures. I developed the app to store data in password protected zip files, upload data to a specific password protected server, and store backup data on the phone in a way that was also password protected and only accessible to the app developers on the team.

Designing the app meant being able to design the user interface to fit the workflow. Collecting data from patients in hospitals made it important for any interactions with the app to be quick and require little attention so as to not compromise the interactions between the patients and doctor or nurse. The intrusiveness of the data collection process may impact not only the experience for everyone involved, but also whether people would decide against participating should the demands be too high.

By making a custom app, I was able to dictate how and where the app stored data not just for the aforementioned security reasons but also to make it fit into the workflow for data analysis. I also tailored the app to link every set of images with the corresponding study ID and other relevant data, which simplifies the workflow for matching photos with other information like TSB results.

As one of the goals of the studies is to inform and test BiliCam's data capture methods, making a custom app also enables me to test and iterate on that very capture method. This process has repeatedly informed the data capture method and user interface. I share these insights in several subsections of this chapter (e.g., subsection 3.4.3, subsubsection 3.5.5.3, and subsubsection 3.6.1.4)

3.3.2.2 App Features

While the data collection apps have many differences between studies, they all have several important features in common.

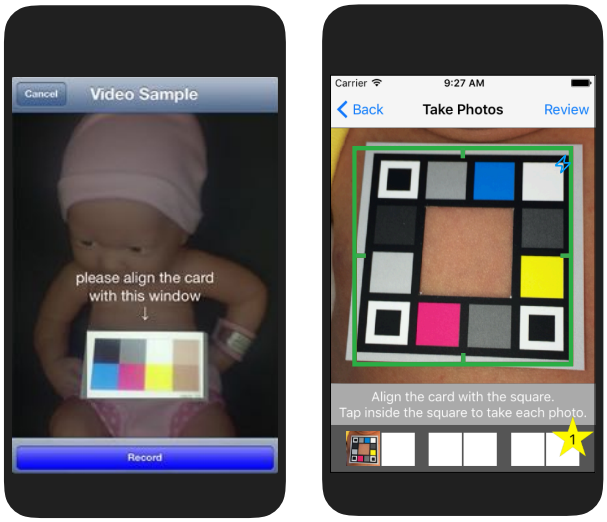


FIGURE 3.3: Screenshots of different version's of the viewfinder in the data collection application, with (A) used in the local studies and (B) used in the nation-wide study.

To constrain the position of the phone relative to the baby and card, the app overlays the camera view with a rectangular "window" (i.e., a viewfinder) which users need to align with the color calibration card, as shown in Figure 3.3. For further consistency, it also offers instruction for where to place the card on the newborn. This combination of

features enables the app to maintain consistent phone positioning relative to the card and baby for each image (in all three cardinal directions and roll, pitch, and yaw).

The app uploads all collected data to an encrypted server and also stores encrypted back-up copies on the phone for redundancy. The server also regularly creates back-ups (which was critical one time when the server was unexpectedly wiped!). As I cannot rely on consistent WiFi connectivity, uploads would happen in the background and the app would try to re-upload any samples that hadn't yet made it to the server over intervals in time.

The app offers people access to a log of all recorded samples. This log enables people to check samples for mistakes (e.g., incorrect sample ID entries), delete samples (which can be important if images accidentally displayed sensitive information, such as information written on the newborn's hospital wristband), and check each sample's upload status.

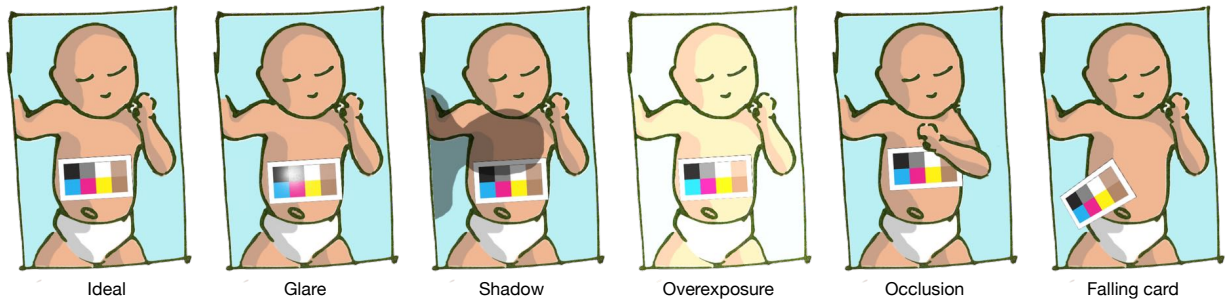


FIGURE 3.4: Illustrations of a few of the common problems for BiliCam's image quality, with an illustration of the "ideal" photo on the far left for reference.

One of the first and most important insights from collecting data from real newborns was that capturing clear photos of the skin and card is difficult. A few of these problems are summarized in Figure 3.4. Babies can move a lot, creating motion blur or occlusions with their hands. Among other issues, the phone and data collectors can cast shadows, the light of the flash LED or surroundings can create specular reflections or glare on the card, adults may try to hold the card and unintentionally bend it or partly occlude it with their fingers, body motion and poor adhesion can make the card start sliding off, poor camera focus can make the image blurry, and poor lighting conditions can clip the range of recorded colors of the color calibration card through overexposure or underexposure. While recognizing and avoiding these issues is intuitive for the technical members of the BiliCam team with computer vision experience, it was not intuitive for the data collectors. Over the course of the data collection studies, I iterated on methods to control for image quality – a process that resulted in significant improvements to the data quality which I believe impacted the accuracy of the algorithm.

These methods for quality control changed over the course of the BiliCam studies. To improve quality as soon as possible, I first spoke with the data collectors in person and found that walking through example images together helped them understand what I mean by image quality. I further created a pamphlet and built a photo-quality tutorial into the app for data collectors to reference for themselves or when on-boarding new data collectors. Examples of this tutorial from BiliCam's nation-wide study are shown in Appendix A.

Once I had these initial methods in place, I began building methods to automatically detect image quality problems. For both the formal local study and the nation-wide study, I built automated quality control natively into the app, which leverages the image segmentation algorithms (described in subsubsection 3.5.1.4 and subsubsection 3.6.1.4, respectively). This automated quality control developed in a progression, as dictated by the constraints on the nature of the data and time constraints for developing and testing the code. It progressed from providing automatic feedback *after* the app finished recording samples (which was necessary when I recorded videos) in the formal local study, to providing visual feedback at the same time as displaying the camera view during the early parts of the nation-wide study, to automatically capturing images based on image quality instead of asking people to manually take them for the majority of the nation-wide study.

3.3.2.3 App Development Process

It's important to make a well-designed app for multiple reasons. In addition to making sure it meets the aforementioned reasons behind the developing the app in the first place, its usability is also critical for running successful studies. If anyone struggles to use the app, they may not capture the right kind of data (if any data), miss important features in the app, or fail to transfer the data for analysis. These UI failures would lead to failures in the development and evaluation of BiliCam – both for its machine learning model and its data capture methods. Hence, I want to be very intentional about the app design and implementation.

To illustrate this point with an example: during one of the studies, one of the data collectors struggled with the user interface to capture photos of a participant. I had recently incorporated quality control by requiring images to fit within certain tolerances, but didn't include the option to upload images that didn't meet these tolerances, had made the tolerances too tight for this situation, and didn't give the user feedback on what to adjust – adjustments that may be intuitive for computer scientists working on computer vision, but would be unfair to expect others to know. Throughout the course of repeatedly trying to re-take these photos, the newborn participant screamed and cried (common for newborns who suddenly had their clothes removed). Our data collector's struggles with the app

took too much time – over the course of these user interface failures, the parents no longer wanted their child to experience such discomfort and decided to withdraw from the study. It was an emotionally stressful situation for everyone involved: the child, the parents, and the data collector. Our data collector was under a lot of pressure in this situation, and had even come in on a holiday to collect this data, forgoing time off of work to meet the timing of this child's birth. In addition to this burden on time, effort, and emotional stress, this user interface problem was also costly from a monetary perspective. When dividing the cost of the grant over the number of participants in a study, we're looking at roughly $1,000 per participant. In other words, every lost piece of data from user interface problems cost at least that much.

Our app development process involved both conducting formative work before building the data collection app and an iterative process over the lifetime of all the data collection studies.

Early formative work included conducting semi-structure interviews and focus groups with pediatricians and nurses to understand the app's needs, as well as contextual inquiry at the local neonatal intensive care unit (NICU) to better understand the context in which people would use the app for data collection. Some of the important insights from this process include how the app's workflow needs to be fast and require as little thought as possible in order to fit in the workflow of the hospital and interactions with patients. It also needs to handle inconsistent WiFi connectivity, as was the case for the study site. More details on this process and its results are documented in subsection 3.4.3.

Before developing data collection apps, I created paper prototypes, user-tested them with different people, conducted feedback sessions with the team's medical experts and data collectors, and iterated on its design accordingly. Further user-testing of the data collection apps after development helped fine-tune and correct usability before deployment. This entire process applied to both before the pilot study and before the nation-wide study. A result of this process for the nation-wide study was a complete re-design and rebuilding of the data collection app, incorporating valuable improvements (detailed in subsubsection 3.6.1.4).

Over the course of the local data collection studies, I regularly asked the data collectors for feedback on the app. By creating an open channel for communication and inviting critical feedback, I were able to quickly address challenges and adjust the app. It also enabled me to iteratively implement, try, and improve new features such as incorporating different levels of image quality control.

3.4 Pilot Study

Preparing for and conducting a pilot study was the first step to determining if BiliCam is feasible. This exploration began with developing a study, developing a data collection method, running a pilot for data collection, conducting contextual inquiry and feedback sessions to improve future data collection, and some value sensitive design.

3.4.1 Data Collection

As data collection is an integral component to this research, I invested much of my time in developing a data collection app and method. A pilot study in collaboration with two pediatricians and one nurse collected samples from 40 newborn participants at two local hospitals.

3.4.1.1 Collected Information

Initial data collection includes two photos and a 10-second video per sample. One photo uses flash, the other does not. The video is composed of 5 seconds without flash followed by 5 seconds with flash turned on. The photos and videos are compressed as little as possible. I encouraged the data collectors to capture at least two sets of these images, in case one of them had poor quality (a valuable practice!). When possible, a photo-video set would be taken of a newborn within their first 24 hours of life, to serve as a baseline sample to compare a later sample against. I anticipated that future data collection may require less information (as was the case), but I wanted to cover as many bases as reasonable until I have insights from the later technical component of bilirubin-estimation.

For the pilot study, I used whichever medical measurements the hospital staff already took of their patients. I therefore had a mixture of measurements from a transcutaneous bilirubinometer (TcB) and Total Serum Bilirubin (TSB) blood tests. I required measures of bilirubin levels to be taken within a 2-hour window of the photo and video samples. TcB measurements are nearly instantaneous, so their results can be recorded in the data collection app with the photos and videos. TSB measurements require drawing blood and post-processing, which can take several hours. Hence, TSB results needed to be recorded separately from the photos and videos.

In addition to images and TcB measurements, the app also asks people to record a study ID. Because TSB measurements need to be recorded separately from the photos and multiple samples can come from the same newborn, particularly when there are baseline samples, the study ID serves as an anonymous way to match samples accordingly. Only the pediatricians and nurses would know which study ID matches to which newborn.

The app also asks people to record a time of birth in order to calculate the newborn's age; BiliCam must know number of hours a newborn has been a live at the time of measurement to determine how dangerous a bilirubin level is, as well as recommend a course of action.

Additionally, the app automatically records a unique identifier for each phone, as it could help identify any systematic discrepancies resulting from using different devices.

3.4.1.2 App Prototyping

A data collection application was designed through iterative paper prototyping. You can see subset of the earliest paper prototypes in Figure 3.5. Paper prototypes fleshed out the vast majority of the user interface's details. User studies with several members of the lab and an overview with collaborating pediatricians provided feedback, upon which the prototype was changed.

Several objectives determined design choices throughout the prototyping process. One objective is to make the collection process streamlined, as the medical practitioners collecting data are already very busy and prefer a straightforward and quick application. Because future data collection would involve many practitioners, another objective is to be intuitive enough for all of them to understand and use it effectively without guidance. A third objective is to adhere to promises made to the Institutional Review Board (IRB) and protect a participant's privacy. Another objective is to keep as many variables constant as reasonable, such as the location of the standardized color card in images or the distance people hold the phone from the newborn during image capture.

3.4.1.3 First App Implementation

The application sketched out by the prototype was implemented as an iPhone application, available for both iPhone 4S and iPhone 5, as they had the highest quality cameras available at the time of the study and also both have flash capabilities (a feature that was not a given for smartphone as the time of this study). I also decided on using iPhone over Android because the iPhone had more consistent hardware (versus the wide variety of Android phones for the same OS) and had a larger user base with which to disseminate BiliCam at the time.

Here are some highlights on what the data collection app, dubbed BiliSampler, looks like and how it works.

Upon hitting a button to take a new sample, a sectioned table view prompts the data-collecting practitioner to fill out the necessary information Figure 3.6. The TcB result is optional, as some samples would depend on TSB results

FIGURE 3.5: A subset of paper prototypes from the early design of BiliCam's data collection app.

FIGURE 3.6: Left: the opening view for taking a new sample, with nothing entered yet. Right: an example of data entry.

or be baseline photos, which have no medical bilirubin tests associated with them because the pilot study does not explicitly have IRB approval to make additional medical tests beyond those already in the standard of care.

All data entered into the application can be modified, initiated by tapping it, before submission. The "Submit" button is enabled only when the sample entry is complete (with TcB left optional).

Upon selection to record a video sample, the screen displays a brief reminder of how the practitioner should position the standardized color card and newborn Figure 3.7a. The following view (Figure 3.7b) is mostly a direct feed from the phone's camera, tinted save for a small window to align the standardized card with. This window constrains the card and newborn to a general position and distance from the phone. The system auto-focuses the camera to whatever is visible within this window, repeatedly in moderate intervals, until recording begins.

Upon hitting the Record button, the system automatically records both photos and videos both with and without flash. It also displays a progress bar to keep the practitioner informed, followed by text informing its progress with saving the video. The Record button becomes a Stop button during recording, in case something happened to prompt the practitioner to start over (e.g., card sliding off newborn). Upon recording and saving completion, the Cancel button becomes a highlighted Done button while the Record button grays out into a subtler Re-Record button, to prompt the practitioner to return to the screen showed in Figure 3.7c.

Submitting a sample uploads the entered information to a password-protected server, as well as saving a local copy

FIGURE 3.7: The views (B) to record a video sample, (A) immediately preceding it, and when submitting a sample, both (C) in progress and (D) upon completion. A drawing substitutes the newborn to maintain privacy for newborns.

to the phone in case something goes awry in the uploading process (e.g., internet connection breaking). Views during the submission process (Figure 3.7d) entail all the entered information, to let someone confirm the sampled data and notice any details that require correction. If submission was unsuccessful, the system continues trying for a short duration. Continued unsuccessful attempts ends the process and results in a short message to indicate that uploading was unsuccessful, but reassures that it's alright because a local copy was saved to the phone.

3.4.2 Study Results

Regression results using data from the pilot study (i.e., preliminary results due to the small sample size) is shown in Figure 3.8. It yields a Pearson correlation coefficient of 0.81 between the estimated and measured bilirubin levels. The mean residual error of this fit is 1.6 mg/dL, with 90% of the estimated bilirubin levels falling within 2.6 mg/dL of those measured by TcB or TSB.

Given that TcB values have an inherent variation with TSB values, the predicted bilirubin levels may also contain added variation from using an overwhelming amount of TcB measures. By replacing TcB values with TSB, the model has the potential to correlate more closely with the TSB ground truth.

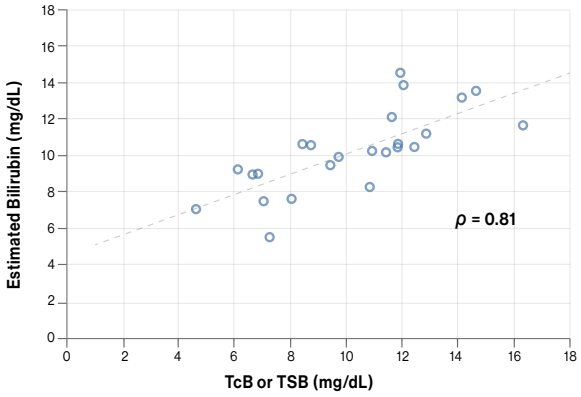


FIGURE 3.8: A scatterplot comparing measured bilirubin levels (TcB or TSB depending on participant) against regression predictions.

These results are promising; they offer preliminary evidence that signals or correlations exist between data that BiliCam can capture and a newborn's bilirubin level. These results justify conducting a larger study, with which to capture more data and better evaluate BiliCam.

3.4.3 Insights for Next Study

Observations from the pilot study inspired improvements for the next version of the data collection application.

3.4.3.1 Fixing Samples Post-Submission

One improvement is to enable fixing samples post-submission. In the pilot version of the app, the system rendered any submitted samples completely inaccessible to anyone but researchers on the technical end accessing the password-protected server or the contents of the phone with a tether and the developing environment with the appropriate provisioning profiles. However, not all data-entry is perfect (both from people, who can make mistakes, and software, which embarrassingly enough, can crash). During the pilot, practitioners occasionally emailed corrections. An improved application could allow practitioners to access recently recorded samples to update them accordingly. These samples can be accessible for a limited time, so as to keep the newborn's images private.

3.4.3.2 Clearer User Interface

Another improvement was to make a clearer user interface. The fact that entered data can be edited was not immediately apparent to one of the collaborating practitioners. A later design would make this edit-ability more obvious.

3.4.3.3 Better Privacy

A later version of the app could maintain better privacy. Recording a newborn's age doesn't require recording his or her time of birth and the time of the sample. These two pieces of information can let someone figure out who a sample refers to. Hence, the next version of this application would either ask only for a newborn's age (in hours), or compute this age and record only that.

3.4.3.4 Automatically Detecting Issues

A major improvement is automatically detecting issues in image quality. Initial data collection involved in-person feedback on the image quality, bringing specific undesired conditions to attention. Automatically detecting these conditions and prompting the photographer would improve image quality for scaled data collection, during which constant, direct feedback with collaborators is not as easy. Some of these conditions include underexposure, overexposure, occlusions, overhead or uneven shadows, glare on the card, and the card losing a good position (e.g., falling off or getting bent).

3.4.4 Contextual Inquiry

Contextual inquiry at the UW Medical Center with one of the pediatricians during data collection raised a number of insights. In addition to a richer understanding of the context of this work, the experience pin-pointed some potential areas for data collection improvement.

The process for gaining consent from parents took much more time than the physical data collection itself. Introductions can take a while, as the practitioner needs to not only explain the project but also introduce him/herself and at times chat for a bit to help parents feel more at ease. A stranger walking into the room asking to run experiments on their newborn would otherwise seem much more intimidating than it should. The timing of these introductions and requests for consent also makes the situation trickier. Mothers, having just given birth, are exhausted. Hence, practitioners need to find an appropriate time to come in and talk, as these mothers could be taking some much-needed

rest and not be ready for communication (e.g., sleeping). Additionally, their exhaustion can make the process for explanations go a bit slower.

A proposed improvement was to ask for consent a good while before mothers give birth, during which they could be more alert and ready to listen. Additionally, recruiting their pediatricians to explain the study could help save overall time spent on introductions and talking to make them feel comfortable with the researchers.

Overall, most parents appreciated and were supportive of this research project. The kinds of concerns they expressed during the contextual inquiry were more focused on waking up sleeping newborns or making them fussy when exposing skin for the sample.

3.4.5 Value Sensitive Design

The value sensitive design I conducted focused on the project's ultimate goal: a smartphone-based solution for parents. The identified primary stakeholders were parents and their newborns. Indirect stakeholders include doctors, TSB drawing nurses, visiting nurses, hospital management, companies that produce medical equipment for TcB or TSB tests, and practitioners in a low-resource setting. A subset is detailed below.

3.4.5.1 Parents

Assuming the system works perfectly, it can help provide parents peace of mind. One parent, who had his first child two years ago, expressed how worried he felt when judging the yellowness of his son's skin during his first few days. He continually questioned himself about his concerns, wondering, "Is this new-parent syndrome or not?" making him uncertain. In the end, it turned out that his son did have abnormally high bilirubin levels and required treatment. Being able to check with his phone would have made his decision to visit the hospital easier and less worrisome.

Additionally, a perfect system could save parents grief, money, and time if it catches a case early. Cases not treated quickly enough could lead to extreme treatments like blood transfusion or leave permanent damage.

However, if the system does not work perfectly, it can have some very adverse effects. The worst-case is having a false negative, which could lead to an untreated newborn and all of its implications. Too many false positives can also be a concern, as returning to the hospital for further testing could waste money, time, and effort. It would make one of the main purposes of the system useless, as the result would not be too different from the current status quo. Additionally, too many false positives might heighten fear for parents thinking about what horrible things could happen to their newborn because the system warned them to visit the hospital.

Another potential harm is that parents could start developing a false sense of security for all medically-related phone applications if this system worked for them and leaves them with good impressions.

3.4.5.2 Newborns

A perfectly working system would benefit a newborn's health and number of pin pricks from medical exams. However, too many false positives could lead to more pin pricks than necessary. False negatives could lead to the terrible consequences described above. Another potential harm is a privacy leak of their personal information or photos, thus the system must be well-designed to protect that.

3.4.5.3 Doctors

With a properly working system, the doctors could have a potentially reduced mental workload. For example, they could receive easier or more objective feedback from families who return for another test. It could potentially reduce the number of patient check-ups. Catching dangerous levels of bilirubin early could also help keep treatment cleaner and cheaper (e.g., avoid the dangerous and expensive process of blood transfusion).

Potential harms could come from system inaccuracies, as described for both direct stakeholders above. They could also have their reputations hurt if they recommended a system that did more harm than good.

3.4.5.4 Nurses

The nurses operating the medical tests for bilirubin could have lessened work regarding bilirubin on a much larger scale than for doctors. If the system works well, it could mean fewer necessary TcB measurements, blood-draws, and TSB processing on the blood. Some hospitals instate visiting nurses to travel from one home to another for the sole purpose of taking these bilirubin measurements 3-4 days into a newborn's life (they measure other aspects too, but mostly just because they are already there). A perfectly working application could dramatically reduce the need for such visits. The lessened bilirubin-related work could open time for other medical needs.

An inaccurate system, however, could reduce their peace of mind, in addition to the aforementioned issues. One nurse explained how even an inaccurate TcB measurement has kept her up at night, out of concern for the newborn.

3.5 Formal Local Study

From the promising results of the pilot study, I were able to justify, finance, and prepare for a formal study with a more complete dataset. I partnered with two local medical centers to collect data from 100 newborn participants, and yielded a 0.85 rank order correlation with the gold standard blood test (TSB) in the evaluations. Unlike in the pilot, this dataset included both TSB and TcB measures for each newborn as well as baseline photos for each newborn. There were other key differences from the pilot, including revisions to the data collection app. I published the results of this study in UbiComp 2014 [51].

3.5.1 Data Collection

To evaluate and inform the design of BiliCam, I conducted a clinical study at two sites in Seattle, the University of Washington Medical Center (UWMC) and the Roosevelt Pediatric Care Center, to create a dataset of image samples paired with ground-truth bilirubin levels from TSB tests. I collected images within two hours of the TSB blood draw to ensure that bilirubin measures were as accurate as possible.

3.5.1.1 Enrollment

Parents of newborns born at the UWMC gave informed consent to participate in the study within 24 hours after delivery. Photo samples were taken within these first 24 hours of life as a baseline and once more between 2.5 to 5.5 days of life for a follow-up. I limited enrollment to English-speaking parents of newborns who were born at more than 35 weeks of gestation (i.e., full term newborns). Of the 134 newborn participants who opted into the study, a total of 100 completed the study. Participants who required phototherapy prior to the follow-up became ineligible, due to the effect of phototherapy on skin color, which is a known issue for the TcB [121]. I also noted which blood samples were effected by hemolysis, a condition that affects the accuracy of TSB readings [26].

Medical professionals collected all the images on iPhone 4S smartphones using a custom data collection app and the built-in camera. I chose to use an iPhone because it has the most standardized hardware of the current smartphone platforms available. The design of this study was informed by a pilot study I ran with 40 newborn participants. The pilot study data is not included in the evaluation of BiliCam due to significant differences in study procedure.

Participant Demographics (N=100)	
Age at follow-up (hours) (mean, range)	86 (60 – 129)
Bilirubin Levels (mg/dL) (mean, range)	9.9 (0.8 – 21.1)
Hemolysis (n, %)	19 (19%)
Reported Ethnicity (n, %)	
American Indian/Alaska Native:	6 (6%)
African American/Black:	15 (15%)
Asian:	20 (20%)
Latino:	9 (9%)
Pacific Islander/Native Hawaiian:	3 (3%)
White:	79 (79%)
Other:	2 (2%)
Multiple Races:	24 (24%)

Table 3.1: Demographic information for participants. Note that participants may report multiple ethnicities.

3.5.1.2 Data Collection Timeline

I structured the study to consist of two sets of image samples per newborn: a baseline and a follow-up. The baseline was taken at the UWMC within the first 24 hours of life, during which the newborn's bilirubin is typically very low. The follow-up was taken at either study site when the newborn was 2.5 to 5.5 days old. Within two hours of the follow-up image, two medical bilirubin measurements were taken: a TSB blood sample and a TcB. The TSB provided ground-truth data and the TcB as a source of comparison. The TcB measures came from a Philips BiliCheck or Draeger Jaundice meter JM-103. I assigned a unique study ID to each participant to match the participant's medical results and image data while maintaining confidentiality.

After receiving the samples from the study phone, I segmented each image to extract the pixel values of the sternum, forehead, and color patches on the color calibration card. The sternum and forehead are the primary locations of interest for skin samples for several reasons. Medical practices standardize TcBs, which are also light- based, to take readings from these two locations. Both the forehead and sternum also offer prominent, flat regions of skin on which I expect even lighting. I expect the whites of a newborn's eyes to be more consistent across skin tones and potentially a better location of interest for a visual system. However, the eyes are closed (e.g., while sleeping or crying) the majority of the time. Even when open, the whites are hard to discern, given their small size compared to the iris.

3.5.1.3 Color Calibration Card

I designed the color calibration card to take the form of a business card for easy manufacturing and its appropriate size on newborns. The card has eight square patches with the following colors: black, 50% gray, white, cyan, magenta,

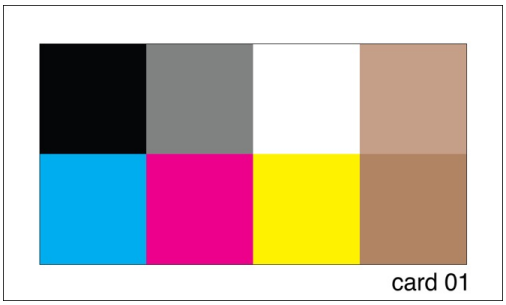


FIGURE 3.9: Design of the color calibration card used in BiliCam's local studies.

yellow, and two skin tones (see Figure 3.9). The cyan, magenta, and yellow were printed such that only the respective inks were used on each patch; the colors did not contaminate each other. The cards were printed by a Konica Minolta Bizhub PRO c6501 printer on Cougar 100lb uncoated paper. I used a new card for each sample to prevent spreading disease. When taking a sample, I affixed a small, skin-safe adhesive to the back of each card for stable placement just above the newborn's navel without bending or occluding the card from the user's fingers.

3.5.1.4 Data Collection Application

Figure 3.10 shows screenshots from the custom iPhone application that medical professionals used for data collection. For each sample, the app records the study ID, time of birth, whether the sample is a baseline or follow-up (i.e., a sample type), and one or more sets of photographs and videos (Figure 3.10a). For taking photographs, it first instructs the placement of the color calibration card on the newborn (Figure 3.10b) and prompts the user to make sure there is a clear view of the card, sternum, and forehead. The phone then provides a live view from the camera with an overlaid "view finder" to align with the calibration card (Figure 3.10c). These cues constrain the distance of the camera from the newborn and comfortably fit the card and newborn's sternum within the image. The data collection application then captures a set of images. In case BiliCam needs to account for brief changes in skin color, such as subtle flushes from blood flow, it includes a 10-second video sequence. The phone's "flash" LED is on during the first 5 seconds and turns off for the last. The system also takes a high-resolution photograph in the middle of each 5-second segment, capturing one image "with flash" and one image "without flash."

The system first analyzes the captured images to check for sample quality. It detects problems with the images such as positioning issues, occlusions, or inconsistent lighting, by applying a threshold on the standard deviation of pixel values for each color patch on the card. It then displays the captured images and recommends retaking them if any

(A) (B) (C)

FIGURE 3.10: Screenshots of data collection application with (A) the view for entering basic sample information, (B) instructions prior to recording images, and (C) a live feed from the camera with a "view finder" to align the color calibration card.

problems arose, additionally listing possible reasons the image failed the quality test. Upon submitting a completed sample, the system uploads the sample data to a server through the phone's Internet connection. It also stores a local copy on the phone as a backup.

While much of this data collection app is the same as in the pilot study, a major improvement came from the ability to automatically check images for sample quality and recommend retaking samples when necessary. After the app capture images and records its 10-second video, an image segmentation algorithm (described in subsubsection 3.5.2.1) extracts patches from the card from two representative still images. If the standard deviation exceeds a threshold or any of the captured values are at the minimum or maximum possible values (which would suggest clipping), the app then prompts the user to take the photos again. People can still submit samples without retaking photos (as has been necessary in some circumstances).

Other changes improved the app's workflow based on observations and feedback from app usage in the pilot study. The app no longer asks people to enter the participant's TcB reading; the data collectors already recorded TcB levels separately and entering them on the phone turned out to be cumbersome. It interrupted interactions with the participant and their family, so some data collectors waited until after they left the family to enter the TcB – keeping the app open to the unsubmitted sample all the while, risking the loss of that sample. Furthermore, the TcBs only

applied to follow-up photos and hence did not even apply for half of the samples.

Recording the sample type was also a new addition, as baseline photos became mandatory and incorrectly entered dates of birth – which happened often enough – can make it difficult to infer the sample type correctly. This addition improved the workflow for analysis on submitted data.

I also changed the ordering of when the app automatically turns the phone's flash LED on or off for each corresponding BiliCam photo. I reordered the photos to start with those that had the LED turned on. That way, when the app initializes the camera, it also turns on the LED, which enables people to align the light of the LED over the newborn at the same time as the camera without creating a sudden and startling flash. People can also maneuver the phone to avoid shining the LED directly into the newborn's eyes as they bring it over the child.

3.5.2 Segmentation and Extraction

I hypothesize that the visual characteristics of a newborn's skin can estimate his or her bilirubin levels. Considering that the collected images (and thereby any extracted features) could vary considerably with different lighting conditions, the images need to be color balanced before feature extraction. The main goals, then, are (1) to color balance the images, (2) extract intensities of various reflected wavelengths and other chromatic and achromatic properties from the skin, and (3) estimate bilirubin levels using machine learning. I explain each stage in turn and show a method outline in Figure 3.12. As an overview:

- Image segmentation isolates the colors on the card, sternum, and skin.
- Color balancing is carried out using the calibration cards captured in each image.
- For each skin patch, I estimate the mean red, green, and blue values, and the gradients of colors in. the patch I employ various color transformations to approximate properties such as hue, gamma, and saturation.
- Extracted properties are used as features in a stacked regression and classification algorithm, which results in a final estimate of the bilirubin value (see section 3.5.3).

3.5.2.1 Image Segmentation

In order to use the card, the system first needs to identify the location of the card and each of its color patches. Although full automation and segmentation is not the focus of this proof of concept for BiliCam, I developed an algorithm to segment the card which I used in automatic image quality feedback when collecting data. I segmented skin patches by hand to reduce confounds.

The data collection UI constrained the card to a specific region on the image (as shown in Figure 3.10c). Hence, the algorithm can ignore the pixels outside of this region to reduce the search space. It then locates at least two color patches on the card and extrapolates the rest of the card from these patches.

To identify the color patches, the algorithm applies thresholds to the image. The system takes advantage of the fact that the cyan, magenta, and yellow patches have very distinct hues and high saturation. Hence, it converts the image to the hue, saturation, and value (HSV) space and applies empirically determined thresholds on the hue and saturation channels. Performing a bit-wise 'AND' operation of the two thresholded images separates the patch from the rest of the image. Figure 3.11 shows an example of thresholding for a yellow patch in this manner. Because that the system is aware of the approximate size of each patch, it can differentiate the patch from further noise in the image. This is done by using edge detection and morphological operations; specifically, I use an opening operation and Canny edge detection. The algorithm then uses contour- detection to identify the patch's boundary from the detected edges and smooths them using the Douglas-Peuker algorithm [102]. After the system finds two of the color patches, it calculates the orientation of the card. It then extrapolates the locations of the remaining patches from these found corners.

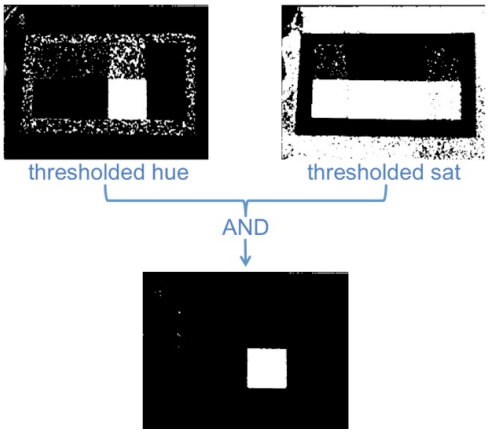


FIGURE 3.11: Segmentation of the yellow patch from the color calibration card. Thresholded versions of a BiliCam card image by hue and by saturation can combine via a bit-wise 'AND' operation to isolate a yellow patch.

3.5.2.2 White Balancing

I derive the features from the observed skin color, which can vary in different lighting conditions. To mitigate some of the effects of different lighting, I compute normalized red, green, and blue values. I calculate these normalized

values by dividing each color channel value by the sum of all three channel values. Normalization alone, however, is not sufficient to counter color variations of illumination sources (i.e., the differences in halogen, fluorescent, or incandescent bulbs that can cause images to seem more "yellow" or "warm"). Hence, I include the color calibration card in each image for further color balancing.

I experimented with a number of white balancing techniques and most effective one to be an algorithm used by many popular image-editing tools. It uses the observed red, green, and blue (RGB) values of the white color patch to adjust the RGB values of the skin. More precisely, given (R', G', B'), it computes the adjusted (R, G, B) by

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 255/R'_w & 0 & 0 \\ 0 & 255/G'_w & 0 \\ 0 & 0 & 255/B'_w \end{bmatrix} \begin{bmatrix} R' \\ G' \\ B' \end{bmatrix}$$

where (R'_w, G'_w, B'_w) is the average observed color of the white patch on the color calibration card [138].

3.5.2.3 Feature Extraction

Elevated levels of bilirubin result in a yellow discoloration of the skin. In order to better detect this subtle discoloration, BiliCam transforms the original RGB values into the YCbCr and Lab color spaces. I calculate their mean values for each color channel, resulting in 9 features.

In addition to color transformations, I also calculate the change in color across the image patch using a linear color gradient. The gradient is calculated by running a 3×3 Sobel gradient filter across each color channel, and then averaging the outputs inside the patch. This is performed in the R, G, and B color planes, resulting in 3 additional features.

The data collection app for BiliCam captures 2 images in each test: "with-flash" and "without-flash," as described earlier. I use mean color features from both images and the color gradient features from the "with-flash" images, resulting in a total of 9+9+3=21 features. These features are used to train a custom machine learning regression algorithm with leave-one-out cross validation. For each fold, the training set features are transformed to have unit variance and zero mean (scaling). I also use principal components analysis (PCA) to decrease redundancy (e.g., the redundancy between YCbCr and Lab color spaces) and reduce the dimensionality to six component features. It learns these transformations only from the training dataset.

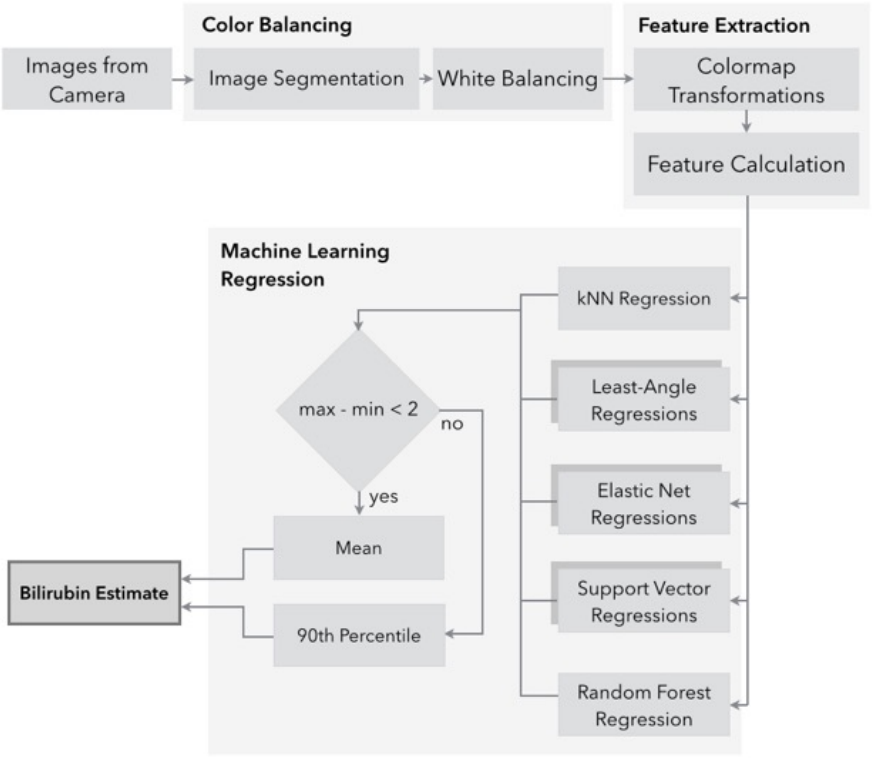


FIGURE 3.12: Flowchart of the algorithm.

3.5.3 Machine Learning Regression

The regression algorithm employs an ensemble of different regressions. Each regression is chosen to give a slightly different perspective of the feature data. First, the scaled or PCA transformed features are used in each regression to obtain separate estimates of the total bilirubin level. Then the outputs of each regression are combined based upon the agreement in the ensemble, resulting in a single value for the bilirubin level. Figure 3.12 shows a flowchart of the machine learning process.

BiliCam used an ensemble of five different regression algorithms. Each regression is discussed in turn. Most regressions are carried out using the scikit-learn toolkit [98] in Python. In order to avoid overfitting, this evaluation of BiliCam used leave-one-out cross validation in all levels of learning. That is, no images from the training sets are used in the testing sets for any of the regressions.

3.5.3.1 k-Nearest Neighbor

The first regression algorithm is an encapsulated k-Nearest Neighbor regression (k = 7) [53]. Intuitively, this regression takes a more "local" estimate of the bilirubin level based upon training points that have similar feature values. In this regression, I have a database of known features and bilirubin values. When an unknown test vector is analyzed, the k-nearest neighbors are found around the test vector in the database of features. The features for finding the nearest neighbors are the first two components of the PCA transformation. I use the L1 norm to calculate the nearest neighbors. Feature points from the neighbors are used to train a linear support vector regression. A new regression is built each time a new test point is analyzed.

3.5.3.2 LARS

The second set of regressions uses least angle regression (LARS) [40]. LARS regression uses a variant of forward feature selection to decide what features are most useful. Intuitively, this regression helps eliminate redundant features, while creating new features based on their correlation to the chosen features. Essentially, the best predictor from the feature set is chosen by developing a single-feature, linear regression from each feature. The most correlated output is chosen as the "first" feature. Then, the algorithm attempts to find another feature with roughly the same correlation to this prediction's residuals as the first feature to the output. It then finds the "equiangular" direction between the two estimates, and finds a third feature that maximizes correlation to the new residuals along the equiangular direction. Features are added in this way until the desired accuracy is met. To experiment with transformations, I train two LARS regressions: one with scaled features and another with PCA components.

3.5.3.3 LARS-Lasso Elastic Net

The third regression uses the elastic net algorithm [147]. Intuitively, this algorithm also eliminates features, but in a slightly different way than LARS. This regression is a combination of Lasso regression (highly related to LARS for forward feature selection) and ridge regression (which uses an L2 regularization). In this way, forward feature selection and the L1 and L2 norms are employed in the regression objective function. This makes it related to LARS and Lasso regression, but with certain "backoff" regularization so that it becomes more stable. The parameters are chosen based on a grid search of the training set (but never the test set). As with LARS, I train two regressions, one with scaled features and another with PCA components.

3.5.3.4 SVR

All the regressions up to this stage were linear regressions. In order to capture the possible non-linear relationship, I employ two support vector regressions [122]. The idea behind the support vector regression (SVR) is that a linear regression function can be found in a high dimensional feature space. Then, the input data can be mapped into the space using a potentially nonlinear function. I train two SVRs: the first uses a linear kernel and the second uses a nonlinear sigmoidal basis function.

3.5.3.5 Random Forest

The last algorithm uses random forest regression [24] with 75 trees. A random forest is a collection of estimators. It uses many "classifying" decision trees on various sub- samples of the dataset. The outputs of these trees are averaged to improve the predictive accuracy and control over-fitting. Each tree is created using a random sub-sample (with replacement). Intuitively, the random forest regression can learn nonlinear or complex relationships in the data, which may be different from the regressions discussed up to this point. The random forest uses scaled features only.

3.5.3.6 Final Output

There are a total of eight regressions trained from the five algorithms. The agreement between the ensemble for a given test value is assessed from the difference between the minimum and maximum values from the ensemble. If the difference is less than the empirically derived threshold of 2.0 mg/dL, the ensemble "agrees" and the mean is chosen. If the difference is greater than 2.0 mg/dL, then the second highest bilirubin value (i.e., the 90th percentile) is chosen. This helps to bias the regression algorithm to selecting a large bilirubin value when the ensemble does not agree; when used as a screening tool, it is more acceptable to have a false positive than to "miss" a potentially high bilirubin.

3.5.4 Results

I break down the results into two subsections. The first section, *Predicting Bilirubin Levels*, quantifies the performance of the machine learning regression. The second subsection, *Predicting Newborn Risk*, quantifies the effects of BiliCam as a screening tool.

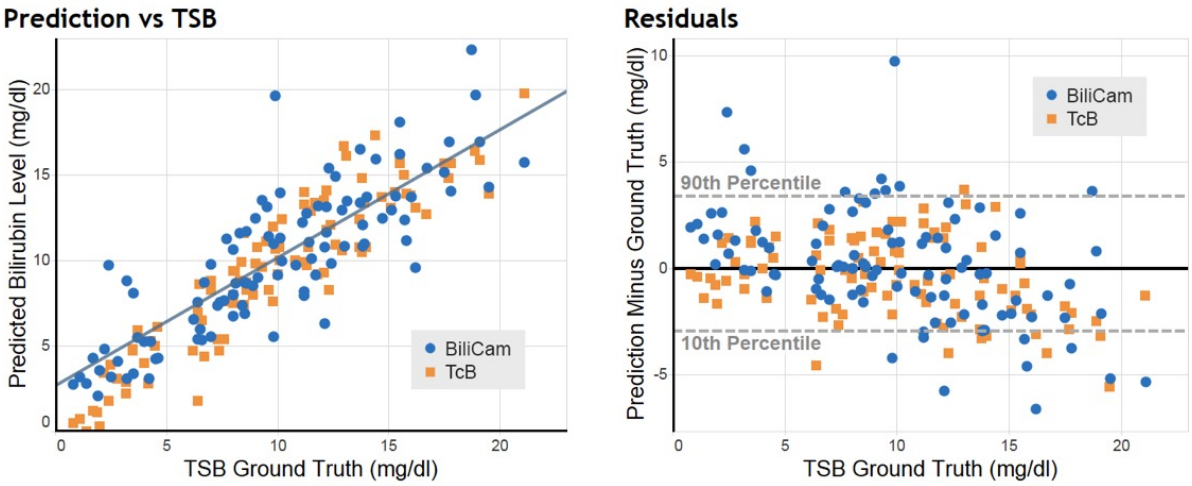


FIGURE 3.13: Left: a comparison between predicted bilirubin levels and TSB. Blue circles represent predictions from BiliCam, orange squares predictions from the TcB. Right: a plot showing residuals of predicted bilirubin levels against the TSB.

3.5.4.1 Predicting Bilirubin Levels

The individual regression algorithms performed similarly in terms of correlation with the TSB (rank order correlations ranged from 0.82 to 0.85). However, a closer inspection reveals that the algorithms perform quite differently on individual samples. The linear methods, in particular, tend to under-report bilirubin levels when the values are above 12 mg/dL despite their high correlations. The ensemble method includes non-linear methods and can improve the overall accuracy of the system. Therefore, I only focus on the performance of this ensemble for the rest of the chapter.

Explanation: Figure 3.13 shows a scatter plot of BiliCam estimates (blue circles), calculated through leave-one-out cross- validation, compared to the TSB. It also shows a chart where residuals (BiliCam – TSB) are plotted against the TSB. For comparison, each plot also contains predictions from the TcB (orange squares). **Results:** The predicted bilirubin levels correlate with the TSB by a rank order correlation of 0.85 (linear correlation of 0.84), with a mean error of 2.0 mg/dL. I also compared the results from the TcB with the TSB and found a rank order correlation of 0.92 (linear correlation of 0.92) and a mean error of 1.5 mg/dL. Note that all results from the TcB used one fewer data point because the TcB would not provide a reading for one participant. **Implication:** Under the constraints of the study, BiliCam is effective at estimating the bilirubin levels and compares favorably with the TcB.

A Wilcoxon signed rank test and an F-test of the residual variances failed to show statistically significant differences between the BiliCam estimates, the TcB estimates, and the TSB estimate ($p>0.05$). An N-way ANOVA on the residual

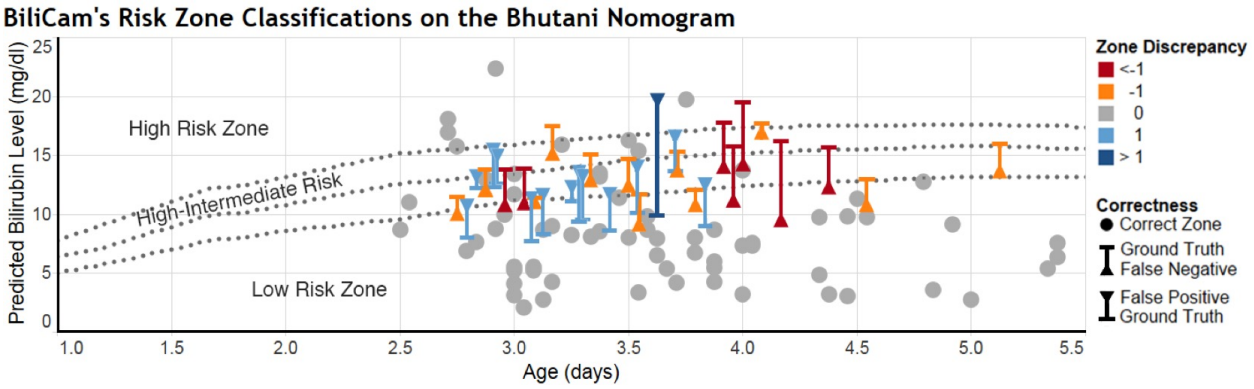


FIGURE 3.14: Risk zone classification of BiliCam using Bhutani nomogram. Colored lines map incorrectly classified points to the ground truth classification and bilirubin level.

magnitude (|BiliCam-TSB|) also did not reveal statistically significant effects on the residual magnitude from race, age, and hemolysis ($p<0.05$).

Note that all presented results thus far use only features from the follow-up sessions. They do not include features from the baseline images for each newborn. Including these features failed to show a statistically significant difference in means based on a two-tailed t-test ($p=0.05$). For comparison, including the baseline features yields a rank order correlation of 0.83 (linear correlation of 0.82) and a similar mean error of 2.2 mg/dL.

I note that BiliCam has more outliers than TcB, the top five of which come from non-white participants. However, there are no consistent attributes (such as race, hemolysis, or observable image quality) for why all the outliers exist. I need more data to characterize the existence of outliers.

3.5.4.2 Predicting Newborn Risk

To evaluate how well BiliCam assesses a newborn's risk from jaundice, I applied classifications from the Bhutani nomogram to the predicted bilirubin levels. The nomogram divides bilirubin samples into four risk zones: *low*, *intermediate-low*, *intermediate-high*, and *high* [14]. The classifications are based on the predicted bilirubin levels and participant age at the time of sample, using risk zone boundaries defined by BiliTool™ [18, 75].

Explanation: Figure 3.14 shows the results of plotting the bilirubin levels predicted by BiliCam against age over a Bhutani nomogram. Where the points fall with respect to the risk zone boundaries determines their classification. Colors and directed symbols encode incorrect classifications based on classifications from the corresponding TSB (see Figure 3.14's legend). Gray circles denote correct risk zone classification. **Results:** 67% of the results exactly match the

Bhutani classification from TSB, 19% are false negatives, and 14% are false positives. Of these misclassifications, 76% were off by one zone (i.e., misclassified into an adjacent zone). For comparison, the TcB yields 68% exact matches, 22% false negatives, and 9% false positives. 87% of the TcB's misclassifications were off by one zone. A suggested method for screening with the TcB is to administer a TSB to catch high risk cases if TcB readings fall into the *intermediate-high* or *high risk* categories [79]. To compare the effectiveness of BiliCam as such a screening tool, I consider the *high risk* cases classified into this combined risk category. Of the 9 samples that should classify as *high risk*, BiliCam classified 2 false negatives (hence, missing 2/9 or 22% of the *high risk* cases) and 8 false positives. In comparison, the TcB classified 2 false negatives (missing 2/8 or 25%) and 5 false positives. Note that there is one fewer TcB measure because the device would not offer the corresponding reading.

Implication: BiliCam demonstrated statistically equivalent performance as the TcB in its ability to catch high risk cases in the dataset. These results indicate that BiliCam could have a very similar utility to TcB as a screening tool, with the advantage of greater accessibility.

3.5.5 Commentary

Our analyses of BiliCam compare favorably with the TSB and the TcB. BiliCam cannot replace TSB testing, but can be used like the TcB as an effective screening tool to determine whether TSB testing is necessary. The ubiquity, portability, and cost of smartphones also offer advantages that may make BiliCam more appropriate for screening in home environments where TcBs are not available. Despite these advantages, there are limitations that require more research before I can fully characterize BiliCam as an in-home jaundice screening tool.

3.5.5.1 Limitations of Local Study

Data collection from the local studies were done solely on iPhone 4S devices. To be more accessible, BiliCam should function on multiple devices and platforms. Different brands and models employ different cameras, lenses, filters, and color corrections. All these factors can affect the collected data. I have yet to investigate the feasibility and necessary adjustments to make the system available on other devices. The color calibration cards are another unexplored variable. Every card I used in the study was printed at the same shop using the same printer and paper. The level of variation for ink, printers, and paper permissible for accurate results is yet to be tested.

To be practical on a global scale, BiliCam results must also address diverse populations. Hence, data needs to be collected from a large variety of participants of different races. The diversity of the dataset is inherently limited as

more than half of its participants are white. BiliCam would also benefit from more data to help characterize its outliers. This need was a primary driver for conducting the next, nation-wide study.

Additionally, the quality of images has a major bearing on the output of BiliCam. I often captured multiple, back-to-back sets of images per sample-taking session to compensate for varying degrees of image quality (i.e., in case of blur, occlusions, graininess, etc.). Randomly drawing from these images can drop the rank order correlation with TSB to as low as 0.80 in the dataset.

3.5.5.2 Samples and Feature Selection

Our final algorithm did not include the baseline photos because it failed to demonstrate a statistically significant difference; using features from both baselines and follow-up samples is as equally decisive as using those from follow-up samples alone. Not needing these baselines is a major benefit — it means that BiliCam can predict newborn bilirubin levels from a single session. However, baselines are still a worthwhile option to explore as they may help to adjust for skin tone differences in more diverse populations.

Although I segmented images for patches of skin on both the forehead and sternum, I ultimately focused solely on the sternum. There are several possible explanations for why the sternum yielded better results. I expect inconsistent lighting to be the primary reason. With the forehead being much further away than the sternum from the color calibration card, compounded by the head's large range of motion, the skin on the forehead experienced different lighting conditions than the card and sternum. Additionally, the sternum is a preferable location for other reasons: sunlight can mildly reduce bilirubin concentration in the skin and the sternum tends to experience less light exposure than the forehead. Studies suggest the same effect for TcB readings, which are also optically based, and explicitly recommend taking TcB measurements from the sternum over the forehead [101].

Given the quality of the initial results from using just the still images, I focused on those and have not investigated the videos in the dataset. Still images are also preferable from a logistical point of view: they can be uploaded and processed by a server in a matter of seconds, enabling BiliCam to offer instant results. Videos are also difficult to take and more susceptible to image quality issues, which I discuss further in the next steps.

3.5.5.3 Next Steps

By the end of this formal local study, I decided that the next steps would focus primarily on further data collection to reduce the system's limitations. In addition to acquiring more data points and increasing the diversity of the samples,

I also came up with the following ideas for improvements.

Redesigning the color calibration card could facilitate smoother data collection. A hole in the center of the card can frame the skin patch of interest. It would force the skin and card to lie immediately next to and flush with each other for more consistent lighting, whereas the current system lets the angle of the skin and card vary freely. Constraining the skin to lie within this hole would also make automating the segmentation process much more straightforward. Currently, a number of complications like unexpected shadows, occlusions, and body positions make automated skin segmentation non-trivial to the point where I prefer to segment them by hand. The card could additionally benefit from having a peel-off back that exposes a gentle, skin-safe adhesive.

The data collectors expressed difficulties in taking images of the newborns that I would like to alleviate in future data collection. Positioning and holding the phone at the right distance, watching the newborn, aiming for the card, and reacting to the newborn's movements for a 10-second video can be surprisingly overwhelming, particularly if the newborn is crying. Given how promising the results are using still photographs, one improvement could be to take a series of clearly punctuated still images instead of a video. I expect that taking these images is significantly easier as it does not require continual tracking.

The image quality feedback mechanism also has room for improvement. The current system only lists possible reasons that an image can fail the quality test, so the reason for a particular failure is not obvious. A future system could automatically determine and report the source of image quality issues (e.g., highlight instances of glare or shadows). It can also improve by checking images in real time, to alert the photographer to potential issues before and throughout the sample collection process, or automatically recognize and capture images with passing quality.

Further into the future, BiliCam could benefit from having both server-connected and stand-alone versions. There are interesting trade-offs between running BiliCam' s algorithm directly on the phone versus on a server. Computation on the server can retain tighter control of how the system runs the algorithm, based on a growing central database of clinical samples to train on or algorithmic breakthroughs. It can also guarantee that BiliCam uses the most up-to-date algorithm. However, computing entirely on the phone offers the ability to use BiliCam without any Internet connection. A stand-alone version may be the way to disseminate this medical system in low resource settings with incomplete or inadequate cell coverage. I believe a server-connected version is otherwise preferred.

3.6 Nation-Wide Study

After the formal local study demonstrated a proof of concept for BiliCam, I decided to evaluate BiliCam on a larger population to encompass more data points for verification and a wider demographic range. Based on insights from the early studies and additional analysis I conducted of the entire study procedure, I also improved the process and focused on collecting more targeted and higher quality data.

We acted on these goals through conducting a nation-wide data collection study with seven clinics around the United States. The study involved 600 newborn participants, 552 of which presented usable data for the final model. A subset of this work was published in the Official Journal of the American Academy of Pediatrics in 2017 [132].

3.6.1 Study Design

This nation-wide study incorporated a number of important changes. These changes to the study design made data from this study incompatible with data from previous studies.

3.6.1.1 Study Procedures and Enrollment

Participants were recruited for the study at 7 sites across the United States, including the University of Washington Medical Center (UWMC) in Seattle, Washington; Thomas Jefferson University Hospital (TJUH) in Philadelphia, Pennsylvania; Seattle Children's Hospital in Seattle, Washington; Kaiser Permanente San Leandro Medical Center (KPSL) in San Leandro, California; Maricopa Integrated Health System (MIHS) in Pheonix, Arizona; McKay-Dee Hospital (MKD) in Ogden, Utah; and Truman Medical Center (TMC) in Kansas City, Missouri. Study participants were enrolled between October 2014 and July 2016.

A variety of enrollment procedures were used. For all study sites, participants were healthy, newborn infants less than $<$7 days old who were born at $\geq$35 weeks' gestation. Neonates who had received phototherapy were ineligible. At UWMC and TJUH, newborns were enrolled when they were $\leq$24 hours old, with a follow-up study visit when they were 3 to 5 days old. At the follow-up visit, a set of BiliCam images was obtained, and blood was drawn for a TSB level. At KPSL and MIHS, BiliCam images were obtained from participants at the time of a blood draw for a TSB level that was ordered because a newborn was clinically jaundiced. Finally, newborns at MKD and TMC were enrolled and BiliCam images were obtained when TSB levels were measured as part of routine screening or if a neonate was clinically jaundiced. Attempts were made to obtain BiliCam images within 2 hours of the blood draw for TSB determination. Blood samples were assayed by the clinical laboratory at each site. The TSB assays at each of the participating sites were

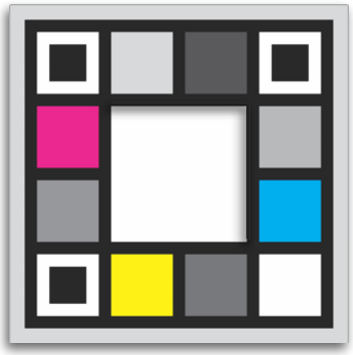


FIGURE 3.15: Design of the color calibration card used in BiliCam's nation-wide study. Note that the central or largest patch on the card is actually a cut-out to show the newborn's skin.

run by using the following platforms: at UWMC, Beckman AU680 Total Bilirubin; at TJUH, Roche Cobas 501 Bilirubin Total; at Seattle Children's Hospital, Ortho Vitros 4600 BuBc; at KPSL, Beckman AU680 Total Bilirubin; at MIHS, Ortho Vitros 5600 BuBc, at MKD, Abbott Architects c8000 or c4000 Total Bilirubin; and at TMC, Roche Cobas Bilirubin Total.

At all sites, parents of participants were asked to provide the race of their infants at the time of enrollment; data on participants' birth dates and times were abstracted from medical records. At UWMC, MIHS, and TJUH, TcB measurements were done for almost all participants at the time of the blood draw for TSB; TcB measurements were obtained for selected newborns at TMC. Both BiliChek (Philips Respironics, Monroeville, PA) and the Draeger Jaundice Meter JM-103 (Draeger Inc, Telford, PA) brands of TcB meters were used for TcB measurements. When possible in study recruitment, samples also included baseline photos (as described in subsubsection 3.3.1.2) but they were not a requirement for study participation, as baseline photos did not improve the model from the formal local study (see subsubsection 3.3.1.2) and would often be a barrier for recruitment in a study that prioritized collecting data from a large population. The study was approved by the institutional review boards at each participating institution, and written, informed consent was obtained from the parents of study newborns.

3.6.1.2 Color Card

We redesigned the color calibration card based on insights described in subsubsection 3.5.5.3. A hole in the center of the card frames the skin patch of interest. It forces the skin and card to lie immediately next to and flush with each other for more consistent lighting, whereas the previous version lets the angle of the skin and card vary freely.

Constraining the skin to lie within this hole and adding black or white fiducial markers to the corners of the card also made it more straightforward to automate the segmentation process, as described in subsubsection 3.6.2.1.

This new card uses a different set of colors than that of the previous studies. It still includes cyan, magenta, and yellow as they can be printed with standard printer ink without mixing. Rather than including any skin tones, as it is impossible to represent the full spectrum of human skin, it instead includes 5 shades of gray in addition to white and black (in the fiducial markers) to represent an even spectrum of red, green, and blue reflectance. I arranged these colors to offer a high contrast between adjacent patches to improve segmentation.

We worked with ColorGraphics, a printing company, to print and die-cut these cards. By working with this company, I could scale the card production to match the needs of collecting data at the 7 data collection sites (whereas previously, I manually printed and hand-cut cut each card). The cards are printed on specially coated paper to reduce glare, and color accuracy is checked during the printing process to insure batch-to-batch stability.

3.6.1.3 BiliCam Images

Images collected for BiliCam in this study had several key differences from prior studies. For starters, I collected data on iPhone 5S devices instead of iPhone 4S, as it was the newest model at the time.

Rather than capturing both skin from both the sternum and forehead in any one image, images in this study focused exclusively on the sternum; the formal local study provided evidence that the sternum was the more suitable source of data.

As videos in prior studies were difficult to collect (at times, prohibitively cumbersome for quality data) and were not necessary for generating the prior results, I no longer recorded videos and only focused on still images.

Focusing on the sternum and eliminating the lengthy process of taking (and re-taking) videos enabled us to investigate capturing photos from multiple distances. One new distance was just a little further than the phone's focal distance, as well as a distance within the focal range of other types of phones (we call it the "near" distance). To make sure to continue capturing photos from the same positioning as the previous studies, one of these distances was the same as before (the "far" distance). A third distance was in between these two (the "middle" distance).

The new card design and the focus on one skin patch also enabled us to center the light of the flash LED to the skin patch and distribute lighting more evenly on the card. However, data and results from the previous studies do not center the LED's light in this manner, so I maintained the same offset of light location in the "far" distance while centering the light in the "near" and "middle" distances.

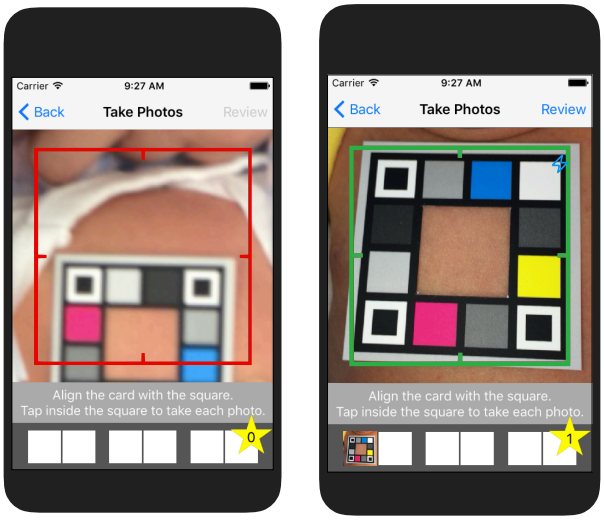


FIGURE 3.16: Screenshots of the data collection application when capturing BiliCam photos with (A) the card not aligned, as denoted by the red square, and (B) with the card aligned and free of visual obstructions, as denoted by the green square.

3.6.1.4 Data Collection App

In addition to the necessary changes for collecting the new types of data, the data collection app underwent a complete redesign for the nation-wide study. Interviews and user studies with all members of the data collection team after the formal local study and design iterations with the rest of the team informed these changes. This redesign both improved the app's usability for both workflow and improving data quality.

The new app ordered tasks differently. When one initiates taking a sample, the app immediately opens the view for takings photos rather than an intermediate view that also has options to enter study ID or date of birth; many data collectors entered study ID and date of birth after taking photos so this UI better suited this workflow.

When taking photos, a square marks where to align the color card, as shown in Figure 3.16. The square is red unless the image quality meets the standards, in which case the square is green. The hues for red and green come from the colors of traffic lights, which are more color-blind friendly (in RGB, red is (234, 3, 0) and green is (33, 176, 63)). I also tested these red and green hues with a red-green colorblind colleague to make sure they are distinguishable.

Over the course of taking photos, the app automatically turns the flash LED on or off, and digitally zooms for the different photo distances while showing the same interface. These photos only capture visual information of the card, the skin patch, and a small border of skin around the card, excluding all other information (e.g., face, arms, etc.) by

design.

For visual feedback, thumbnails of each photo fill in empty squares at the bottom of the screen mark the progression of data capture, instead of using a progress bar, as shown in Figure 3.16b. The app visually pairs thumbnails from the same photo distance. This progression begins with pairs of photos (with and without flash) from the near distance, followed by pairs from the middle distance, pairs from the far distance, and then repeats these three pairs (a total of 12 photos). Although it still offers data collectors the option to collect even more photos, the key change here is that the request to take two sets of each photo is built in, rather than requiring us to verbally clarify to each data collector the importance of taking multiple photo sets.

The view for reviewing collected data has several changes as well. Because I collect many more photos than before, rather than deleting the entire sample to retake photos if they accidentally contain sensitive information, the app offers the option to delete individual photos. Because manually reviewing and labeling each photo for possible image quality issues is very time-consuming during data analysis downstream, the app also offers an interface to flag photos for any noteworthy issues, along with checkboxes for common issues to make it easier.This ability to flag photos and list issues was also intended to make quality concerns explicit. In practice, the data collectors never used this feature – something to keep in mind when deciding where to place efforts when building other apps like this one.

During the early part of conducting this nation-wide study, I introduce the ability for the app to automatically capture photos. Initially, data collectors needed to manually capture photos by tapping the screen when the square was green. Data from the early part of the study enabled us to tune and validate the tolerances for this green square, after which I could trust these tolerances to automatically trigger image capture. For stability, it waits to capture images until at least a few images in a row fit within these tolerances; a later image sometimes has higher quality than the first, which may be slightly blurrier from movement or the camera adjust focus. Just in case these tolerances are too tight for a particular data collection session, the ability to capture photos manually was still available so that it does not block data capture. Introducing auto-capture improved the workflow, as it freed attention for data collectors to spend with the families the were working with, and made it easier to capture photos from the far distance (it can be difficult to view the screen when holding the phone up high).

Much like in the local studies, I continued to make instructions and tips accessible in the app. For the full set of instructions, refer to Appendix A.

3.6.2 Algorithm

3.6.2.1 Card Segmentation

The card segmentation algorithm in the nation-wide study was quick enough to run in real-time for automatic photo capture. Introducing the QR-code like black fiducial markers in three corners of the card and the white patch in the fourth corner enabled this quick extraction method.

The process begins by finding the fiducial markers. It resizes the image to standard dimensions, as photos taken from different distances would have different sizes. It then isolates the card's fiducial markers by applying an upper and lower threshold on a gray scale version of this image, additionally applying a morphological open to reduce noise. After that, it searches for contours of these markers using the contour-finding algorithm developed by [126], then checking which of these contours fit the general shape, expected size, expected locations, and expected quantity for these markers. If it finds them, it simplifies the contours to have only 4 points (one for each corner) using the Douglas-Peuker algorithm [102].

Using the locations of these markers, it locates the white patch and determines the corners of the card. It first orders the marker contours relative to each other and the white patch. After approximating the location of the white patch based on the arrangement and positions of these markers,it finds the corresponding contour of the white patch in the image. It then extrapolates the corners of the card by finding the center of all these markers and choosing that point on each marker that is farthest away from this center. Because the corners of the black markers are slightly inset from the edges of the card, some final tweaking adjusts the corners by offsetting them accordingly.

The corners of the card and the orientation determined by the ordering of black fiducial markers are enough to segment the rest of the card. It transforms the coordinates of these corners of this resized image back into the coordinates of the original image. It can then use a homographic transformation from the standard locations of each patch to calculate where they are in the image while accounting for distortions from perspective. To incorporate some wiggle room, it adds some padding around each card patch. It finally extracts the regions of interest within the padded portion of each patch.

3.6.2.2 Feature Extraction

BiliCam aims to estimate bilirubin concentrations through observing differences in skin coloration, so the feature extraction process focuses on different ways to characterize representative colors from each sample. This process

generates many features for the modeling process to choose from, with the expectation that a number of them will be discarded.

To select colors of each patch of skin and each color patch on the card, features focus on the median pixel values of a patch's region of interest, calculated separately for each color channel. It uses the median in favor of the mean to isolate colors that are representative of clear skin, especially in the presence of uneven colors like that of mottled skin, rashes, or other subtle and undetected occlusions of the prevalent skin color. In the event that the median value misses other information from the skin, other feature options include the 32$^{\text{nd}}$ and 68$^{\text{th}}$ percentile of pixel color channel values (i.e., ± one standard deviation).

There are several options for preprocessing the skin patch to "correct" for lighting conditions. They calculate a transformation from the *observed* values of card patch colors to the expected or *reference* values, then applies this transformation to the observed skin color to calculate the *adjusted* skin color.

One such option is the white balancing method used in the formal local study, which only uses the white patch and is popular in many image-editing tools. To repeat this method's description from earlier, given observed skin RGB values (O_R, O_G, O_B), it computes the adjusted (A_R, A_G, A_B) by

$$\begin{bmatrix} A_R \\ A_G \\ A_B \end{bmatrix} = \begin{bmatrix} 255/O_{R_w} & 0 & 0 \\ 0 & 255/O_{G_w} & 0 \\ 0 & 0 & 255/O_{B_w} \end{bmatrix} \begin{bmatrix} O_R \\ O_G \\ O_B \end{bmatrix}$$

where $(O_{R_w}, O_{G_w}, O_{B_w})$ is the median observed color of the white patch on the color calibration card and $(255, 255, 255)$ is the reference color of the white patch [138]. By treating the entire image as a matrix of RGB values, the image matrix can replace the vector of observed red, green, and blue values for efficiency.

Another method to "correct" observed colors uses all the card's reference colors, which was used in a related work called BiliScreen [83]. In this case, the process calculates a calibration matrix using information from all the observed card patch colors such that

$$\begin{bmatrix} A_R \\ A_G \\ A_B \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix} \begin{bmatrix} O_R \\ O_G \\ O_B \end{bmatrix}$$

where the individual C's represent elements of the calibration matrix. It calculates this calibration matrix using an iterative least-squares approach by Wolf [143] and the CIEDE2000 color error [115] to quantify color difference during

these iterations. It initializes C as

$$C_{initial} = \begin{bmatrix} \frac{\text{mean}(R_{R_i})}{\text{mean}(O_{R_i})} & 0 & 0 \\ 0 & \frac{\text{mean}(R_{G_i})}{\text{mean}(O_{G_i})} & 0 \\ 0 & 0 & \frac{\text{mean}(R_{B_i})}{\text{mean}(O_{B_i})} \end{bmatrix}$$

before these iterations, where R_{R_i} represents the reference red values of the i card patches, O_{R_i} the observed red values of the i card patches, and so forth. Iterations repeat until the sum of squared errors converges below the tolerance of $1\text{e}{-3}$. If that fails, it instead initializes the calibration matrix to the identity matrix and tries again. To account for gamma encoding of images, observed values undergo gamma correction (raising observed colors to the power of γ) before applying the calibration matrix to compute adjusted colors. A reversion of the gamma correction (by raising adjusted colors to the power of $1/\gamma$) results in the final adjusted colors.

For any one BiliCam model, all of its skin-related features underwent the same preprocessing pipeline (if any); the different color balancing methods were not combined. The options I explored for model features include features from skin that was exclusively white balancing using only the white card patch; exclusively color balanced using all card patches; exclusively color balanced using only the white, gray, and black card patches; or exclusively not processed at all. For each of these options, any additional features from the card patches as direct inputs to the model had no such preprocessing with the expectation that their unprocessed values provide information about the lighting conditions. Combinations to explore include incorporating all patches as features, only using the skin patch for features, only the skin and white patch for features, and only the skin and two card patches (white and light gray) for features. Feature sets that exclude the card entirely can help investigate the relative model performance of using card-free BiliCam images, and feature sets that only use the skin and white patch could help investigate the relative model performance of depending only on the paper of the card (not the ink).

While cameras express colors in RGB color space (i.e., with red, green, and blue color channels), channels from other colorspaces highlight and express different properties of a color. To take advantage of potential differences in signal strength of channels in other color spaces, features include conversions to different color spaces that separate chroma and luminance: HSV (hue, saturation, value), CIE XYZ, YCbCr, CIE L*a*b*, and CIE L*u*v*. Some of these color space conversions involve non-linear transformations from RGB color space – a transformation that a linear regression model could not represent or reproduce with only RGB. Representative values (i.e., the median and percentiles) for each color space come from entire image patches converted to the respective color space; they are not conversions of

the representative RGB values, which would yield different numbers.

Bilirubin primarily absorbs specific frequencies of blue light – a property that the TcB uses, and is also what causes the yellowish appearance of jaundice [55]. Some channels from these aforementioned colorspaces better isolate the relative blueness in chroma, compared to the absolute blueness from the blue channel in RGB. For its potential to directly isolate the relative blueness from the RGB channels, two additional features are the ratio of blue to green and the ratio of blue to red from the representative RGB values.

All features were scaled to have unit variance and zero mean before used to train models. As these different color spaces encode similar information, I also investigated whether models would benefit from reducing this feature redundancy through applying principal component analysis (PCA). Within applying PCA, I investigated two options: applying PCA once to all features (both skin and card), or applying PCA separately for skin features and card features. The idea here is that it may be better keep information from the skin and cards separate. All scaling and PCA were part of machine learning pipelines and fitted using only training data in cross validation folds. The number of components for PCA was chosen programatically to be the lowest number of components such that the explained variance was at least 99%.

As discussed in subsubsection 3.6.1.1 and subsubsection 3.6.1.3, the image dataset include baseline and follow-up photos; photos with and without flash; and photos from near, middle, and far distances. Combinations to examine include whether to include baseline images (2 options); using both flash options together as well as either exclusively one flash option or the other (3 options); and using all photo distances together, exclusively one distance, or only the near and middle distances because the far distance caused usability challenges (5 options). Together with whether to include color percentiles (2 options), the different ways to apply color balancing (4 options, including no preprocessing), what patches to include (4 options), and choice of PCA usage (3 options) results in $2 \times 3 \times 5 \times 2 \times 4 \times 4 \times 3 = 2,880$ options of feature combinations for feature selection and modeling. Rather than compute the outcomes of using each possible option for each possible regressor, the modeling process explored a subset of these options at a time (as described next, in subsubsection 3.6.2.3).

3.6.2.3 Bilirubin Estimation Modeling Process

As discussed in subsection 3.2.1, BiliCam involves a regression model to determine a BiliCam-estimated bilirubin (BCB) in mg/dL, using the TSB for ground truth. Cross validation plays an integral role in evaluating BiliCam models to take advantage of the full dataset. For any one cross validation split, the full modeling pipeline (complete with feature

scaling, any PCA, hyper parameter tuning with further cross validation folds, and algorithmic feature selection) is fit on the training folds, then applied to the test fold to get prediction results for model evaluation. While a composite of test predictions from each split cannot represent test results for any one tuned model, it can represent test results for the modeling method and hence a way to evaluate BiliCam while taking full advantage of the dataset.

All test predictions were constrained to fit within an interpretable range for bilirubin levels: a BCB of 0 replaced any predictions below 0 and a BCB of 28 mg/dL replaced any predictions higher than 28 mg/dL, as 28 mg/dL is already considered an extremely high bilirubin level for any human being of any age.

I considered a number of different regression models, including linear models (Lasso [133], Ridge [59], Bayesian Ridge [76], LARS, LARS-Lasso [40], Elastic Net [147], Generalized Additive Model (GAM) [56]) and non-linear models (SVR [122], KNN [53], random forests [24], AdaBoosted trees [49], gradient boosted trees via XGBoost [31], and neural networks [108]). In the case of neural networks, each image (regardless of distance or flash setting) was treated as a separate training sample with its own TSB with only the median, un-adjusted RGB values for each skin and card patch as features. Predictions from test samples were aggregated by study ID by either taking the mean or the max BCB predicted from the 6 images per participant. For many of these algorithms, I also explored using cross-validated recursive feature elimination (RFECV) [10] as part of the pipeline in training folds to select features.

To include an additional baseline to compare models against, I also trained the simplest model I could think of: an ordinary least squares linear regression (with no regularization) using 6 features: only the un-adjusted median red, green, and blue values of skin patches photographed at the near distance both with and without flash.

Rather than attempt to train every model with of the 2,880 possible feature combinations for feature selection (which would take a month of non-stop computation if each model unrealistically took only one minute to train and tune), I investigated only a few of these different dimensions at a time.

Some features will hold more predictive power than others for estimating bilirubin levels. I operated under the assumption that features with a higher signal strength will offer more predictive power for any model using it, and that including features with poor signal strength would weaken the performance of any model using it. Hence, I used one type of regression model to investigate several of the different dimensions for the feature combinations described in subsubsection 3.6.2.2: whether to include percentiles, whether or how to incorporate PCA, which photo distances to include, what flash options to include, and which card patches to incorporate (if any) – a total of $2 \times 3 \times 5 \times 3 \times 4 = 360$ combinations. To choose which regression model to use for this exploration, I trained, tuned, tested, and evaluated all the aforementioned regression models with all BiliCam features excluding baseline photos. The regression model

for this feature exploration came from selecting the fastest model to train among the regression models that tied in BCB predictive performance. As the performance of this model on one feature set did not change significantly between using 3, 10, or more cross validation folds, I used only 3 cross validation folds while exploring the 360 feature combinations for time efficiency.

After testing the BCB performance of the chosen regression model on the 360 different feature combinations, I selected four of the highest performing combinations to explore the remaining options (i.e., whether to include baselines and what color-adjustment preprocessing method to use). As these highest performers included both combinations with and without PCA, I required two of these chosen combinations to exclude PCA in order to enable further analysis on the predictive power of individual features.

To evaluate whether baselines improve model performance, I compared models trained both with and without data from baseline photos, using all four of the aforementioned feature combinations (a total of 8 model comparisons). Only 247 of the samples have baseline photos. To do a direct comparison on model performance with and without baseline features, I only used these 247 samples for training and testing both models that use baseline features and comparison models that do not. I increased the number of cross validation folds from 3 to 30, as this sample size of 247 is smaller than even training set from 3-fold cross validation on 530 samples (i.e., 350 training samples per split). I validated this number of cross validation folds by checking that the performance of training without baselines on this smaller dataset matches the performance from the larger dataset. Model comparisons for the four different color-adjustment preprocessing methods incorporated results of whether baselines improve model performance.

After this exploration of feature combinations, I trained and evaluated a subset of the different regression models yet again, but on the four feature combinations that offered the strongest signal and using 10 cross validation folds instead of 3.

As the far distance demonstrated poor predictive power, I re-incoporated samples that were previously excluded for missing a far-distance photo, which increased the dataset size from 530 participants to 552. I further experimented with treating near and middle distance photos as separate samples (resulting in 1104 samples) and applying ensembles to their predictions (namely, taking the mean and the max of predictions for samples from the same participant). I was careful to split validation folds such that samples from the same participant would always be in the same fold.

The highest performing model was chosen as the model for evaluating BiliCam as a concept in subsection 3.6.3 and for the additional analysis detailed in subsection 3.6.4. A 100-fold cross validation process with this model produced the final results for analysis.

3.6.2.4 Evaluation Methods

The primary outcome was the linear correlation between BCB and TSB values; subgroup analyses were conducted for newborns from various racial groups. The mean (± SD) difference between paired BCB and TSB levels was also calculated. A Bland-Altman analysis was performed; mean bias and limits of agreement were calculated. Similar analyses were done comparing paired TcB and TSB measurements.

I assessed the utility of BiliCam and TcB as screening tools for identifying newborns with significant jaundice using 2 recommended decision rules [79]. First, BiliCam and TcB levels were plotted on the Bhutani TSB Nomogram [14]. A positive test result was a BiliCam or TcB level of $\geq 75^{th}$ percentile on the nomogram (ie, in the high-intermediate or high-risk zone), with a positive results being a TSB level of $\geq 95^{th}$ percentile (high-risk zone) [13, 79].

The second decision rule is to use a cut-off predicted bilirubin level to identify newborns with a TSB > 17.0 mg/dL. To be consistent with medical literature and practice, I used a cut-off of 13.0 mg/dL for evaluating TcB (i.e., a positive test for TcB was fixed at TcB level > 13.0 mg/dL) [78, 79]. As BiliCam does not have standard practices for cut-off levels yet, BiliCam's cut-off level was tuned to be the highest possible cut-off level to achieve a sensitivity of 100%.

The sensitivity, specificity, positive predictive value (PPV), and negative predictive (NPV) value were calculated for BiliCam and TcB for each decision rule. I further compared the utility of BiliCam and TcB as screening tools for the second decision rule by constructing receiver operator characteristic curves (ROC curves).

3.6.3 Results

3.6.3.1 Study Demographics

As stated at the beginning of section 3.6, from a data collection study with 600 newborn participants, 552 presented usable data for the final model. BiliCam images were obtained for 580 enrolled newborns. A matching TSB level was missing for 8 participants; in 2 newborns, there was a laboratory problem, parents of 2 participants declined the blood draw, and no matching TSB level was obtained for for 4 infants. In addition, data on 3 newborns were excluded because of issues with the consenting process. From the remaining 569 participants, a complete set of BiliCam images was obtained for 530 newborns (93% of those who were eligible). Loosening this restriction to be complete sets of images taken only from the near and middle distances (with no regard to the far distance) results in a dataset from 552 newborns (97% of those who were eligible).

The race of the 552 newborns with complete TSB and BiliCam data are summarized in Table 3.2 and Figure 3.18. The mean age of these participants at the time that BiliCam images were obtained was 74.7 ± 29.1 hours, with a range

Participant Demographics (N=552)	
Age at follow-up (hours) (mean, range)	74.7 (12 – 163)
Bilirubin Levels (mg/dl) (mean, range)	10.4 (0.6 – 24.8)
Reported Race (n, %)	
African American/Black:	116 (21%)
American Indian/Alaska Native:	15 (3%)
Asian:	116 (21%)
Latinx:	150 (27%)
Pacific Islander/Native Hawaiian:	15 (3%)
White:	312 (57%)
Multiple Races:	148 (27%)
Not Reported:	6 (1%)

TABLE 3.2: Demographic information for participants. Note that participants may report multiple races.

of 12 to 163 hours (see Figure 3.18b for more details). The time between the TSB blood draw <2 hours for 97% of the participants; the time difference was <3 hours in the remaining 2% of participants.

The mean TSB value in the 552 study participants was 10.4 ± 4.4 mg/dL, with a range of 0.6 to 24.8 mg/dL. There were 68 participants (12.3%) with a TSB level in the high-risk zone on the Bhutani Nomogram. 86 participants (15.6%) had TSB levels of >15.0 mg/dL, and 30 participants (5.4%) had TSB levels of >17.0 mg/dL. Figure 3.18a shows a full distribution of the study participant TSB levels.

TcB was collected for only 312 study participants (subsubsection 3.6.1.1 describes which ones). In comparison, they had a mean TSB value of 9.4 ± 4.3 mg/dL with a range of 0.6 to 19.1 mg/dL. Also for this subset of partipants, 12 (3.8%) had TSB levels in the high-risk zone on the Bhutani Nomogram, 29 (9.3%) had TSB levels of >15 mg/dL, and 7 (2.2%) had TSB levels of >17.0mg/dL. It is important to keep in mind that this set of participants had a smaller proportion of high TSB levels when examining results.

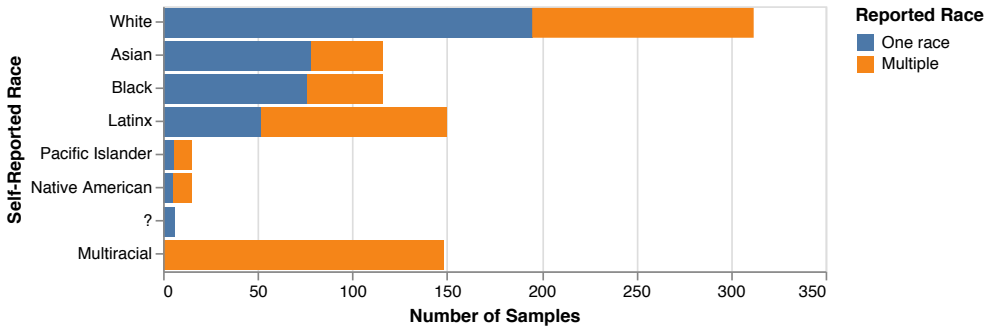
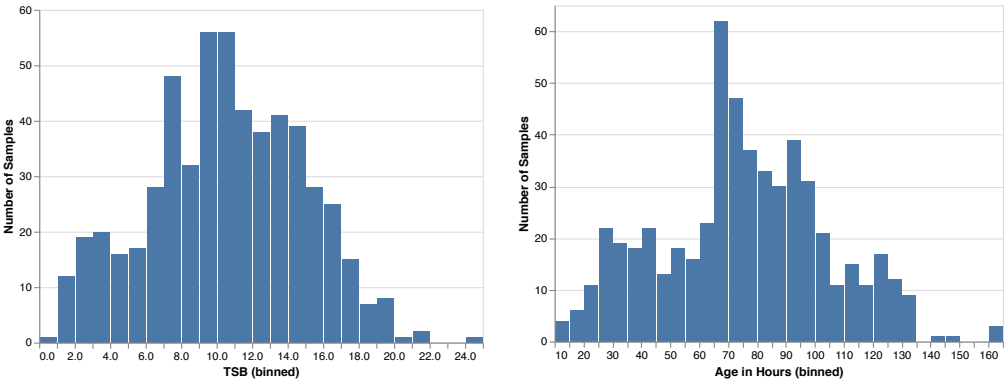


FIGURE 3.17: A histogram illustrating the distribution of self-reported racial demographics in the dataset, color-coded by whether anparticipant associated with more than one race.



(A) TSB distribution (B) Age distribution

FIGURE 3.18: Histograms illustrating the distribution of TSB and age of participants in BiliCam's nationwide study.

3.6.3.2 BiliCam Model

The linear correlation between predictions and TSB for many different models and feature combinations were either identical or very similar, failing to show statistically significant differences from each other. The evaluation metric that varied most between models was the maximum specificity at 100% sensitivity for classifying TSB >17.0 mg/dL.

The resulting model uses a Lasso regression model, with features from BiliCam photos taken at the near and middle distances, without flash, with data from all card patches in addition to the skin, without applying any color adjustments, and without PCA. This model does not including data from baseline photos as, like in the formal local study, including them failed to demonstrate a statistically significant difference in performance. It also only uses the median values from card patches, as introducing additional percentiles does not make much of a difference.

The underlying Lasso regressor in the model treats photos from different distances as individual samples. A BCB from the overall model is the Lasso regressor's maximum prediction from photos for the same patient. By using the maximum of these predictions, we can err on the side of caution to help avoid false negatives.

Not using photos taken from the far distance improves the usability of BiliCam. They were the most difficult photos to take, as it often required holding the phone high above a newborn that is already elevated in a crib and, unless the data collector is very tall, it can be difficult to simultaneously watch the phone's screen and operate the app. This challenge made the request to remove photos for the far distance the most frequently requested change to study procedures.

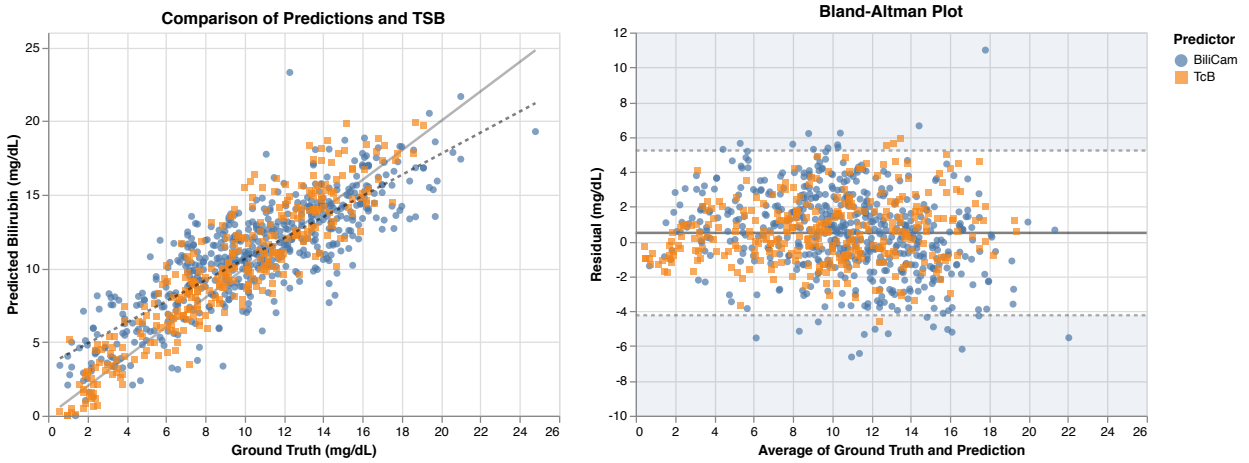


FIGURE 3.19: Bilirubin estimation results for BiliCam and TcB. **Left:** The relationship between paired TSB and predicted bilirubin levels, with blue for BiliCam's predictions and orange for TcB. A dashed line illustrates the line of best fit for BiliCam's predictions, and a solid line illustrates the ideal line of fit. **Right:** A Bland-Altman plot of paired TSB and predicted bilirubin levels, with blue for BiliCam's predictions and orange for TcB. A line along the x-axis denotes the mean difference of TSB and BiliCam's values (in mg/dL). Dashed lines and shaded regions denote the limits of agreement (at a 95% confidence interval) for BiliCam's predictions.

3.6.3.3 Bilirubin Estimation

Based on estimates calculated by using 100-fold cross-validation, the correlation between the BCB level and the paired TSB measurement was 0.83 (95% confidence interval 0.8-0.93). The correlation between BCB and TSB is shown graphically in Figure 3.19. The mean difference between BCB and TSB was 0.5 ± 2.4 mg/dL, with a range of -6.6 to +11.0 mg/dL; 77.2% of BCB values were within 3 mg/dL of the paired TSB level, and 59.4% were within 2 mg/dL. A Bland-Altman plot summarizing the differences between BCB and TSB is presented in Figure 3.19; as shown, the limits of agreement were -4.2 to +5.2 mg/dL.

TcB measurements were made for 312 study newborns. Among these infants, the correlation between TcB and TSB was 0.92 (95% confidence interval 0.89–0.93). The mean TcB-TSB difference was 0.54 ± 1.8 mg/dL, with a range of -4.6 to +5.9 mg/dL; the limits of agreement were -3.1 to +4.2mg/dL.

The correlations between BCB and TSB values are consistent with results of published studies on the accuracy of TcB in estimating bilirubin levels in newborns. The reported correlations between TcB and TSB range from 0.77 to 0.97 [13, 21, 39, 41, 42, 81, 82, 106, 110–112, 120, 129, 144]. Most of these studies were conducted on newborns during their birth hospitalizations. Although there are limited data on the accuracy of TcB measurements in outpatients,

Device	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Predicting high-risk zone on Bhutani Nomogram				
BiliCam	82.4	71.7	29.0	96.7
TcB	74.8	81.3	13.9	98.8
Predicting TSB level > 17.0 mg/dL				
BiliCam	100	69.9	16.0	100
TcB	100	74.8	8.3	100

TABLE 3.3: Utility of BiliCam and TcB as screening tools to identify newborns with high TSB values, defined as a TSB level in the high-risk zone on the Bhutani Nomogram or as a TSB level of >17.0 mg/dL.

in 2 studies in which researchers focused on neonates after hospital discharge, correlations between TcB and TSB were found to be 0.77 and 0.78, respectively [42, 82]. A possible reason for the lower correlations found in outpatient newborns is that TSB levels tend to peak after newborns are typically discharged from their birth hospitalizations [27, 35, 36, 47, 66, 80]. TcB levels have been found to progressively underestimate serum values in neonates with higher TSB levels, particularly >15.0 mg/dL [41, 77, 129].

3.6.3.4 Classification for Screening

The utility of BiliCam as a screening tool for identifying newborns with significant hyperbilirubinemia is summarized in Table 3.3, alongside values for TcB collected in the study. Note that a formal comparison of TcB and BiliCam as methods for screening neonates with TSB values in the high-risk zone (>17 mg/dL) was limited to 312 participants with TcB readings that had a smaller proportion of high TSB levels compared to BiliCam's 552 samples. As is shown in Table 3.3, BiliCam had a sensitivity of 82.4% for identifying newborns with a TSB level in the high-risk zone on the Bhutani Nomogram and a specificity of 71.7%. For BiliCam to identify a neonate with a TSB level >17.0 mg/dL at 100% sensitivity, it could achieve a 69.9% specificity by applying a cut-off of 12.6 mg/dL.

For comparison, the TcB measurements in this study yield a 75.0% sensitivity and 81.3% specificity for identifying newborns in the high-risk zone on the Bhutani Nomogram, and a 100% sensitivity and 74.8% specificity for identifying newborns with a TSB >17.0 mg/dL. For additional points of reference, here are additional TcB performance reportings in medical literature. Bhutani et al. reported that the BiliChek-brand TcB meter had a sensitivity of 100% and a specificity of 88.1% in identifying newborns with a TSB level in the high-risk zone on the Bhutani Nomogram during their birth hospitalizations [13]. In a study evaluating both the BiliChek and JM-103 brands of TcB meters in 2 populations of newborns assessed during their birth hospitalizations, the sensitivity of TcB screening was 94.1% and 91.9%, respectively, for this same outcome [128]. However, in a study of outpatient newborns with higher TSB levels, the sensitivity and specificity of TcB screening for identifying newborns with high-risk zone TSB levels were 79% and 84%, respectively [42].

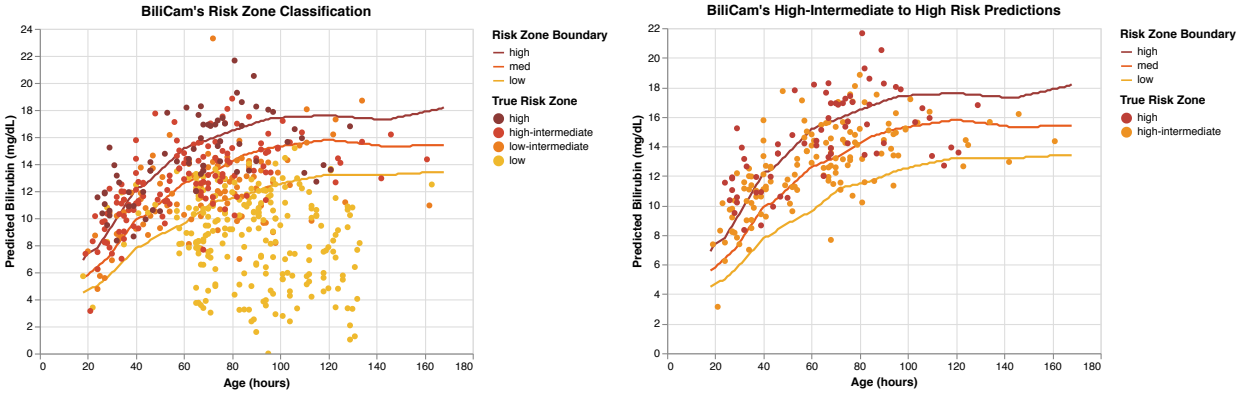


FIGURE 3.20: Classification results using Bhutani Nomogram. **Left:** BCB plotted on the nomogram for predicted risk zones, colored by the ground truth risk zone. **Right:** the same chart but showing only samples that are in high and high-intermediate risk zones.

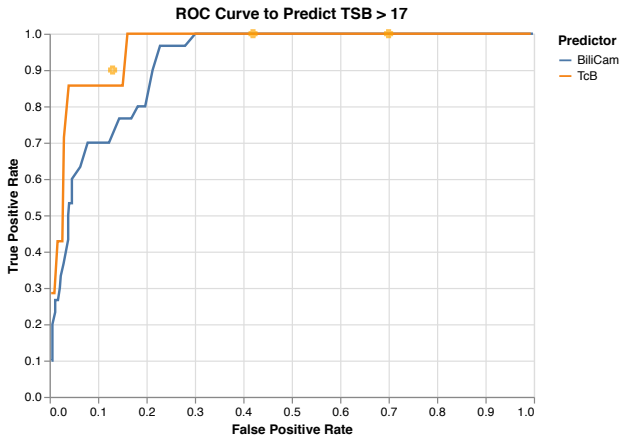


FIGURE 3.21: ROC curve to predict TSB > 17.0 mg/dL by adjusting the cut-off BCB. Additional orange points mark TcB performances reported in literature [42, 82, 106]

In this same study using a cutoff value of >13.0 mg/dL to define a positive TcB screen, the sensitivity of TcB screening was 100%, and the specificity was 58% for identifying outpatient neonates with a TSB >17.0 mg/dL.

3.6.4 Further Analysis

This dataset of BiliCam images with TSB measures for 530 newborn participants (the fruits of 4 to 5 years of labor) enables additional analysis that was not possible with earlier, smaller datasets. It has enough data to explore performance differences between different self-reported racial groups and between different lighting conditions, as reflected by card patches.

3.6.4.1 Performance by Race

Individual correlation coefficients for reported racial groups are summarized in Table 3.4. Among the reported racial groups represented by at least 10 BiliCam samples, the correlation was lowest among Asian newborns.

Reported Race	Count	Percent (%)	Correlation	95% CI
African American/Black	74	13	0.82	0.73-0.88
Asian	78	14	0.78	0.68-0.85
Latinx	50	9	0.85	0.75-0.91
White	189	34	0.82	0.77-0.86
Multiracial	144	26	0.82	0.76-0.87
American Indian/Alaska Native	5	1	0.99	0.85-0.99
Pacific Islander/Native Hawaiian	6	1	0.73	-0.02-0.97
Not Reported	6	1	0.07	-0.79-0.83

TABLE 3.4: Correlations between TSB and BCB among study newborns from different reported racial groups. For the purposes of these calculations, the Multiracial group exclusively represents all participants with more than one reported race.

To help examine these different performances between racial groups, Figure 3.22 and Figure 3.23 plot comparisons between BCB and TSB for these individual racial groups. As illustrated in these plots, BiliCam tends to underestimate bilirubin levels for Asian newborns and overestimates bilirubin levels for black newborns.

One confounding factor in this analysis is that the BiliCam dataset has a higher proportion of high TSB levels among Asian newborns and a higher proportion of low TSB levels among black newborns (which you can see in Figure 3.22 and Figure 3.23). As documented in medical literature [74], it is common for Asian newborns to have a higher average TSB, and for black newborns to have lower average TSB. To tease apart whether or how much BiliCam is biased by race, a more informative approach is to examine the prediction error by TSB level. To do so, I calculated the five-number

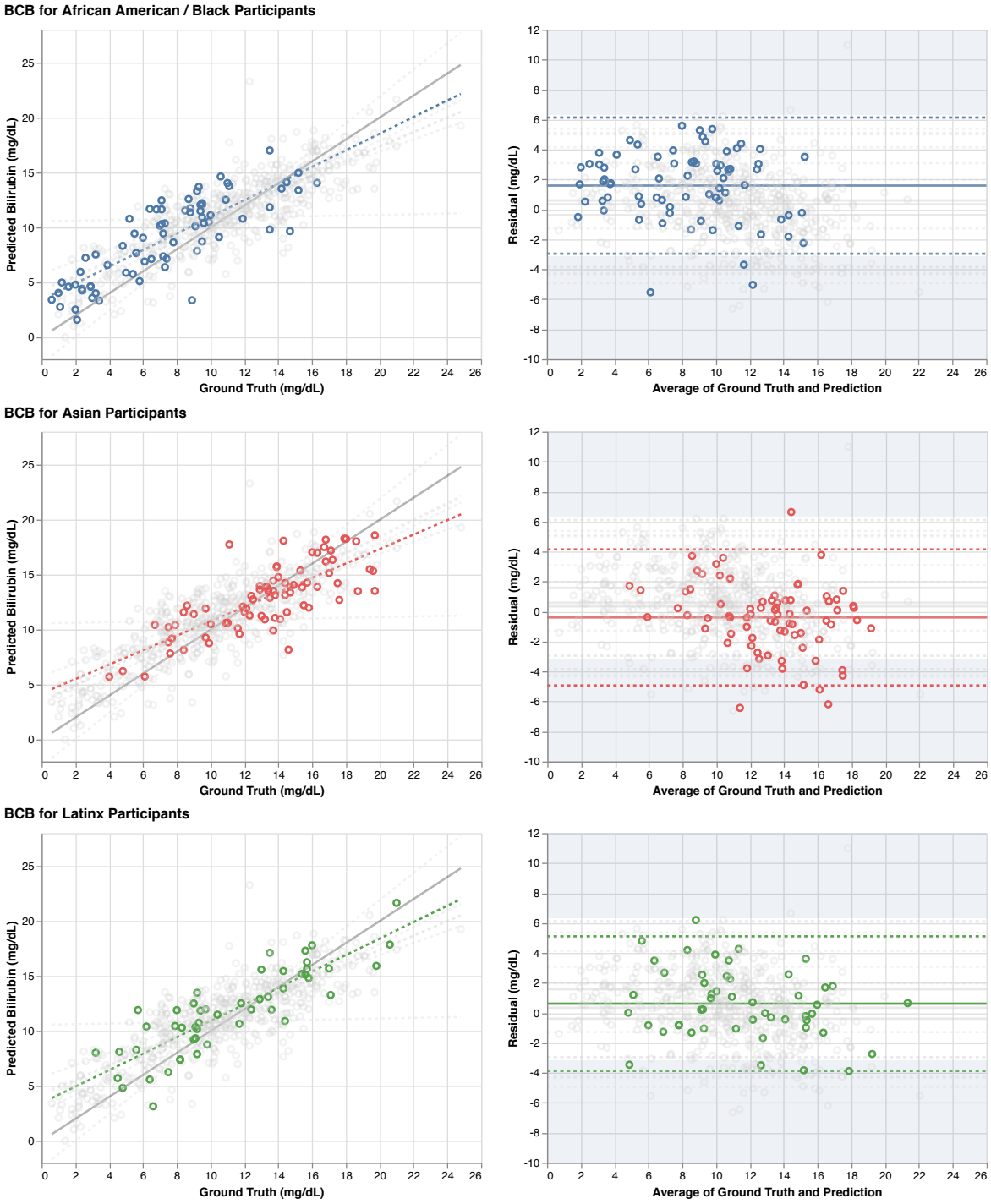


FIGURE 3.22: Regression results presented in a similar way to Figure 3.19, but focusing on predictions for samples that are reported as African American / black, asian, and Latinx. The shaded areas of the Bland-Altman plots refer to the limits of agreement for the aggregate BiliCam model, whereas the dashed lines refer to the limits of agreement for a particular racial group. For the purposes of these calculations, the Multiracial group exclusively represents all participants with more than one reported race.

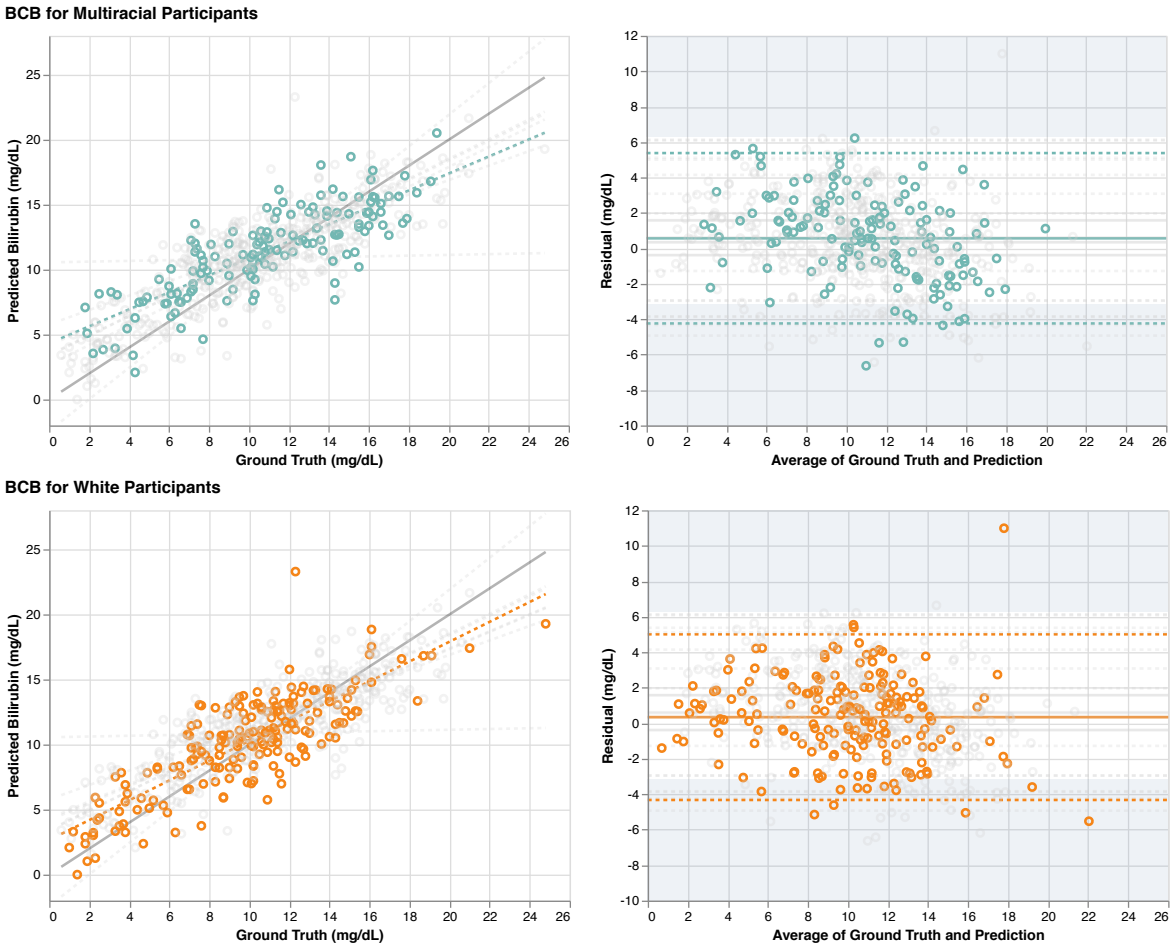


FIGURE 3.23: A conditiuation of Figure 3.22, focusing on predictions for samples that are reported as multiracial and white.

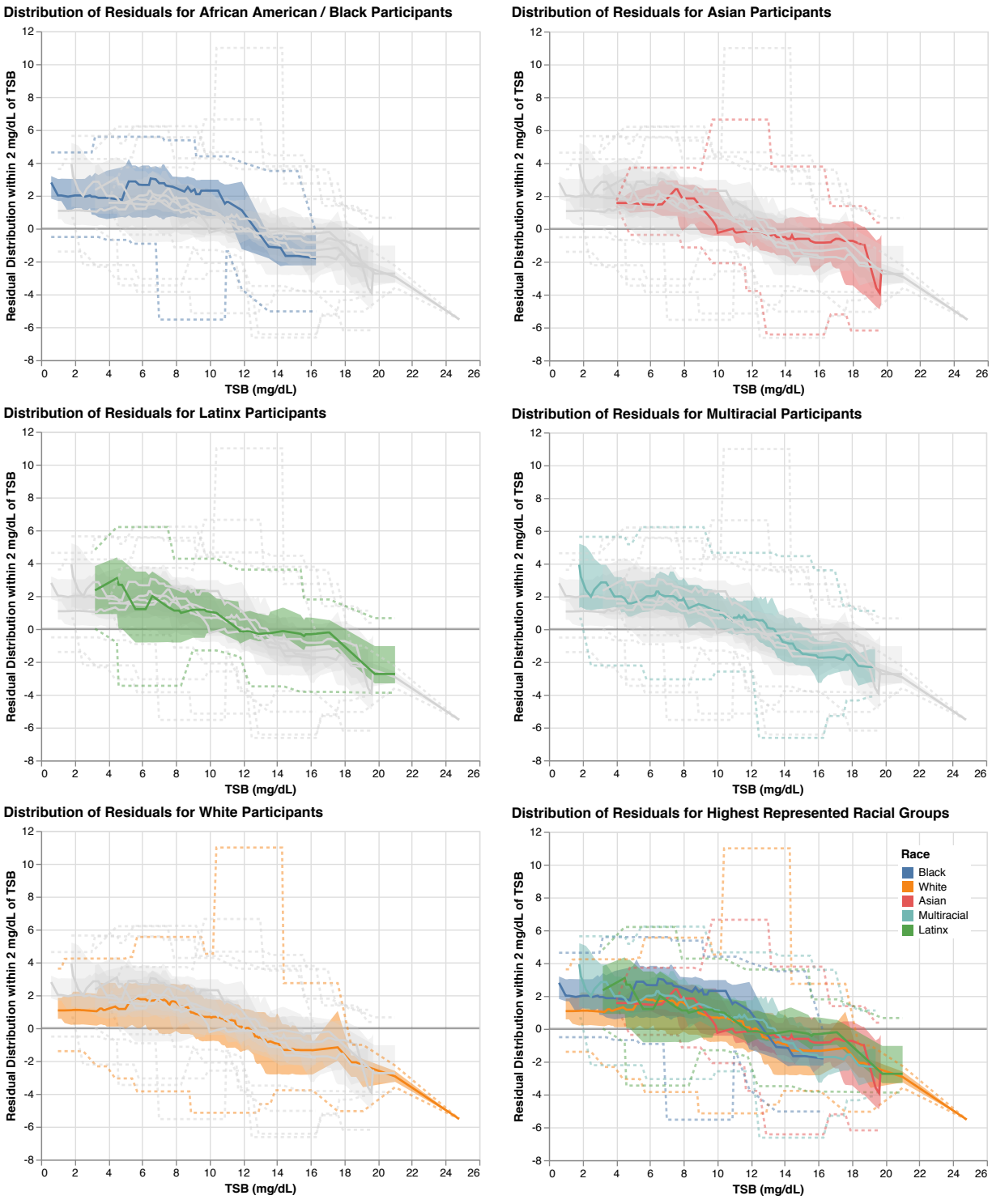


FIGURE 3.24: Five-number summaries of residuals (BCB-TSB) within a sliding window of 4 mg/dL, separated by racial group. Colored and light-gray solid lines denote the median residual, the shaded areas denote the lower to upper quartile of the residuals, and the dashed lines mark the minimum and maximum residuals within 2 mg/dL of the TSB.

summary (i.e., the minimum, lower quarter, median, upper quartile, maximum values) of the residual in a sliding window over different TSB values, in intervals of 4 mg/dL at a time. The results are plotted in Figure 3.24.

Upon examining these windowed residuals, we can see that BiliCam's mean error within these 4 mg/dL windows is unremarkably different for samples reported as Asian in our dataset. Hence, their lower overall performance is more likely due to their skewed TSB distribution than it is due to differences in BiliCam's ability to interpret their photos.

Among the plots in Figure 3.24, the residuals for samples reported as Black in our dataset stand out for having noticeably larger absolute error. The chart shows how BiliCam overestimates their BCB more than for other racial groups when they have low TSB levels. It also shows that BiliCam underestimates their BCB more than for other racial groups for the highest TSB levels represented represented in their data. This evidence hints that this version of BiliCam may not predict bilirubin levels as accurately for black newborns, despite the high reported linear correlation. Unfortunately for analysis (though perhaps fortunately for these newborns), none of the samples reported as black have TSB levels of 17 mg/dL or higher; we cannot explore the relative BiliCam performance for black newborns at these critical high levels, nor can we make a proper evaluation on how well BiliCam performs for black newborns.

3.6.4.2 Lighting Conditions

Another way to break down the data is by lighting condition. I decided to do that by examining the observed colors of lightest gray card patch captured without flash (i.e., illuminated only by the environmental lighting conditions). I chose this light gray patch over the white patch because some of the observed colors of the white patch experienced clipping (i.e., some of the red, green, or blue values of the white patch were higher than the camera could capture, so their recorded values were cut off at 255, the maximum). I explore lighting conditions by examining the median color channel values reflected off of this light gray card patch for different color spaces. I further compared BiliCam's performance on samples grouped by slices of these observed values in any one color channel.

One of the most interesting dimensions along which to explore different lighting conditions is the correlated color temperature (CCT) of the light reflected off of this light gray patch. CCT, measured in degrees Kelvin, is one way to describe the spectrum of "warm" to "cool" lighting conditions (i.e., the spectrum of orange, yellow, white, and blue tints to what we perceive as white light).

I calculated the representative CCT for each sample's light gray card patch using the method by Hernández-Andrés et al. [58] and patch's median values in the XYZ color space. The histogram for CCT values is shown on the right side of Figure 3.25.Because a more even distribution of samples binned by lighting condition would help with

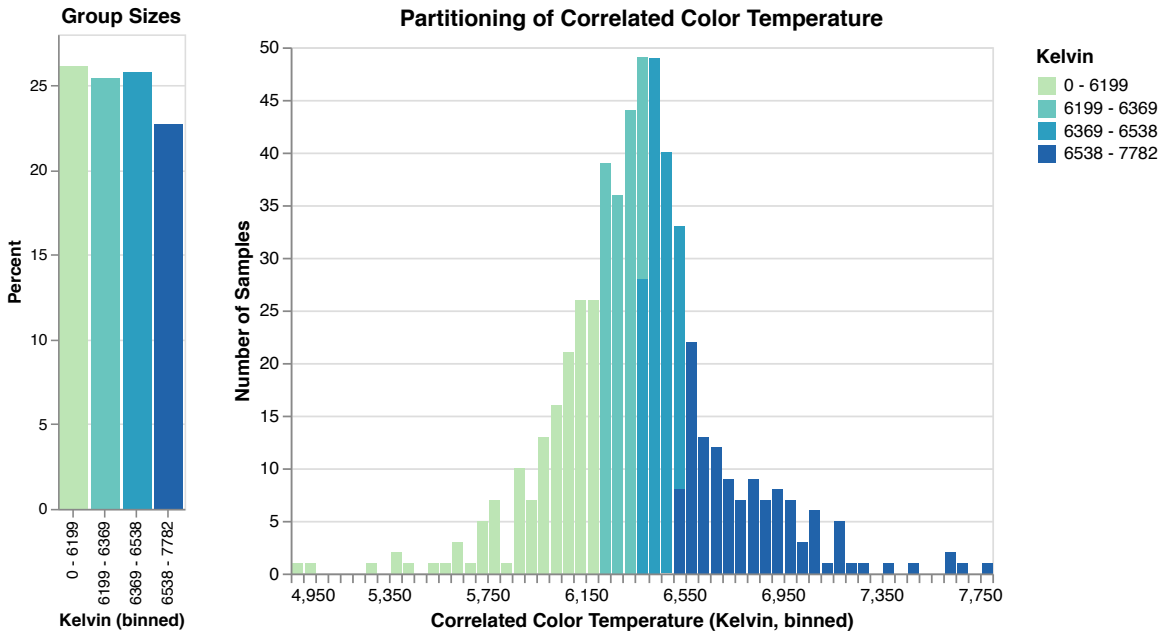


FIGURE 3.25: Histograms illustrating the partitioning of correlated color temperature. **Left:** A histogram of what proportion of study samples each partition represents. **Right:** A histogram of correlated color temperature, colored by partition to illustrated each partition's context and TSB distribution. The correlated color temperature was observed from the lightest gray card patch from photos taken without flash from the middle distance

analysis, I partitioned them into 4 bins separated by the following CCT boundaries defined by the mean and standard deviation of our CCT values: 0, mean $- 0.5 *$ std, mean, mean $+ 0.5 *$ std, max. The histogram of samples using this partitioning, shown on the left side of Figure 3.25, demonstrates that this distribution of samples is more even.

Figure 3.26 summarizes BiliCam's performance broken down by this grouping of CCT. The differences in correlation coefficients are not that exciting; they all fall close to BiliCam's aggregate correlation coefficient. However, the differences in their predictive power of high risk newborns is more interesting. For the decision rule based on the Bhutani Nomorgram, the sensitivity improves with partitions of increasingly high CCT, with the largest difference of a 71% sensitivity to 92% sensitivity between the partitions representing the lowest and highest CCT, respectively. For the decision rule based on predicting TSB $\geq$ 17 mg/dL, performance differences in our evaluation metric (the maximum specificity at 100% sensitivity) stand out more. While this maximum specificity for partitions representing < 6538 Kelvin is between 68% and 77%, the maximum specificity for the partition representing $\geq$ 6538 Kelvin is 96%.

To help examine these different performances between racial groups, Figure 3.27 plots comparisons between BCB and TSB for individual CCT partitions. As illustrated in these plots, BiliCam tends to underestimate bilirubin levels when TSB levels are high for samples in CCT partitions representing < 6538 Kelvin. Meanwhile, BCB for samples with high TSB for samples in the CCT partition representing $\geq$ 6538 Kelvin lies closer to the line of ideal fit.

These performance outcomes for the CCT partitions suggest that BiliCam may perform better under lighting conditions with the higher CCT of 6538 to 7782 Kelvin. One possible explanation is based on the physical properties of bilirubin: bilirubin primarily absorbs blue light [55], and these higher Kelvin correspond to a blue-ish hue for white light. Lower Kelvin would have a less blue-ish hue, thus may offer less signal strength for observing how much the skin absorbs the bluer frequencies of the ambient.

3.7 Discussion

3.7.1 Findings

BiliCam's results suggest that a technology based on the analysis of images obtained by using an app on a commodity smartphone provided reasonably accurate estimates of TSB values in newborn infants. Although BiliCam's correlation of 0.83 does not match the study's TcB correlation of 0.92, it does fall within the 0.77 to 0.97 bounds of TcB correlations documented in literature [13, 21, 39, 41, 42, 81, 82, 106, 110–112, 120, 129, 144].

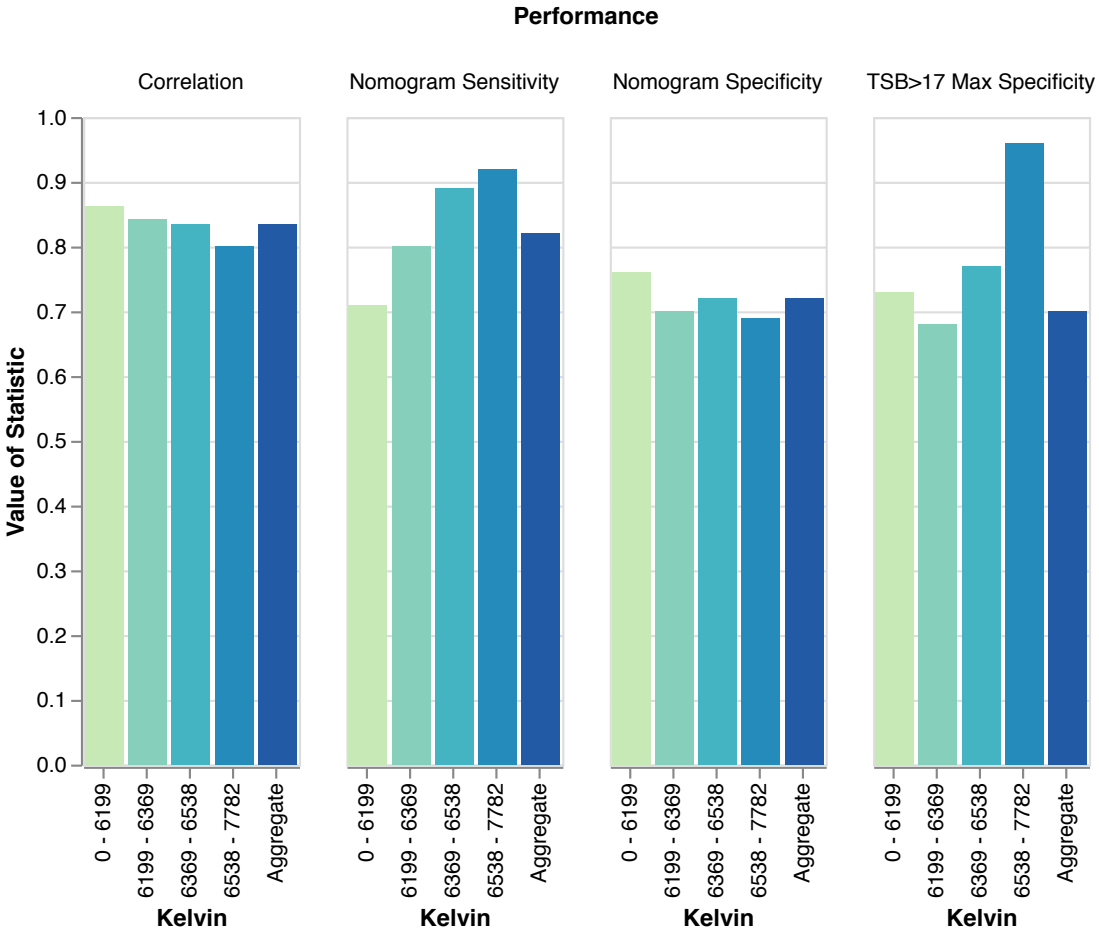


FIGURE 3.26: Bar charts illustrating BiliCam's performance for each CCT partition: the liear correlation, the sensitivity and specificity from using the Bhutani Nomogram decision rule, and the maximum possible specificity to achieve 100% sensitivity for predicting TSB >17.0 mg/dL. Note that the latter was tuned to different BCB cut-off values for each CCT partition. An additional bar shows BiliCam's aggregate performance.

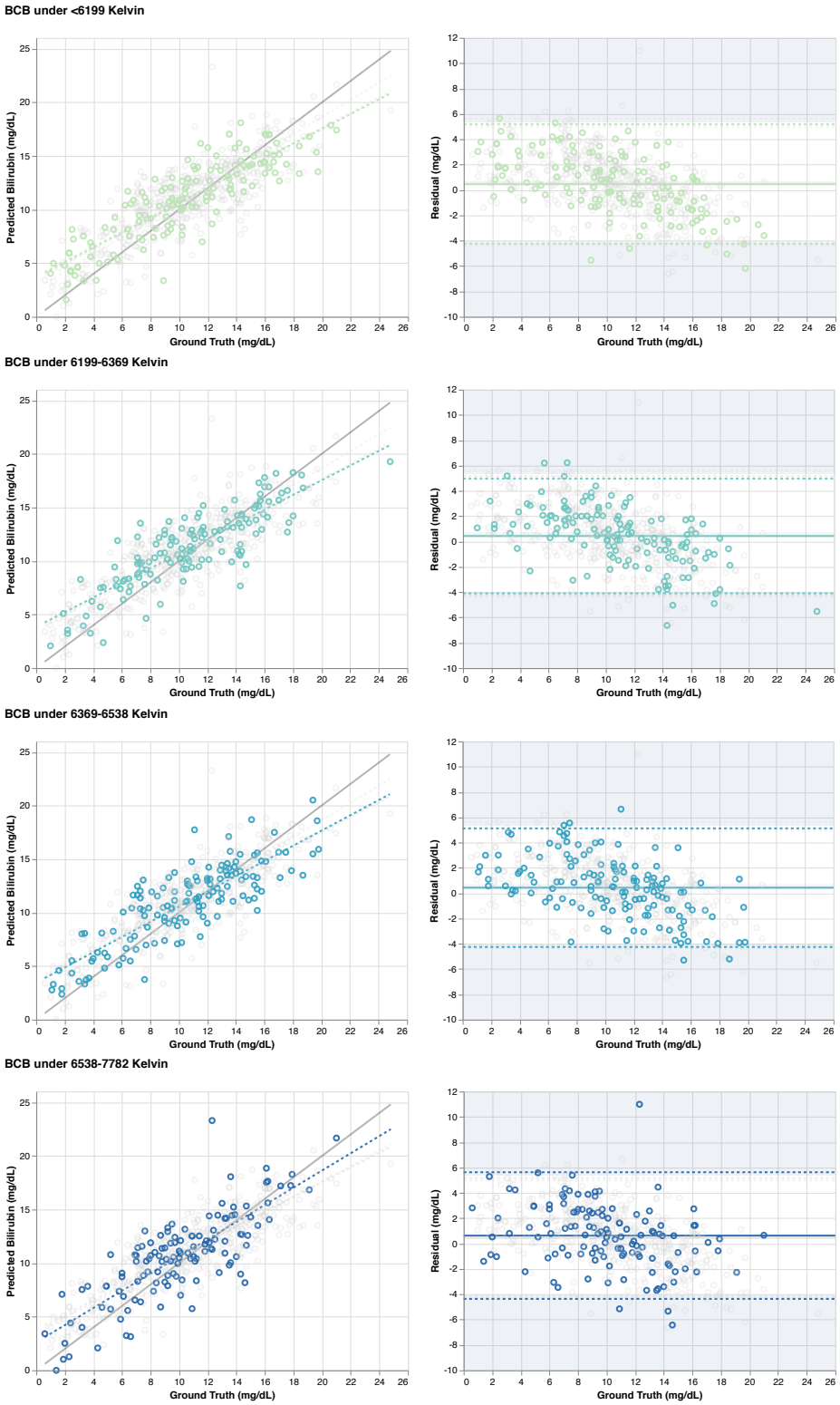


FIGURE 3.27: Regression results presented in a similar way to Figure 3.19, but separated by the four partitions of correlated color temperature. The shaded areas of the Bland-Altman plots refer to the limits of agreement for the aggregate BiliCam model, whereas the dashed lines refer to the limits of agreement for a particular CCT partition.

The results suggest that BiliCam does not have adequate accuracy to serve as a standalone methodology to assess jaundice in newborns. Rather, as with TcB meters, BiliCam is best suited as a screening device to aid in determining which neonates require a blood draw for a TSB level, with treatment decisions being based on the TSB level. Because of this, perhaps the most clinically relevant comparison between BiliCam and TcB is their utility as screening tools for identifying newborns who require a TSB level while obviating the need for a blood draw in most.

As discussed in subsubsection 3.6.2.4, one decision rule is to predict high-risk cases using the Bhutani Nomogram. For the data collected in this study, BiliCam yielded a higher sensitivity and lower specificity than the study's TcB. This trade-off may be preferable for BiliCam; TcB measurements happen in hospital contexts that benefit from the additional sensitivity introduced by visual assessment from trained staff, whereas the vision for BiliCam is to take measurements in contexts where people likely have little to no such experience.

The second decision rule discussed in subsubsection 3.6.2.4 is to use a predicted bilirubin cut-off level to identify newborns with a TSB > 17.0 mg/dL. As discussed in subsubsection 3.6.3.4, to achieve a 100% sensitivity in this dataset, BiliCam yielded at most a 69.9% specificity. Using a standard TcB cut-off at 13 mg/dL, the study's TcB achieves the same 100% sensitivity and a higher 74.8% specificity. Although BiliCam's specificity did not match the TcB collected in this study, it did outperform specificities of 30% and 58% for TcB documented in literature predicting TSB > 17 mg/dL [42, 82], as illustrated in Figure 3.21.

Overall, these results suggest that BiliCam may be as effective as TcB in identifying newborns in need of a blood draw for a TSB level.

3.7.2 Limitations and Future Work

There are important limitations that require more research before anyone can fully characterize BiliCam as an outpatient jaundice screening tool.

All of the data collection studies each used only one type of phone. While such a constraint is important for conducting a proof of concept, it is not tractable for BiliCam's real-world use (especially as the phones I used in the studies have since become outdated). Enabling BiliCam on other phones requires addressing differences in the phones' sensors, camera filters, gamma correction, any other firmware- and software-based image adjustments, focal distance, colors and brightness of the flash LED, and relative positioning of the LED and camera with respect to each other. All these factors can affect the collected data. I have yet to investigate the feasibility and necessary adjustments to make the system available on other devices.

The color calibration cards are another unexplored variable. Every card I used in the study was printed to high specifications from the same company on the same paper. The level of variation for ink, printers, and paper permissible for accurate results is yet to be tested. Requiring the card itself is also a limitation to the system – a method that does not require the card, or uses paper without requiring ink, can improve the accessibility of BiliCam.

Another limitation is that TSB levels were measured by using different assays from different laboratories. There is more variability in TSB measurements related to laboratory methodology than is generally appreciated. It is possible that the correlations between BiliCam predictions and TSB would have been higher if only a single laboratory for TSB assay had been used.

The generalizability of the results are also limited by the choice of study sites; I only collected data from medical clinics in the United States. This limitation means I cannot use this data to validate whether it would work in other settings, such as in the home or outside, that have different lighting conditions not represented in the dataset.

This corpus of data was also collected by the same set of people who were all briefed on the study, whereas an anticipated use case for BiliCam includes parents at home who would not develop this level of experience for collecting data. I cannot conclude whether there is a learning curve to using BiliCam that effects model accuracy.

Although the nation-wide study strove to collect data from a wide demographic, it still encompasses a limited demographic. White still represents more than half of all the self-reported races in the dataset. The dataset represents a very small proportion of some self-reported racial groups (i.e., American Indian/Alaska Native and Pacific Islander/Native Hawaiian). The study's racial grouping does not distinguish between groups that tend to have different skin tones like South Asian and East Asian. One approach to improving the representation of skin color in the dataset is to internationalize data collection, and make sure it targets different specific regions of the world that have high representations of skin tones that are different from each other and/or deficient in current datasets.

It is also worth noting that race is ultimately a social construct, not a definitive categorization of physical skin properties. This reporting method limits the ability to fully evaluate how well the data represents different skin tones – ultimately, the breadth of data I want to represent lies in skin color, not race. Instead of documenting self-reported race, another idea is to use methods for measuring skin tone (something that exists in beauty and cosmetics industries) for evaluating the diversity of a dataset.

An open question that future work must address before BiliCam can enter medical practice is how to present its measurement or estimation results and what decision rules to use for screening. Note that because the contexts and use cases for BiliCam is different from those of the TcB, it does not make sense to assume BiliCam should be used the

same way. For instance, how should BiliCam present its estimates to parents? Many parents know how to interpret a thermometer's temperature reading, not a bilirubin level. How would framing around the estimate take uncertainty into account? How does it guide parents into action without causing unnecessary stress, both to parents observing high estimates and to medical providers when BiliCam offers false positives? What actions should BiliCam recommend, and what decision rules should it base these recommendations on? How do answers to these questions differ when accounting for different cultures and regional needs? For example, differences in the availability of care, the effort required to access care, the degree to which jaundice screening is the limiting factor for adequate treatment (e.g., it is less relevant in regions that practice giving all newborns solar-based phototherapy), and other cultural practices for newborn health vary around the world. Future work needs to address these questions before BiliCam can be developed into a complete tool.

While the design of the nation-wide BiliCam study enabled some analysis on dimensions like photo distance and lighting conditions, there are futher unexplored variables that can inform BiliCam's design for better accuracy or usability. For example, a limitation to my analysis on photo distance is that they are constrained to only three distances in the dataset. The latest card design and image segmentation algorithm can enable data extraction from every frame taken of a video of the card, and with the right user interface, such a video can capture many more distances and even angles between the phone and the card or skin at little cost to the data collection experience. While my analysis on lighting conditions characterized color using the lightest grey patch on the BiliCam card, an open question for analysis lies in characterizing lighting conditions by how evenly it illumatines the card in any one sample. Does the lighting's evenness across the card correlate with better predictions? I designed the card to enable this analysis by including a light grey border along the outside of the card. Segmenting and analyzing the lighting distributions on this border can enable one way to characterize this evenness of card illumination.

Another option question is in determining BiliCam's physical limitations. I expect that the laws of physics (particularly optics) and the physiology of newborn skin limits how close BiliCam's predictions can align with ground truth measurements. Precisely how close BiliCam can be is yet to be determined, and could also be an interesting question to investigate in future work.

3.7.3 Impact

Neonatal jaundice poses a greater challenge in resource-poor areas of the world where the necessary medical tests are unavailable. In some countries, kernicterus is the second or third leading cause of newborn death, as well as an

important contributor to long-term disabilities [121]. Because of the seriousness of these complications, the social return on investment for this technology in resource poor areas is considerably elevated.

Because BiliCam requires no extra equipment besides a smartphone and the color calibration card, it has the potential to transform the outpatient management of jaundiced newborns. Health care providers evaluating newborns shortly after hospital discharge could use the technology to efficiently determine which infants require a blood draw for a TSB level. BiliCam could be used both in office settings and by nurses and other health care professionals evaluating newborns during home visits. Perhaps most importantly, in low- and middle-income countries with limited resources, BiliCam could be a low-cost technology that is used by health care workers to screen large numbers of newborns for jaundice and effectively identify the few that are at significant risk for life-threatening bilirubin levels. In combination with low-cost phototherapy devices that have now been developed [17, 25, 46, 61], BiliCam could thus be part of a system of care that could significantly reduce the morbidity and mortality related to high bilirubin levels in these areas.

To help make it a reality, I have guided the work on BiliCam into commercialization. I integrated steps toward commercialization through the work from the end of the formal local study and onward. During this time, I authoring two patents [130, 131], working with the University of Washington's Center for Commercialization (a.k.a. CoMotion), and initiated discourse and procedures with the FDA so that I could incorporate their insights and requirements in the work and begin the process for medical clearance. As BiliCam was one of the first mHealth projects to work with the FDA for clearance, it participated in helping the U.S. Food and Drug Administration (FDA) define how to approach clearance for such app-based medical devices. In line with the FDA's requirements, I actually collected more data in the nation-wide study than I used for analysis and reporting the findings in order to sequester a portion of the data exclusively for the FDA's final analysis on efficacy.

BiliCam became the pioneering app for a start-up called Senosis Health, which works on mobile health monitoring and develops apps to measure, diagnose, and manage diseases. Google has since acquired this start-up, and continues to develop BiliCam – conducting additional studies, investigating how to generalize to other phone models, experimenting with and evaluating new machine learning models, and other work to bring BiliCam to fruition. Members of the original BiliCam team continue their involvement on the project at Google.

Our work on BiliCam has also supported or opened doors for other research projects. A project called BiliScreen, for example, built on BiliCam to estimate jaundice levels in adults [83]. Insights and experience from working on BiliCam has informed study and technical design for ongoing projects at the UW Ubiquitous Computing Lab. I hope to further make some of these insights available outside the lab by documenting many of them in **??**, section 5.1 and section 5.2.

Another contribution is that this work on BiliCam played an active role in helping the FDA define their standards for approving software as a medical device (SaMD). By helping set these standards, BiliCam further supported and helped opened doors for the realization of other software-based medical devices.

Chapter 4

AI Opportunities for CHAMP: Helping Monitor Infants with Single Ventricle Heart Disease

CHAMP (Cardiac High Acuity Monitoring Program) is a system for monitoring infants with single ventricle heart disease during the critical months between their first and second heart surgeries. Through CHAMP, a hospital care team examines daily records of these patients, such as weight and oxygen saturation, that parents collect and upload on a tablet PC. While the creation of CHAMP in the last few years has dramatically decreased mortality rates, most patients still experience costly emergency room trips and unplanned hospital readmissions.

In this chapter, I discuss approaches for leveraging AI to help care teams discover acute medical complications sooner, enabling even more proactive interventions that could mitigate these issues. I also include results from an example machine learning model and a proposed way it could be used in the clinical setting.

My contributions include a design ethnography, early data analysis, and initial algorithm investigation to improve care and reduce the burden of a rare and complex disease. My early findings also changed the medical standard of care.

4.1 Background and Significance

This section provides necessary background information to understand and contextualize my work. It summarizes background on the medical condition, the CHAMP system, alternative approaches, and challenges in this research

space. I compiled this information through a combined approach of reviewing medical literature and applying HCI techniques (focus groups, semi-structured interviews, and contextual inquiry) to collaboratively unpack the expertise of medical experts.

4.1.1 Medical Background

4.1.1.1 Single Ventricle Heart Disease

Single ventricle heart disease is a rare congenital heart defect (i.e., present from birth) in which someone is born with only one functional ventricle in their heart. A heart normally has two ventricles: one to pump blood to the lungs, the other to the rest of the body. They separate oxygenated from deoxygenated blood into serial circulation (i.e., blood flows through the lungs and body in series). Patients with single ventricle heart disease, however, cannot separate blood this way. To survive immediately after birth, oxygenated and deoxygenated blood mix in their sole functional ventricle and result in parallel circulation (i.e., it connects the lungs and body in parallel). This mixing is enabled by a hole between left and right chambers of the heart (known as a septal defect) and/or connections between pulmonary and systemic arteries.

It is worth mentioning that there are many types of single ventricle hearts. The fact that each patient has a unique heart anatomy and therefore unique progressions, complications, and needs, makes this condition even more challenging to both treat and monitor. To illustrate how much these anatomies can vary, here is the list of variations documented in the CHAMP database (summarized again in Table 4.2): double inlet left ventricle, double inlet right ventricle, double outlet right ventricle with leftsided stenosis, double outlet right ventricle with pulmonary outflow obstruction, hypoplastic left heart with aortic and mitral atresia, hypoplastic left heart with aortic and mitral stenosis, hypoplastic left heart with aortic atresia and mitral stenosis, hypoplastic left heart with aortic stenosis and mitral atresia, pulmonary atresia with intact ventricular septum, tricuspid atresia, unbalanced AV canal (RV dominant or LV dominant), and single ventricle otherwise not classified (RV morphology, LV morphology, or unknown morphology). Figure 4.1 illustrates the most common variant is hypoplastic left heart syndrome (HLHS), where the left side of the heart is underdeveloped (i.e., hypoplastic).

4.1.1.2 Treatment and Risks

Single ventricle heart disease is a rare condition; different sources estimate its prevalence to be anywhere from 6.1 per 100,000 live births [124] to 8 per 10,000 live births [91]. This disease's complexity and rarity also necessitates highly

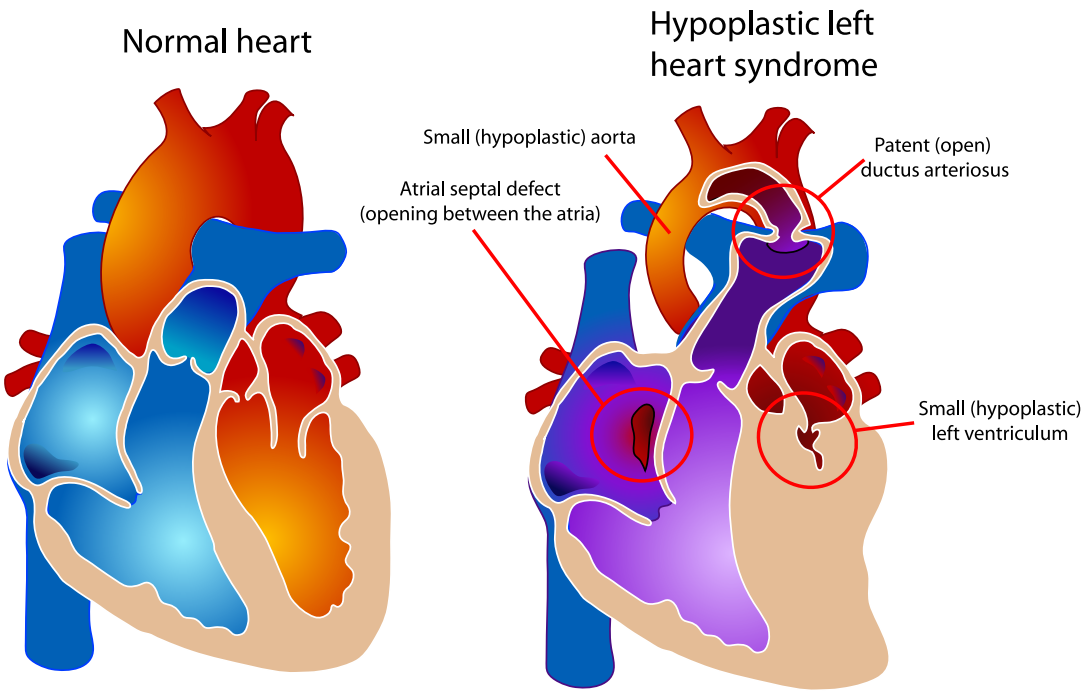


FIGURE 4.1: An illustration of Hypoplastic Left Heart Syndrome (HLHS), the most common type of single ventricle heart disease. Compared to a normal heart, the left side of an HLHS heart is underdeveloped (hypoplastic) and unable to pump blood. An opening between the left and right atria (an atrial septal defect) and a connection between the aorta and pulmonary arteries (patent ductus arteriosus) enable the mixing of blood necessary for initial survival. Attribution: "Hypoplastic left heart syndrome" by Mariana Ruiz LadyofHats is under public domain.

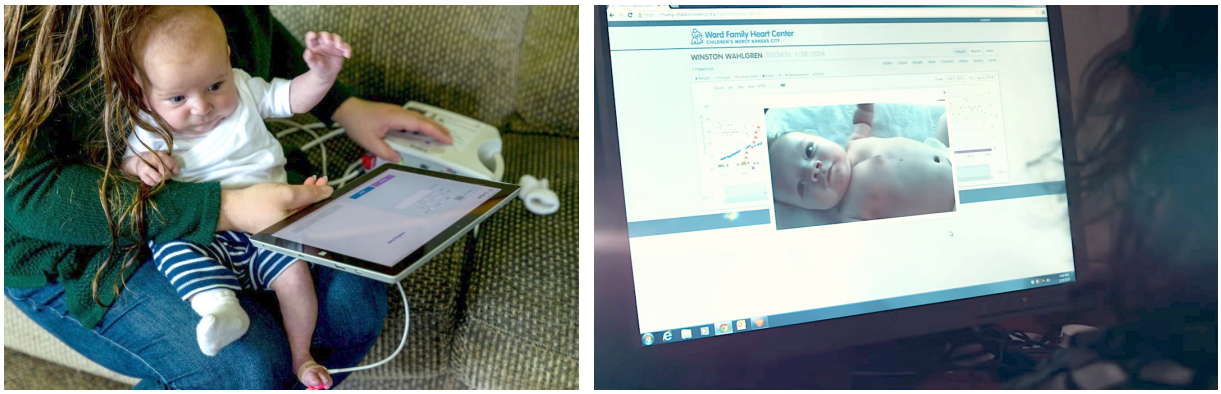
specialized skills and surgical training for any chance of survival. As a result, few hospitals can actually treat this condition. Given that the world's population is not clustered around these hospitals, most families live far away from them, making the accessibility offered by a remote monitoring systems all the more important.

Treatment for single ventricle heart disease involves three specialized stages of cardiac surgery spaced over multiple years. Stage I, which creates or reinforces the openings and connections to enable parallel circulation, happens within the first week of life to stabilize the patient so they can survive long enough for stage II. A common stage I operation is called the Norwood procedure. Completion of stage I has a 10-20% mortality rate. Four to six months typically need to pass after stage I for the patient's body to sufficiently grow and mature to handle stage II, which usually is the Glenn procedure. The period between discharge after stage I and the completion of stage II is called interstage, and historically has a 15-20% mortality rate [19] – the highest among common congenital heart procedures [117]. Patients usually spend the majority of this critical period at home. After stage II, patients continue to grow and mature their bodies over the course of 3-4 years in preparation for stage III. The mortality rate during this period is 2-4%. Hence, interstage is the most dangerous the period and requires the most monitoring at home. My work focuses on interstage patients for this reason.

4.1.2 CHAMP

Cardiac High Acuity Monitoring Program (CHAMP) is a remote monitoring program for interstage patients with single ventricle heart disease, developed by a team at the Kansas City Children's Mercy Hospital in 2016 (CMH) [117]. Over the course of its two years of use with 84 patients at CMH thus far, it has demonstrated its ability to help reduce interstage mortality rates to 4% – a significant improvement over the previous 15-20%. Because this tool is too new for FDA approval, it is currently a study tool. At least seven other hospitals have adopted CHAMP by enrolling in the study to improve interstage outcomes for their patients.

This program provides caregivers with devices like infant weight scales and pulse oximeters, which they use to capture monitoring data at home. CHAMP data includes oxygen saturation, feeding intake and output, infant weight, 15-second videos, and parental concerns. They record the data on a provided tablet PC through a dedicated CHAMP app that provides instantaneous data transfer to a central database. A medical care team can then access the data, which they review remotely to check the patient's health status. The 15-second videos prove to be invaluable; care teams have used them to notice warning signs like chest retractions or other signals of respiratory distress for which the caregivers do not have as much experience or training.



(A) A parent and child using CHAMP. (B) A care team member reviewing CHAMP data.

FIGURE 4.2: Photographs of CHAMP in use. (A): a parent using CHAMP with their interstage child. An infant pulse oximeter provided to the family wraps around the child's foot to measure their blood oxygen saturation while the parent prepares to log the resulting value on a CHAMP tablet. (B): a care team member reviewing CHAMP data. Part of a 15-second video of the child is in the foreground of the CHAMP interface and a graphical representation of physiological measurements like weight and oxygen saturation is in the background.

Before the initial discharge from the hospital, the care teams teach caregivers about what red flags to watch out for. Red flags include numerical values from CHAMP data (e.g., an oxygen saturation below 70%, or not gaining an average of 20 grams per day for 7 days) and qualitative signs (e.g., increased sweating during feeds). The CHAMP app lists these red flags to help remind caregivers to look for them and contact the care team whenever one arises. As an alternative to calling the care team, the app also offers a way for parents to submit concerns through selecting red flags or write explanations via free response. This method of communication, while not as common as calling, is helpful for non-English speaking caregivers because the app lets them select concerns in their native language.

The digitization and central storage of CHAMP data also provides the ability to instantly alert the care team about specific issues. Currently, any data that meet or exceed a preset threshold trigger an instant alert. For example, if a caregiver logs an oxygen saturation that is lower than 70%, the care team would receive an instant alert. Table 4.1 lists all current instant alert thresholds.

The team that created CHAMP developed these instant alert criteria through a combination of using medical intuition and slight modifications over the brief course of CHAMP's existence. Before CHAMP, there were no prior guidelines for these criteria; this resolution of interstage data collection and transmission has not exists before CHAMP and, as I describe in subsection 4.1.4, any guidance from literature on monitoring patients is relatively sparse. Currently, they only account for the detection of 10% of complications that resulted in hospital readmissions.

The detection of medical complications when instant alerts do not occur depends on caregivers noticing warning

Current instant alert criteria
Oxygen saturation <70%
Oxygen saturation >94% (for Norwood patients)
Oxygen saturation >97% (for the other patients)
3 emesis (vomits) in 24 hours
3 diarrhea stools in 24 hours
Heart rate <80 or >180 bpm
Any parent concern

TABLE 4.1: Current instant alert criteria. The entry of CHAMP data that meet any of these thresholds trigger an instant alert that automatically pages the care team.

signs or on the care team reviewing data on their own schedule. In the most recent study, more than half of the initial unplanned readmissions were found by a medical team reviewing CHAMP data instead of the caregiver [19], highlighting the importance of the care team's attention. Ideally, they would detect warning signs before patient condition becomes severe enough to warrant emergency room visits or hospital readmission. Many care teams have hectic schedules as part of their job in hospitals, and reviewing CHAMP data can take too significant of an amount of time to manage at the time of each data entry. Some care team members are able to block out multiple sections of time over the course of their day to review data, while others may have demanding jobs that ask them to review data on their own time (e.g., coffee breaks or after work). In the latter case, sometimes a patient's data will not be looked for more than 24 hours. A result of the current timing for care team data review is a need to assist with patient triage; ideally, patients at higher risk of having a medical complication would receive attention first – especially given how interstage complications can dramatically worsen in a matter of hours.

4.1.3 Other Monitoring Approaches

Typical home monitoring for interstage patients without CHAMP also provides families with a weight scale and pulse oximeter. Instead of using a tablet PC to record any monitoring data, however, they would record daily oxygen saturation, heart rate, weight, and feedings in a notebook or three-ringed binder. A phone call would connect these families with a care team at least once per week to communicate these records and anytime a red flag issue arises [107]. This method of home surveillance monitoring was introduced in 2002 [50]. CHAMP improves on this method by enabling immediate data transfer and review, as well as the ability to send 15-second videos of the patient. These videos

prove to be invaluable, because the care team has used them to notice warning signs like chest retractions or other signs of respiratory distress for which the caregivers do not have as much experience or training.

One publication demonstrates an approach for using machine learning and preexisting sensing hardware to monitor infants with parallel circulation that can predict impending deterioration events within one to two hours [109]. While it also focuses on patients with single ventricle heart disease before stage II surgery, it monitors patients in a hospital's intensive care unit (ICU). This environment enables continuous bedside monitoring with hospital equipment for real-time physiological signals like beat-to-beat heart rate variability, respiration rate variability, peripheral oxygen saturation, the ST segment of electrical heart impulses, atrial blood pressure, core temperature, toe temperature, respiration rate variability, and more. In short, this work was able to leverage a number of sensors and hardware that are not feasible for the outpatient environments I target. Despite these differences, this work offers insights that inform the development of my own methods, such as an approach to categorizing the health status associated with retrospective medical records.

4.1.4 The Challenging Nature of this Research Space

A major challenge for developing methods to monitor interstage patients lies in the intersection of how they differ from healthy infants, the heterogeneity of this patient population, and the sparsity of related work.

Because of how minimal stage I surgery needs to be, interstage patients pump blood very inefficiently. Their resulting baselines vital statistics are very different from those of most infants, rendering existent monitoring metrics unsuitable. A normal oxygen saturation for them may seem dangerously low for other infants. Many of them present a baseline level of cyanosis (bluish skin discoloration from poor circulation or low oxygen levels), unlike healthy babies for whom the mere presence of cyanosis should be alarming. A typical growth rate for these patients also tends to fall below the averages for healthy infants, and as many of the interstage patients have a feeding tube, they also have a tendency to spit up more than healthy infants.

An important implication of these differences is that the literature and guidelines for monitoring most infants do not translate well for monitoring interstage patients. The heterogeneity of the patient heart anatomies (as illustrated in Table 4.2) also affect what signals are normal or distressing on a per-patient level. Furthermore, the rarity of this disease heavily limits the existence of related literature. A result of these limitations, "for this complex population there is little consensus as to what constitutes 'normal' or even 'acceptable' for a given physiologic parameter" [109].

Types of Single Ventricle Heart Anatomies in CHAMP Database
Double inlet left ventricle
Double inlet right ventricle
Double outlet right ventricle with left-sided stenosis
Double outlet right ventricle with pulmonary outflow obstruction
Hypoplastic left heart with aortic and mitral atresia
Hypoplastic left heart with aortic and mitral stenosis
Hypoplastic left heart with aortic atresia and mitral stenosis
Hypoplastic left heart with aortic stenosis and mitral atresia
Pulmonary atresia with intact ventricular septum
Tricuspid atresia
Unbalanced AV canal (RV dominant or LV dominant)
Single ventricle, otherwise not classified (RV morphology, LV morphology, unknown morphology)

TABLE 4.2: A list of different types of single ventricle heart anatomies documented in the CHAMP dataset to illustrate the heterogeneity of the patient population. Note that more anatomies exist than listed here, as expressed by the last row of the table as "otherwise not classified."

With the additional reasons that the disease is so rare, surgical techniques continue to evolve, and that regular digital records for these patients were introduced only a couple of years ago, there is little data on what physiological signals should be cause for concern. The heterogeneity of the patient heart anatomies also affect what signals are normal or distressing on a per-patient level. The resulting lack of empirical evidence on both the population-level as well as for each heart anatomy makes it incredibly challenging to develop guidelines on how to monitor these patients.

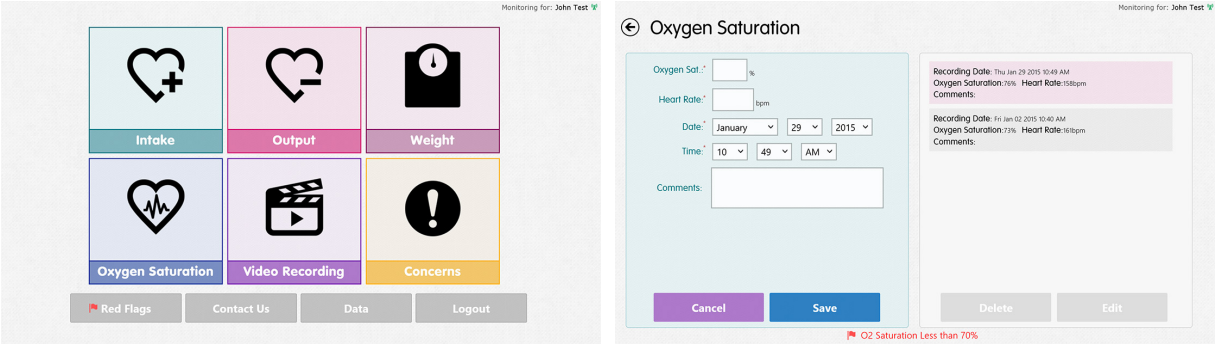
The nature of CHAMP studies themselves introduces additional challenges for distinguishing control groups. They began a study as a randomized controlled trial, in which they randomly assigned some families to using CHAMP after one month and others after 2 months of the alternative binder-based approach. During an interim analysis in their studies, they already began to uncover evidence of CHAMP's stronger efficacy for detecting medical complications. While continuing the separation of these experimental and control groups would simplify data analysis, they prioritized patient safety. They terminated the randomization early and switched all of their patients to CHAMP.

The nature of CHAMP also makes it challenging to disambiguate control statuses on the level of individual patients. CHAMP enabled earlier and subtler interventions; with its available data, the care team interpreted minor warning signs and prescribed medication or feeding changes as early interventions. It is possible that they were able to prevent a number of complications from reaching the severity that warrants a hospital readmission, which is a boon for the patients. On the other hand, the data cannot provide evidence for whether their interpretations were correct or how the interventions altered the course of the patient's health. These ambiguities make it more difficult to delineate different health status for patients in the database.

4.2 Materials

The materials available at the time of writing this thesis proposal consist of data carefully collected from 84 patients at CMH over first 2 years of CHAMP's existence. I categorize the data into two overarching types: "tablet data" and "forms data". I also introduce several important limitations and challenges with this dataset in this section, and discuss more of them in subsection 4.5.2.

4.2.1 Tablet Dataset



(A) The home screen (B) Interface for logging oxygen saturation

FIGURE 4.3: Screenshots of the CHAMP tablet interface for caregivers to record monitoring data of their interstage child.

Tablet data encompasses all data that caregivers record with their CHAMP tablets, using the interface shown in Figure 4.3. Caregivers capture measurements of weight, oxygen saturation, and heart rate using the medical devices provided to them by the hospital. The inclusion of heart rate was introduced halfway through the two-year recordings of CHAMP, so approximately half of the participants represented in the database do not have home heart rate data. Caregivers were asked to record their child's intake every time they feed. Records of feeding intake include two components: the quantity and whether feeding was through bottle, breast, or tube. Note that they cannot know the quantity for intake through breast feeding. They similarly are asked to record their child's output (excrement), which includes urine, stool, or vomit. In the case of stools, they also mark check-boxes for stool quality, such as whether it was watery, red, pale, and so forth. Finally, tablet data includes 15-second videos of the baby, disrobed to show their chest and surgical scar, and any parental concerns logged through the app.

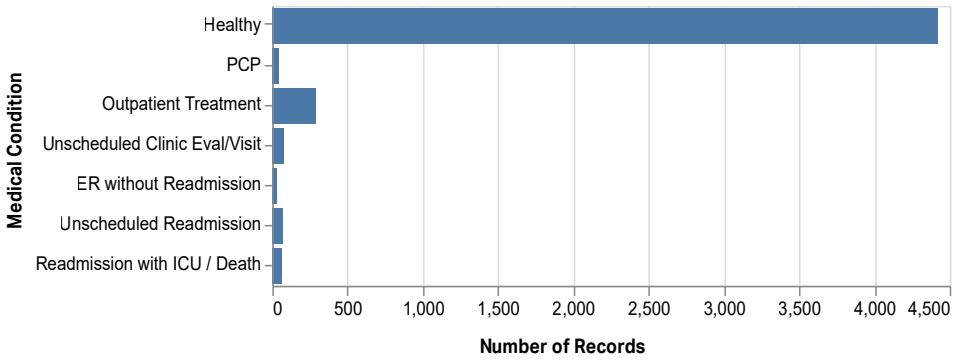


FIGURE 4.4: A bar graph showing the distribution of CHAMP medical conditions in the dataset.

4.2.2 Forms Dataset

CHAMP care teams maintain detailed documentation for every family, visit, and major medical event through a number of different survey-like forms. These forms cover enrollment (e.g., demographics, birthweight, heart anatomy), neonatal surgery and hospitalization, neonatal discharge, instances of red flags, clinic visits, readmissions, Glenn surgery, and death/withdrawal.

Among the hundreds of pieces of information recorded in the forms data, they have records of different medically-related events that serve are the available indications for different states of health. Records of death on the death/withdrawal form correspond to the most severe medical condition. Records on forms for clinic visits that are marked as "unplanned" are an indication of needing medical attention. Similarly, records on the readmissions form marked "unscheduled" indicate the need for severe medical attention. They also have records for how many days the patient stayed in the ICU during that readmission, which can serve as an indicator of the severity of this readmission. The form documenting cases of red flags include a section for the care team's primary recommendation to the family. These recommendations ranked in order of least to most severe are: Review and reassurance (review the data, reassure caregivers it's okay), Visit principal care physician (PCP) (e.g., if it looks like the child has a non-CHAMP related problem), Outpatient treatment (i.e., make feeding or medication adjustments), Cardiology clinic evaluation (i.e., an unplanned clinic visit), Emergency room, and Readmission.

4.2.3 Dataset Challenges

Several challenges from this dataset come from its small size, its noisiness, and how sparse much of the data is. The combination of this disease's rarity and the relative newness of the CHAMP data collection tools results in a small dataset of 84 patients. The smallness of this number is particularly significant in light of how complex and heterogeneous this disease is and the hundreds of possible machine learning features available for each patient.

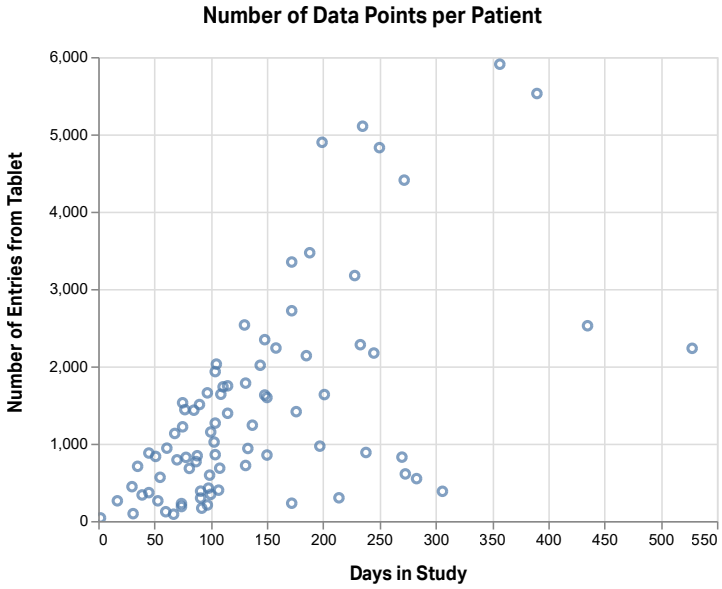


FIGURE 4.5: A scatter plot showing the number of tablet data points against time spent in the study for each patient. The number of days in the study was calculated as the time between the first and last data point. This chart showcases how inconsistent and noisy the tablet data are.

The tablet data is also very noisy. They were recorded by caregivers who are already overwhelmed by not just having a newborn child but also the added stress of caring for a child who is so ill. A common result of this overwhelm is to stop recording data – sometimes skipping just the intake and output because the CHAMP care team anticipates the other information being more valuable, sometimes skipping days of data collection entirely. Because all the measurements are taken in a home setting by people who do not have the same degree of experience as their professional counterparts, there may be additional inconsistencies in the calibration of individual measurements. Both the tablet data and forms data also can contain mistakes. I made a number of data corrections by hand, but the extent of these corrections is limited by what a CS researcher was able to intuit. The database may still have uncaught data errors.

Figure 4.5 showcases how inconsistently caregivers log tablet data. Should all caregivers log tablet data at the same rate, all the points in the scatter plot should fall onto a neat line. Instead, there is a wide spread of data points on or below this theoretical line.

4.3 Applications

My aim is to investigate whether and how machine learning can be integrated into CHAMP to assist the care team with detecting acute medical complications or triaging the patients they review. My approach consists of collaboratively uncovering use cases for machine learning in this context and developing an example machine learning model to illustrate how to approach one of these use cases (the latter of which I discuss in the next section).

I worked closely with the CMH and SCH medical care teams to uncover pain points in their using of CHAMP, and iterate on actionable applications of machine learning that could help address these pain points. This process involved applying both HCI techniques (i.e., a design ethnography involving focus groups, contextual inquiry, many semi-structured interviews to unpack this space) and an understanding of machine learning to assess technical feasibility. Analyses of the CHAMP dataset also inform these applications.

Below are two of the most prominent categories of pain points and use cases I uncovered that machine learning could potential address with the non-video tablet data. I also describe additional opportunities for leveraging machine learning with the video tablet data, although work on these applications require more resources (time, ground truth labels, and more) to investigate and are out of scope for this dissertation.

4.3.1 Emergency Room Visits and Unscheduled Hospital Readmissions

Interstage patients undergo emergency room visits or hospital readmissions when their medical conditions worsen to the point of requiring the skills (e.g., surgery) and equipment (e.g., echocardiograms) only available at hospitals. For these patients, hospital readmissions are what can prevent fatalities when their condition reaches this level of severity. Because any severe medical complication they experience can dramatically worsen in a matter of hours, quick identification of these conditions is paramount to their health outcomes.

Sometimes caregivers recognize the red flags associated with these acute conditions, and reach out to the care team for immediate review and intervention. However, more than half of the initial unplanned readmissions in one CHAMP study were found by a medical team reviewing CHAMP data instead of the caregiver [19], highlighting the

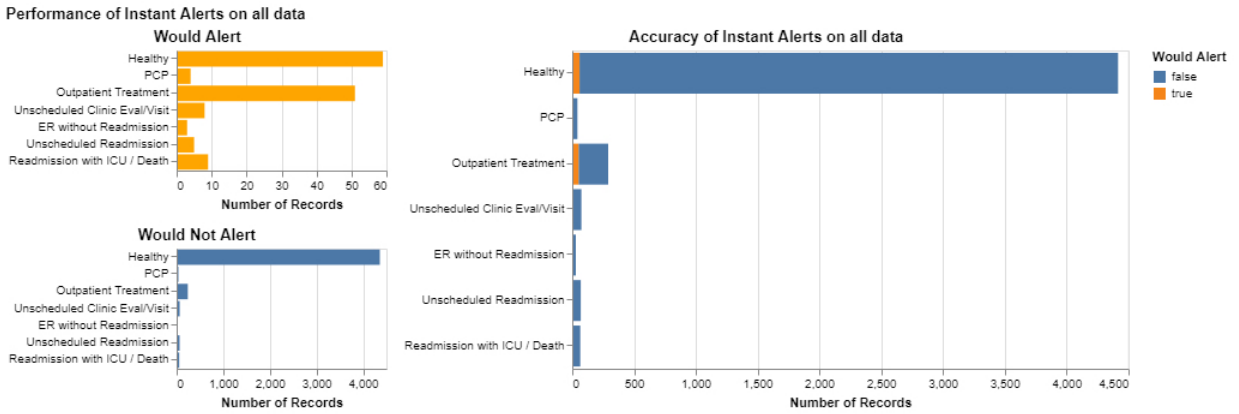


FIGURE 4.6: Bar charts showing the current performance of instant alerts. The bars are color coded by orange for instances that had instant alerts and blue for instances without instant alerts. Top left: the distribution of health statuses at the time of instant alerts. Bottom left: the distribution of health statuses when there were no instant alerts. Right: The combined distributions of health statuses with and without instant alerts. Note that I want instant alerts for the health statuses of "ER without Readmission", "Unscheduled Readmission", and "Readmission with ICU / Death", and do not want them for the statuses.

importance of the care team's attention. Instant alerts currently serve as the work-around to reach care teams quickly when caregivers may not have caught something amiss.

My analysis indicates that these instant alerts corresponded with the detection of approximately 10% of instances when CHAMP patients required emergency room visits or hospital readmissions (i.e., a sensitivity of 10%). 90% of the instant alerts did not correspond with these acute conditions. Figure 4.6 illustrates a breakdown of instant alert performance. These low numbers indicate that there is a lot of room to improve the accuracy of instant alerts.

Past experiences from the CMH care team demonstrate how challenging it is to improve the instant alert criteria. They formerly used criteria that had wider bounds for thresholds; multiple vomits over only 12 hours instead of 24 hours used to trigger an instant alert, as did an oxygen saturation above 94% for any patient – not just Norwood patients. While the looser oxygen saturation threshold resulted in a higher true positive rate of 12%, its false positive rate of 5% was higher than the care team could manage due to alarm fatigue. By adjusting the thresholds to what they are today, they were able to drop this false positive rate to a more manageable 2.5%. I illustrate this difference in Figure 4.7.

Because accurate instant alert criteria has proven to be very difficult to develop by humans alone, it is one area which machine learning has the potential to help. Classification algorithms that introduce additional or alternative set of heuristics that improve the instant alert's sensitivity without reducing its specificity could help care teams identify readmission-worthy medical complications sooner.

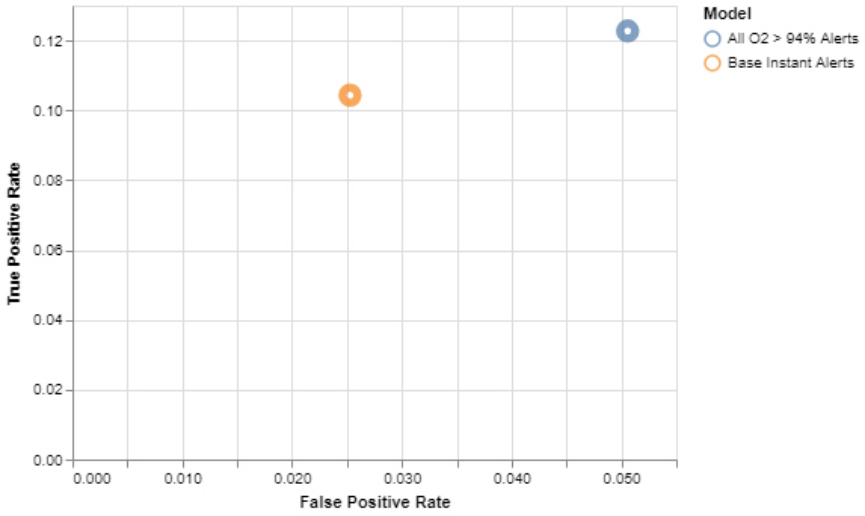


FIGURE 4.7: A chart comparing the performance of an old and the current instant alert criteria.

4.3.2 Patient Prioritization

While hospital readmissions may save a patient's life, they are also costly, distressing, and are the result of high-risk states of health. Because many families life far away from the hospitals that are capable of operating on single ventricle heart patients, the commute can be especially costly. For example, there is a family that lives a 7-hour drive from such a hospital. When there was a medical emergency, they and the care team decided that they needed to send a helicopter to bring the patient to the hospital as quickly as possible.

Earlier, softer interventions have potential of lowering the likelihood of needing unscheduled hospital readmissions. Care teams want to prioritize their attention first to patients who may need emergency room or unscheduled hospital visits. For the remaining patients, they want to prioritize those who need these earlier interventions.

From interviewing nurses at different hospitals, including one who is the data manager for multiple sites, I learned that the way care teams can prioritize patients vary from hospital to hospital. Some hospitals, like Seattle Children's Hospital (SCH) have fewer patients than CMH and ask caregivers to record less data (i.e., no intake or output, and only one measurement of oxygen and heart rate per day instead of CMH's two measurements per day). The SCH care team have sufficient availability to review data points as they come in during the work day, and usually have enough time to review data that came in outside of these hours as their first task in the morning. In contrast, another hospital has different constraints on staffing, data quantity, understanding of CHAMP, and work culture. Rather than offer nurses dedicated time to examine CHAMP data on a daily basis, they ask nurses to review the data on their own time (e.g.,

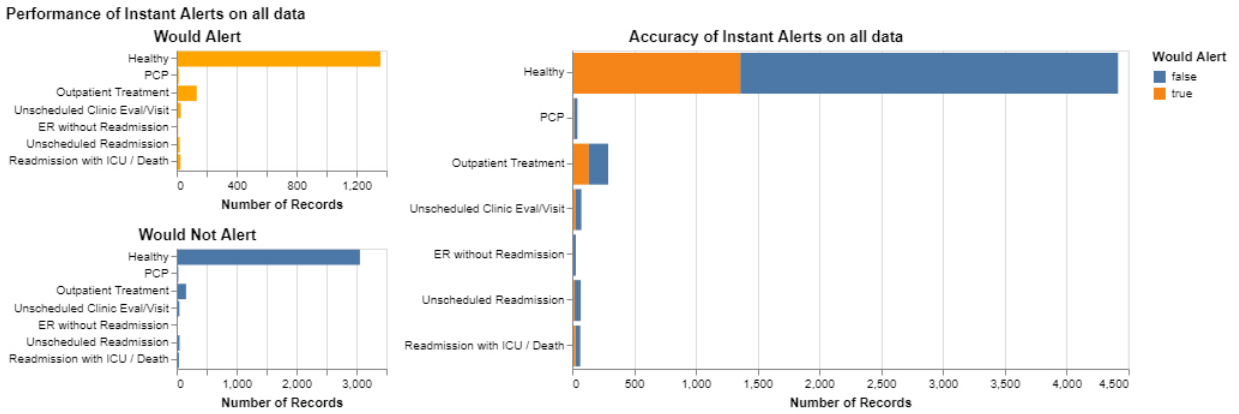


FIGURE 4.8: Bar charts showing the current performance of automated red flags. The bars are color coded by orange for health statuses that had red flags and blue for those without. Top left: the distribution of health statuses at the time of red flags. Bottom left: the distribution of health statuses when there were no red flags. Right: The combined distributions of health statuses with and without red flags. Note how many red flags correspond to patients who do not need CHAMP-specific medical attention (i.e., "Healthy" or "PCP").

during coffee breaks or after work) to manage their resources. An unfortunate effect of these constraints is that these nurses do have the bandwidth to consistently review every patient every day. CMH's process falls in between in that their nurses do have time built into their duties to review CHAMP data, but there are too many patients and incoming data points to feasibly review them upon receiving them. One of the nurses describes her process as reviewing several patients at a time in multiple chunks spread across the day. Some patients may not have their data reviewed until half of the day has gone by, and the order in which to review these patients is often not obvious.

Particularly for situations with a high data to resource ratio, patients and their care teams can benefit from assistance with triaging which patients to examine first. The currently system offers "red flag" markers for any data that corresponds to instant alerts plus two more criteria (whether a patient gained less than 20 grams per day over 4 days, and whether they had fewer than 4 wet diapers in 24 hours). Figure 4.8 shows how red flags compare to the health statuses they try to help prioritize. The majority of these red flags mark healthy patients, so red flags do not fully disambiguate which patients to prioritize.

Machine learning algorithms can search for alternative criteria for prioritizing patients. Classification algorithms can try to distinguish patients who need any kind of CHAMP-related medical attention from those who do not. Regression algorithms can try to estimate "risk scores" to aid triage, much like those described in subsection 2.1.2.

4.3.3 Video Analysis

The time spent analyzing the 15-second CHAMP videos is a significant contributor to how long it takes a care team to review CHAMP data. Although each video is only 15 seconds long, it takes multiple views to estimate important metrics like respiratory rate or carefully search for changes qualitative signals like indications of respiratory distress. This process usually involves switching back and forth with several previous videos to establish a relevant baseline for comparison. Cyanosis and other color-dependent properties are also very difficult to evaluate as these videos do not have consistent lighting.

Non-exhaustive examples of what features the care team looks for include: respiratory rate, nostril flare, chest retractions, eye gaze, the general amount of movement, the degree and locations of cyanosis, skin texture (e.g., for mottling), sweat, wheezing, and facial expressions.

Because all the signals available through video data is only accessible through professionals viewing the videos, they currently cannot form the basis for any instant alerts or red flags. Computer vision, signal processing, and machine learning techniques to automatically extract information from these videos have the potential help in this respect.

By working together with the CMH care team, I identified which of these features were the most useful and tractable from both a medical and technical perspective. They are estimating the respiratory rate, detecting chest retractions, and evaluating cyanosis. Be warned that each of these features are very challenging to detect, even for humans. With respiratory rate, the ability to track the inhale and exhale of very breath can also offer insights to breathing variability within the 15-second clip. A healthy newborn has a very inconsistent respiratory rate, so a very even breathing rate may signal a certain level of stress. Chest retractions can occur in different parts of the body, with each location suggesting a different level of severity. A baseline level of cyanosis is common for many of these patients. Instead of the presence of cyanosis in general, it would be useful to evaluate a potential change in cyanosis and/or whether cyanosis becomes apparent around the nose and mouth (a sign of poor oxygenation in the body, versus only poor circulation in the extremities).

Note that to train any algorithms to accurately extract video features, they need ground truth labels for training and validation. These videos currently do not have ground truth labels. As there are hundreds of CHAMP videos which are only accessible to the care team and a handful of researchers who are not trained to identify these features, it would take a non-trivial amount of human resources from these busy medical staff members to create ground truth labels. These concepts are important to keep in mind for any ventures into developing such video-based tools.

4.4 Example Method

To demonstrate how machine learning can be applied to these use cases, I present an example method for how to approach the use case of patient prioritization. In this approach, I apply classification algorithms to find an alternative method of triaging than CHAMP's current automated red flags. I also discuss how this approach can be applied to classifying medical emergencies, though the limitations described in subsection 4.5.2 of this budding dataset prevent conclusive analysis for this use case.

4.4.1 Machine Learning Targets

I collaboratively worked with members of the CMH team to define ground truth labels for the machine learning model. The limitations of the CHAMP dataset requires some assumptions for defining machine learning targets, which are also documented below.

4.4.1.1 Medical Complications

In subsection 4.2.2, I discussed what information in the database can be interpreted as indicators of health and health complications. I worked with the CMH team to uncover 9 indicators located in 4 different tables of the CHAMP dataset and collaboratively determined the following ranking of severity for these indicators:

1. Review and reassurance
2. Visit principal care physician (PCP)
3. Outpatient treatment
4. Cardiology clinic evaluation / unscheduled clinic visit
5. Readmission within 1-2 days
6. Emergency room with no readmission
7. Readmission
8. Readmission with days in ICU
9. Death

It is possible that the care team evaluated a patient to have a lower severity than the actual severity of the patient's medical status, which would manifest with records of a higher-severity evaluation within a day after. To account for these possibilities and to prevent the over-representation of data points from patients with multiple events within a

brief time frame, I combine all medical complications that were within 24 hours of another and user the most severe evaluation as the primary label for that cluster.

Note that the delay between medical complications presenting themselves and timestamps for documentation are likely inconsistent. Also note that much of this documentation is based on recommended courses of action, which can only be a proxy for a patient's health status.

4.4.1.2 States of Health

To define which points patients should be considered healthy, I made that assumption that any period of time that is sufficiently far away from any medical complications is a period of relative health. I define "healthy" states of being to: have no events 1 day before or 3 days after, not be during a readmission, and not be within 5 days of Glenn surgery (as this surgery is sometimes an intervention for several days of physical decline). To avoid over-representing the same data, I chose to equally space all targets for healthy periods to be 2 days apart.

4.4.1.3 Grouping

In collaboration with CMH, I define complications of severity ranked 3 or higher to be relevant for prioritizing patients. Similarly, complications of severity ranked 6 or higher are relevant for instant alerts.

4.4.2 Extracted Features

Through conversations with CMH and SCH, I developed the following machine learning features:

- Weight: rate of change, weight-over-age z-score
- Oxygen saturation: min, max, average, range, trends, percent changed
- Heart rate: same as oxygen but do not have enough to properly evaluate its utility yet
- Intake: quantity in terms of mL/kg/day (not useful when breast fed), whether the patient was tube-fed
- Output: urine frequency, vomit frequency, diarrhea frequency, the presence of red or pale stools
- Heart anatomy/condition: type of surgery, level of ventricular dysfunction, level of AV regurgitation, stress-related hormone levels, age at discharge, weight gain before discharge
- Age: overall age, gestational age, days since initial discharge
- Caregiver context: levels of education, English speaking or not
- Recommended nutrition route: feeding tube vs oral vs both

I did not use the 15-second videos as they do not have ground truth labels for any of the features I would want to extract or evaluate. I also did not use the concerns submitted through the CHAMP tablet because very few of them exist (most concerns were shared through direct phone calls). These tablet-transmitted concerns also already automatically trigger an instant alert.

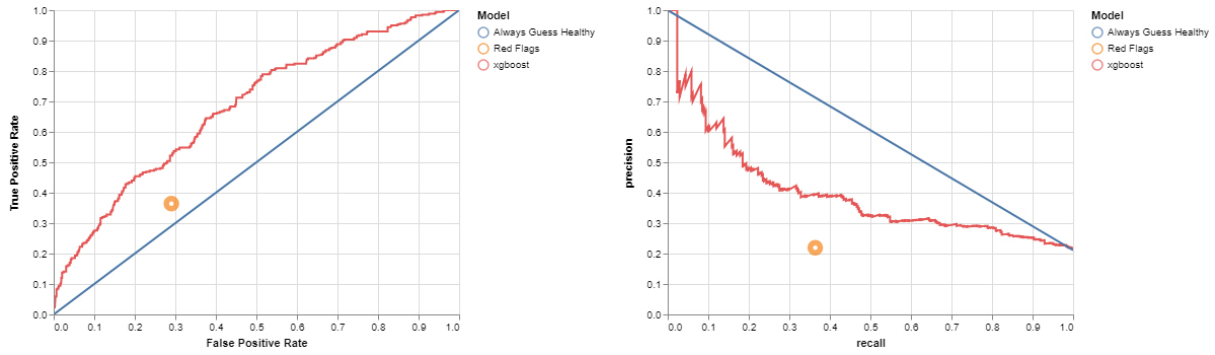
4.4.3 Models

I focused on using gradient boosted classifiers using the XGBoost framework [31] for two main reasons: it can handle CHAMP's many missing values (a major limitation of the dataset) without imputing missing data (an approach that favors smoother and more predictable data) and it offers a degree of explain-ability for its decisions (a "white box" model can better help medical teams than any "black box" model). Favoring any machine learning model that cannot handle missing values results in a dataset size that is too limited to draw meaningful conclusions. A future, larger (and hopefully more consistent) dataset may not have this restriction. I used a 75-25 train-test split and tuned hyper-parameters via grid search using 3 folds on the training dataset.

I want to prioritize maintaining a baseline specificity to avoid alarm fatigue, which could otherwise render a system unusable regardless of its sensitivity. In the case of predicting whether a patient's health warrants an instant alert, I based the target false positive rate off of the past and current instant alert criteria (as described in subsection 4.3.1) to be ideally no more than 2.5% and certainly less than 5%. For detecting whether a patient has a medical complication of any level of severity, I compare the model's performance against the performance of automated red flags for the same classification task.

The resulting XGBoost model for predicting severe medical complications performs comparably to instant alerts when maintaining specificity. I would not recommend using this particular model yet, as it current offers no added benefits and the current instant alerts are more interpret-able. Combining the positive predictions of both the model and current instant alerts would decrease specificity too much.

The XGBoost model for predicting the presence of CHAMP-related medical complications of any severity outperforms CHAMP's automated red flags, as shown in the receiver operating characteristic curve (ROC curve) in Figure 4.9a and the precision-recall curve (PR curve) in Figure 4.9b. The histogram in Figure 4.10 displays how the classification's probability scores for each example in the test set aligns with coarse groupings of medical severity.



(A) ROC Curve (B) PR Curve

FIGURE 4.9: Plots comparing the performance of the model for predicting the presence of any medical complication against the baseline of CHAMP's automated red flags and a "dummy classifier" that always guesses "Healthy".

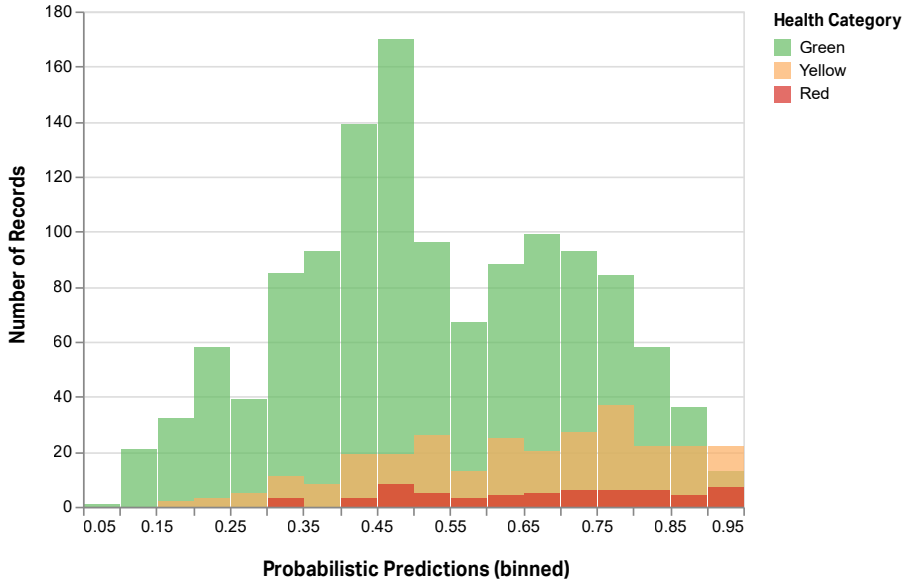


FIGURE 4.10: A histogram of the different probabilities associated with each classification on the test set, color-coded by the level of medical severity. Green represents having no CHAMP-related complications, Red represents warranting an instant alert, and Yellow is everything in between.

4.4.4 Feature Analysis

Both in preparation for developing a prediction model and to better unpack this domain, I conducted feature analysis using several methods, some of which I list here. I charted and examined bivariate plots for all pairs of features to visually look for patterns that link features to severities of medical outcomes, e.g., clustering of different medical outcomes. To confirm that there were little to no obvious clusters, a Kruskal-Wallis H-Test [64] compared samples with different medical severities for differences in feature distributions. Feature ranking from models, such as the XGBoost model, also offered insights on which features offer more predictive power.

Based on this analysis, the most salient feature (i.e., the feature with the highest predictive power when used in isolation) is a patient's number of days since discharge. The longer it has been since a patient was discharged form their initial hospital stay, the less likely they were to have a medical complication. This finding makes sense; more time spent out of the hospital can correlate with a more stable body. This insight, when shared with the CMH CHAMP care team, changed their medical practice. They decided to incorporate this insight when choosing what order to review daily patient data. As described in subsection 4.3.2, prioritizing patients with higher risks for medical complications is one goal for improving CHAMP.

Other findings support the medical intuition behind CHAMP's red flag and instant alert criteria. As discussed in subsection 4.1.4, there has been little data to base the red flag and instant alert criteria on. Instead, they were chosen by medical experts based on their highly educated intuition. For the first time, this CHAMP dataset and feature analysis offered empirical evidence that, while these criteria are not sufficient to capture all medical complications, they do capture patterns that track with medical complications. One example is a patient's oxygen saturation. The red flag and instant alert criteria apply thresholds to identify patients with particularly low or high oxygen saturation. As illustrated in Figure 4.11, while the distributions of oxygen saturation for patients with and without medical complications heavily overlap, there is indeed a higher proportion of patients with medical complications at the extrema of oxygen saturation levels. In other words, past CHAMP patients with particularly low or particularly high oxygen saturations do have a higher likelihood of medical complications.

4.5 Discussion

The project began with speculation that machine learning could help improve the CHAMP monitoring system, and I set out to investigate whether and how that may be the case. To do so, I explored what the opportunities were and

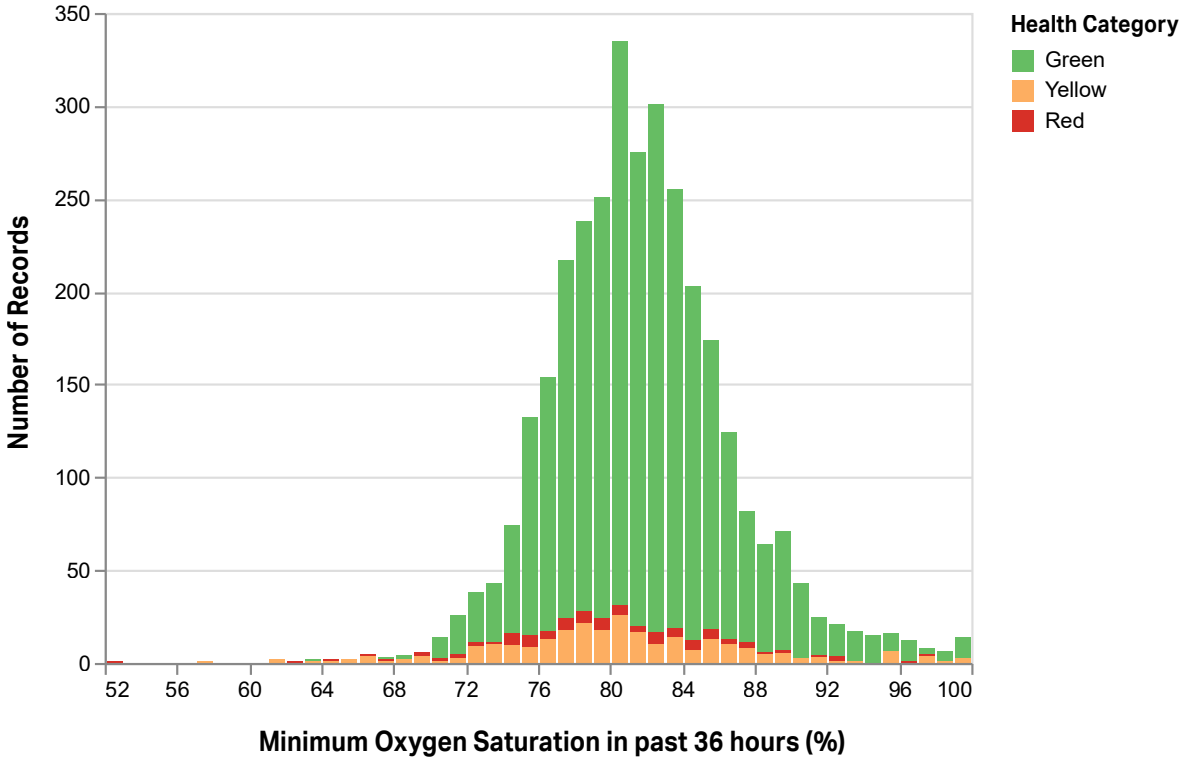


FIGURE 4.11: A histogram of the CHAMP records' minimum oxygen sautrations from 36-hour windows, color-coded by the level of medical severity. Green represents having no CHAMP-related complications, Red represents warranting an instant alert, and Yellow is everything in between.

charted a roadmap for developing such methods.

By unpacking the domain and integrating the expertise of the CHAMP medical teams and computer science researchers, I defined specific applications for machine learning to the degree that they can be treated as problem statements that have also been vetted for real-world relevance. By defining these applications, I uncovered that there were indeed opportunities for machine learning. Developing the infrastructure and approach for an example machine learning model made the definition of two of these opportunities even more explicit and demonstrated that such an application of machine learning was possible.

Defining the framework and building the infrastructure for such a machine learning model also enabled a first-ever empirical evaluation of current automated heuristics for instant alerts and red flags. It enabled hospital care teams to confirm their impressions on the heuristics' effectiveness, quantify the room for improvement, and serve as a baseline to compare any new algorithms against.

4.5.1 Example Use Case for Example Method

The example machine learning model's improvement over CHAMP's current automated red flags for predicting medical complications is a demonstration of how CHAMP could benefit from machine learning.

For proper closure on the example method, here is an example of a concrete and specific method for using the model in the context of helping triage patients (an application described in subsection 4.3.2).

Although a goal can steer what model to develop in the first place, the performance of the model influences its use cases and how I would apply it. The chances that a computer will always make perfect predictions is practically impossible, especially given how diverse, complex, and inconsistent the available data is for predicting medical complications (see subsection 4.1.4 for details) and the current dataset limitations (see subsection 4.2.3 for details). Instead, there are important trade-offs between sensitivity, specificity, and precision that I must address to develop an effective method of using a model.

Much like the alarm fatigue discussed in subsection 4.3.1, the system needs to be trustworthy enough for people to continue using it. A higher average performance than previous methods is only part of the picture for making it useful; if people begin to dismiss its predictions regardless of its overall utility, this lack of trust would render the system useless. Hence, the system design needs to account for precision, specificity, and the way it presents its predictions while maintaining a high enough sensitivity for it to help care teams catch medical complications sooner.

It is important to balance these trade-offs. Concrete descriptions of these trade-offs can help contextualize them. Here is an example use case or method to illustrate this point:

> Every one to two days, suggest looking at one or two patients before looking at the others. Roughly half of the time, these patients have a medical complication that at minimum warrants outpatient treatment. At least three quarters of the medical complications would not be brought up this way. An expected quarter of all medical complications would.
>
> I can change these numbers, e.g., allow the system to suggest more than two patients if they reach a certain predicted likelihood threshold, but they may have medical complications only a third of the time. A trade off for these false-positives is that it would likely catch more than a quarter of the medical complications, but still less than half.

I suspect that intuitive explanations of uncertainty is important not just in joint discussions with the medical team, but also in the system's presentation of its predictions.

4.5.2 Limitations

The generalizability of this work on exploring machine learning opportunities for CHAMP has several important limitations: the dataset size, the representation in the dataset, the representation of expertise, and challenges in the space of single ventricle heart disease as a whole.

For a disease as complex as single ventricle heart disease, the dataset available for this work was too small to draw conclusive evidence for generalization. The dataset comes from 84 patients over the course of two years and has many missing values, as described in subsection 4.2.3. Because several patients have little to no recorded tablet data (as illustrated in Figure 4.5), a more realistic count is approximately 80 patients. With explicit definitions for health categories (see subsection 4.4.1) of 'red' (urgent medical complications that warrant instant alerts), 'yellow' (medical complications that do not need instant alerts but should prioritize a patient), and 'green' (otherwise healthy), I can calculate a bound on how many data points I have when extracting all potential red, yellow, and green instances from all patients.

Table 4.3 displays the number of samples calculated as such. Evidently, requiring all samples to have complete features (i.e., the requisite data entries to calculate the features listed in subsection 4.4.2) results in too small of a dataset to train a model for such a complex disease. Applying an inclusion criteria of requiring at least a skeletal representation of the two most medically relevant pieces of tablet data (weight and oxygen saturation) offers an upper

Inclusion Criteria	Green	Yellow	Red
Any data (null okay)	4712	365	163
No null values	67	27	7
Scarce Requirements	1780	216	62

TABLE 4.3: A table of the CHAMP dataset size, broken down by health category, depending on different exclusion criteria for null values. The listed sample size with scarce requirements only require data points to have a minimum of two weight measurements over the past 4 days and four oxygen measurements over the past 7 days – a deficient feature set for an effective algorithm.

bound on the sample size from the current dataset (the last row of Table 4.3). However, this limited feature set is likely insufficient to train an effective algorithm (if it were, the CHAMP care teams would have figured it out without machine learning).

As discussed in subsection 4.1.4, the population of single ventricle heart patients is heterogeneous, exhibiting many different heart anatomies and thus many different types of medical complications that manifest with different types of symptoms (and hence, express different features). Furthermore, the variety of home environments and types of care from the patients' caregivers introduce more complexity. For a patient population and condition this complex, it is could very well be the case that the current CHAMP dataset only represents a narrow variety of interstage patients.

This dataset size also limits the ability to generate statistically significant results for many types of models. For example, if a model requires all features to predict whether a patient's health falls into the 'red' category, predicting twice as many true positives than the current instant alert system does not mean much with this data set. As the instant alerts barely have over a 10% true positive rate, it would correctly predict approximately one out of the seven 'red' cases with complete data. Correctly predicting twice as many only means predicting one additional 'red' case correctly, which could very well be a fluke.

The reliability of this data is further impacted by noise. Some of this noise comes from the process of data collection. Although I took care to recognize and correct as many errors in the data as possible, there could be more. It contains many gaps from when caretakers did not record data. Similarly, patterns in the data suggest that a few caretakers might have estimated, rounded, or guessed some values for their child's measurements (e.g., when they record the exact same value multiple days in a row).

The dataset also contains another type of noise: changes in methods of care or medical practice over the course of these first two years of CHAMP. As CHAMP was a nascent system during this period of data collection, care teams experienced a learning curve, continually maturing and improving their ability to work from CHAMP data. I anticipate that these changes are most dramatic during the early years of a system, impacting the consistency of this dataset.

The generalizability of the findings are also limited by the data sources, both for machine learning development as well as for the foundational HCI work. As all of the patient data came from CMH, it may not represent different patient outcomes that may arise out of different methods of care or practices from different hospitals using CHAMP. Differences in resources or resource management also impact what pain points different hospital teams experience with CHAMP, as demonstrated by the different needs for patient prioritization at CMH and SCH described in subsection 4.3.2. While real-world experience from two different hospital teams inform the findings, there may be even more differences from other hospitals.

These results were also limited by the duration of the project – all of the work needed to fit within an internship's twelve week period to adhere to the IRB protocol.

4.5.3 Future Work

This exploration of AI opportunities for CHAMP is only the beginning of working in this space; there are many directions for continued work on improving CHAMP with machine learning.

One direction is to build on or improve upon the example machine learning algorithm, be it for predicting hospital readmissions or otherwise triaging patients. More CHAMP data would enable a better investigation, for the reasons listed in subsection 4.5.2. Increasing this dataset could come with time, as CMH would be able to collect more data, or figuring out ways to combine datasets with those of other hospitals'. With a sufficiently large dataset, other modeling methods could also be investigated.

One idea for improving the example model is to further break down the 'healthy' (or 'green') classification into two groups. The histogram of model prediction probabilities shown in Figure 4.10 showcases a bimodal distribution of probabilities for 'healthy' labels. Investigating what may represent this distribution could help improve the algorithm. One starting point could be to compare what heuristics could divide the 'healthy' group to where they fit in this distribution. For instance, the 'healthy' could be broken down into one group for patients who may have recently been 'unhealthy' (i.e., were recently discharged from a hospital admission or recently had an urgent medical complication) and one group for patients who have not.

There are other ML-related directions to explore using the infrastructure I built. One idea is to focus specifically on what labels the current automated criteria (i.e., instant alerts and red flags) misclassify. An analysis on positive results could help bring insights on where the instant alerts and red flags are deficient. Another idea is to train a regression

based on the different rankings of health complication severity (as listed in subsection 4.4.1) to develop a "risk score" for patients – a concept that hospital staff already use for other medical conditions (see subsection 2.1.2 for more details).

One open question lies in understanding the relative importance of different tablet data. As described in subsection 4.2.3, the demands of caring for both a newborn child and a child with single ventricle heart disease are significant enough that parents often do not have the bandwidth to record every measurement for CHAMP. A question then lies in what empirical evidence is there for some data being relatively okay for them to skip (currently, the medical teams expect input and output data to be skip-able) as well as what data deserves extra emphasis (e.g., whether taking one versus two oxygen saturation measurements per day makes an important difference).

Another direction for future work is to explore algorithms for other applications of AI for CHAMP. CHAMP's video data is a trove of unexplored opportunities for automation, predictions, or estimations. subsection 4.3.3 describes what specific tasks or goals are most relevant for information-extraction. To reiterate, the three most valuable characteristics to extract videos are respiratory rate, chest retractions, and cyanosis around the nose and mouth. Note that exploring video data would be a longer-term project that would require many resources (i.e., a substantial amount of care team availability to label data, significant funding and time for developing novel algorithms, and the capacity to have a lot of patience as the project would likely not present benefits in the near-term).

For any CHAMP machine learning model to be fully vetted, it not only needs to fulfill metrics in evaluations of accuracy, but also needs to demonstrate its potential for or level of impact in clinical settings. Given a sufficiently promising model and specific use-case for the model, a next step would be to test the utility of the model when used in clinical practice. One approach could be to conduct AB testing, alternating periods of time (e.g., randomizing stretches of days) in which care teams do or do not incorporate results from the model.

With the aim of improving CHAMP and its effectiveness as a while, it would also be worthwhile to explore directions that do not involve machine learning. Applying HCI techniques to improve CHAMP's user interface could improve the workflow and a care team's ability to analyze data. With the growing maturity of CHAMP, insights could be compiled into a more mature training program for new care teams and care team members. Becasuse care teams in some hospitals struggle to find sufficient time to review patients, we could consider increasing the personnel such that the hospital can afford to set aside dedicated time for CHAMP data review.

Chapter 5

Insights for Working in this Space

Through the exploration of applications for consumer devices to improve monitoring methods in pediatrics, I developed insights that could help inform future work in this direction.

This chapter discusses insights I gained specifically while working on BiliCam or AI opportunities for CHAMP: recommendations for how to decide whether to pursue a particular direction or apply machine learning, and general advice for working in this space.

5.1 When to use Machine Learning (or Technology in General)

There are two important criteria to address for deciding to work on any machine learning, data science, or mHealth approach: (1) its real-world relevance and (2) its technical feasibility.

It's very common for technologists to be excited about building something that uses a particular technology, such as the latest advances in machine learning. However, whatever we build would only have positive impact if it addresses a real-world problem in a meaningful way. Otherwise, we could end up building a powerful system that solves an insignificant (or even non-existent) problem, or worse yet, exacerbate a different problem. For real-world impact, our approach also needs to be technically feasible or it may not materialize. This combination of criteria requires integrating a human-centered approach to ensure relevance with a technical basis that justifies feasibility for an approach to succeed.

5.1.1 Medical Relevance

Relevance is complex and multifaceted. It entails questions about whether an approach address a real problem, its effectiveness, how actionable it is, whether not just the need but the intention and infrastructure exist, and its potential impact on indirect stakeholders. In short, to make positive impact when tackling a real-world problem with machine learning or any other technology, it is critical to approach the problem holistically. The technical skills for developing the technology are not enough.

For these reasons, the methods from human-centered design and HCI are powerful tools for machine learning. As eloquently written by Jeff Bigham [16]:

> Discovery of important problems, mapping them onto computationally tractable solutions, collecting meaningful datasets, and designing interactions that make sense to people is where HCI and its inherent methodologies shine.

Both when deciding whether to apply machine learning as well as throughout its entire development process, be sure to apply the tools of HCI throughout the process for guidance on taking steps in effective directions.

5.1.1.1 Verify Impact

One of the first questions to ask is whether the strategy driving the development of the machine learning algorithm or technology actually addresses the overarching real-world problem.

Let's consider developing some screening tool for some medical condition as an example. Does this screening tool actually help address the problems presented by this medical condition? Or do the pain points of this condition lie elsewhere, such as problems in prevention, treatment, education, adherence, or compliance? Is the condition so common in relevant or target regions of the world that screening would make a negligible difference compared to addressing other pain points? Does the current methods of screening actually present problems? Would the new approach address the problems from these current methods, or would it only offer an alternative without improving upon current methods?

An approach to uncovering whether a strategy would have impact is to flip the question and ask what the problems or pain points are within the problem space. By identifying the pain points, we can consider how well a strategy addresses these problems and whether there is a more impactful problem to work on instead.

This approach of first identifying pain points was how I identified the applications of machine learning for CHAMP described in section 4.3.

5.1.1.2 Consider Alternate Solutions

In addition to evaluating the impact of a particular strategy, it's also important to evaluate the tactics: whether a particular technical approach would be effective at implementing this strategy.

The relevant questions here all pertain to justifying the *why* behind using a proposed technical approach. For example, you could ask: Why use machine learning? Why use mHealth – what does the mobile platform offer that nothing else can? What's the difference between taking an mHealth approach versus developing a cheaper version of an existing device? *Why even use technology in the first place?*

There are many different ways to address a problem. A problem in question can often benefit more from entirely different tactics: user interface improvements, better visualizations, better training, better framing, more human resources, policy change, etc. Think broadly and consider alternative approaches before deciding on one path.

To help uncover these alternative solutions, figure out what the precise challenges are in the problem space. Really drill down, be specific, and get to the heart of it. Sometimes that means conducting contextual inquiry or shadowing someone to observe where their struggles are. Remember that people from different disciplines have different perspectives and framing to recognize challenges, so an interviewee may not have the same awareness to list all the challenges you would be able to notice when observing them at work. When someone brings up one of their challenges, repeatedly ask them or yourself "why." Pinpoint and define these challenges as specifically as possible.

With these resulting well-defined problems, creatively come up with different solutions – ideally from different practices. Get other people's input. Develop and bounce ideas with people who are experts in the space. This process sets us up to be able to consider whether AI is really the better approach.

I used this approach when considered projects in other medical domains, each with interested potential medical collaborators. By asking these questions, I was able to quickly recognize that mHealth not be an effective approach for a number of these projects. Finding out early saved more time, money, and energy than finding out after developing a technology.

5.1.1.3 Ensure that Outcomes are Actionable

While approaches like machine learning can be powerful for making medical predictions, it's also important to determine how you would use these predictions. Are they actionable? In what concrete way? Does the prediction enable a specific intervention, and if so, what? Note that there are different degrees to which we can be specific about how actionable our model outcomes are.

5.1.1.4 Investigate Existing Institutional Forces

It's common to think of technology as a solution, but there is so much more to a solution. Technology ultimately is only a tool that amplifies existing institutional forces, as described in the theory of technology-as-amplifier [135].

For an approach with technology to succeed, it's important to consider whether there are already forces in place that would use the technology as intended. Are there specific people actively trying to address the same problems who would use the technology? Do they have the motivation, intention, and bandwidth to use it? Is there sufficient infrastructure for continued use, such as the ability to maintain the technology, the electrical power to drive it, and business model to sustainably incorporate it into a workflow? Are there political, social, or historical causes behind this problem that could inhibit using this technology?

There a many examples of failure on this front. Many of the documented cases are especially apparent in a lot of work on the front of information and communication technologies for development (ICTD). You can find many illustrative examples, with their context in the theory of technology-as-amplifier, in the book Geek Hersey [134].

5.1.2 Project Feasibility

Ensuring medical relevance is only one part of the equation for deciding to work on a project; it's also important to investigate technical feasibility before investing a lot of time, money, and energy. Feasibility here both means whether there's reason to believe the technical ideas could work and whether you have the resources to carry the project out.

5.1.2.1 Technical Tractability

Like the skills for recognizing real-world relevance, predicting how technically tractable a project is multifaceted: it can come from an understanding of the underlying technology (e.g., a background in machine learning for any ML-based approaches), the underlying principles (e.g., the physics), and related medical approaches to the same problem.

A helpful way to check initial feasibility for addressing a challenge with technology is to consider whether there is a similar medical precedent for the diagnosis or intervention.

For instance, measuring iron levels in the blood non-invasively has the medical precedent of using a pulse-oximeter which uses light waves instead of blood samples. This fact can serve as evidence that it may be feasible to do similar kinds of measurements for iron by measuring light waves through smartphone cameras, as was demonstrated by HemaApp [140]. On the other hand, measuring other vitamins such as zinc of vitamin A have only been possible

through blood tests so far and have no non-invasive precedent, making it significantly more challenging to envision a method to do so with a cheap technology.

Another example of precedent is whether people are already able to recognize and observe a signal with the right tools (e.g., a doctor's ability to recognize jaundice or a nurse's ability to recognize patterns in CHAMP data that suggest a medical complication). Mining data to discover a "hidden signal" that's invisible and unbeknownst to humans certainly has its allure, but counting on such an approach is risky. If people cannot already interpret the data in meaningful ways, there's no guarantee that meaningful signals are even there.

The tractability of a project depends not only on whether the technology seems possible, but also in the ability to validate its effectiveness. What are the requirements and methods for evaluation? What is the ground truth data? What is the baseline to compare against? Is the necessary data for evaluation available or obtainable? Does evaluation require a randomized controlled trial or other such study, and if so, is one possible and ethical? A technology needs rigorous validation before people can justifying incorporating it into medical practice. Asking these questions is helpful for deciding whether to pursue a project and formulating how to develop such a technology.

5.1.2.2 Resources for Development

The success of a project also depends on having sufficient resources. While funding may be one such resource, there are also many other important ones.

One crucial ingredient is a fertile collaboration that includes people with the relevant medical or domain expertise and people with the necessary technical skills. In a successful collaboration, both parties would regularly communicate with each other; frequent check-ins, updates, and clarifications can help address the unconscious assumptions, misinterpretations, or misunderstandings are almost guaranteed to happen before they cause costly problems (see subsection 5.2.2 for examples). It can help to establish how available people are early in the collaboration; doing so enables the team to find work-arounds or decide against a project direction if one party will frequently be too busy to commit much time or energy (e.g., they have a demanding schedule for working with patients).

Another resource is the availability of data or the ability to collect it. If data already exists, can the team have access to it? Does it capture the right information for ground truth and features? Are there enough samples? Are they of sufficient quality? If the project requires data collection, can the team acquire access to the right study participants (in an ethical manner), the relevant study environments (e.g., connections and willingness from the right parts of a hospital

or clinic), and necessary personnel (e.g., data collectors with clearance to work and interact in this environment)? Does the team have any time constraints for the project, and if so, how does that effect the ability to collect data?

These questions guided me on where to focus my efforts for CHAMP. For example, they redirected me away from investigating how to extract data from videos enough though it is a ripe area for research. Although the videos offered rich amounts of data, they did not have the requisite ground truth labels (as mentioned in section 4.3). I considered collecting ground truth labels, but recognized several problems. Only the doctors and nurses on the team had the unique combination the expertise to accurately recognize ground truth and the authorization to access these videos. I could not hire additional help and labeling the data ourselves would take more time than any of the team could afford for the early machine learning work in CHAMP.

5.2 Advice for Developing Medical Technology

After several years of work in the space, I developed insights from my experiences that I would like to share as it may help with other forrays in is area. I summarize them in this section and include ways it has impacted my work on BiliCam and CHAMP.

5.2.1 Ensure *Every* Part of Pipeline Aligns with Specific Human Need

To make a machine learning model or other technology effective, developing a high accuracy is not enough — we need to understand and incorporate the context of the problem it's trying to address. For instance, how do we make sure people apply the sensors correctly to get the right inputs? Or well before people use the model, how do we make sure we collect, train, and test on the correct data? That the features and labels are medically relevant? That when both developing and when applying the model, we have quality data? That we evaluate our models in ways that are meaningful for its real-world application? That people interpret and act on results correctly? But before all else, how do we make sure we're even addressing the correct problem — that the modeling problem maps to a meaningful real-world problem in a very concrete and actionable way?

Every part of the model must align to a specific human need. Otherwise, we would likely end up building a powerful system to solve an insignificant — or nonexistent — problem. Worse yet, it could even exacerbate a different problem.

These questions guided many of my decisions in both BiliCam and CHAMP. Here are examples for each one.

Correctly applying sensors in BiliCam was one of the driving factors behind why and how I developed the data collection app, as discussed in subsection 3.3.1 and subsection 3.3.2. My skills in HCI and user interface design played a major role in addressing this challenge and had important impacts on data quality.

Many facets of taking a human-centered approach impacted the ability to collect, train, and test on the correct data for both BiliCam and CHAMP. I describe several of them in subsection 5.2.2.

Developing methods to collect high quality data in BiliCam ties with ensuring that people correctly apply both sensors and the overall data collection workflow. It drove the development of image quality feedback (see subsection 3.3.2) and ultimately the square color card and automatic photo capture (see subsection 3.6.1). These iterations all depended on applying HCI skills to uncover and address underlying human needs to improve data capture.

To ensure that they were medically relevant, the features and labels in the CHAMP example model were grounded in current medical practice (see subsection 4.4.1). Uncovering what information had medical relevance involved many semi-structured interviews with multiple domain experts; guessing what data to use as ground truth labels would have lead to medically irrelevant models and blindly trying all different combinations of features was neither practical nor computationally tractable.

In making sure the model evaluations are meaningful for real-world applications, I uncovered that typical model analysis with ROC-curves for predicting CHAMP medical emergencies were irrelevant (see subsection 4.3.1). Without understanding the importance and low tolerance of false positives for this application of machine learning, I could make critical mistakes in evaluating which models are better or even whether a model outperforms current practice.

Making sure people use models correctly includes framing results for the right interpretation, balancing risks of uncertainty, and making the appropriate suggested use cases. These concepts guided the development of the example CHAMP machine learning model's use case, as described in subsection 4.5.1. This same concern would also guide any work on developing the user experience of a commercialized version of BiliCam or CHAMP; I would want to make sure it frames and conveys information from model predictions in a way that would best guide people to the right use case.

Properly understanding the real-world problem behind neonatal jaundice and the specific ways it maps to human needs was critical in developing BiliCam. A naïve understanding of neonatal jaundice could have lead my team and I to invest years of research to develop a jaundice-detection system. However, most healthy newborns have visible jaundice and do not need treatment; a mobile app that detects newborn jaundice would be *useless*. Furthermore, it was critical that I understood how age and gestational age factor into the relationship between jaundice and whether a child needs medical treatment. Because of these factors, I understood that developing a classification model for detecting

dangerous levels of jaundice based on images alone would not work. Instead, I needed to develop a regression model to predict the blood concentration of bilirubin, which I could map to age in order to do medical assessments.

Similarly for CHAMP, rather than diving into the project by immediately building models, I invested a significant portion of my time and energy into investigating what needs I could address. This investigation lead to understanding what applications of machine learning can help CHAMP care teams in meaningful ways (see section 4.3), which gave me meaningful direction for my work. Before conducting this investigation, each member of the collaboration had ideas for machine learning applications that turned out to be medically irrelevant or computationally intractable with the time and technology available at that time.

5.2.2 Fully Understand the Problem

When developing technology, it is important to understand the problem it's trying to address. There are many parts to understanding a problem (e.g., from studying its context as I did in BiliCam's formative work described in subsection 3.4.3, or examining a problem from multiple angles as discussed in parts of subsection 5.1.1). Here are a few suggestions to help with a few other aspects to understanding a problem.

Fully understanding a problem requires effort. Read up on medical literature and related works. Understand the underlying physiology behind a medical problem as well as current practices documented in the literature. The more computer scientists self-educate before talking to medical collaborators, the more efficient they can be with their time together, as they could focus the conversation on undocumented or new ideas. Having basic medical knowledge from reading can also help develop a common language between collaborators. For instance, machine learning researchers may frequently refer to 'recall', whereas doctors and nurses referred to the same concept with the word 'sensitivity'. Furthermore, the same words can have different meanings in different domains.

Unpacking knowledge from medical experts on the team is even more important than reading literature. Medical experts have intuition, medical common knowledge, a level of understanding that comes from years of training and experience, and knowledge of practices that may not be published or documented. The understanding and skills to effectively use CHAMP, for example, exceeds what's published in medical journals or guidelines. Medical knowledge for most infants do not apply to CHAMP patients (as discussed in subsection 4.1.4), and it is not possible to distill all the experience and intuition of CHAMP care teams in only a couple publications.

However, tapping into their medical and domain knowledge is challenging. It's unrealistic to ask a doctor or nurse to explain all of their relevant knowledge and then expect to receive all the information you need. Doing so is akin to

asking someone to teach you German in a couple of sittings (particularly when they're a native speaker who studied German literature instead of German education). Imagine you're one of the medical experts — as someone who's been steeped in this field for your entire career, it's really hard to identify what's medical common knowledge to you and unknown to others. Without a machine learning background, it's even harder to figure out what information you need to share to help develop the machine learning model you want — especially if the person you're talking to doesn't have a medical background.

This challenge is one area where HCI skills in conducting semi-structured interviews, focus groups, and contextual inquiry can help. These skills help unpack the domain to both drive to the heart of the problem and uncover its relevant context. It's important to think about and prepare specific questions for informational conversations or interviews with medical experts, as the right questions can prompt them to discuss other valuable components of the problem space or fill in blind spots (be vigilant about identifying those blind spots too!). Visiting and shadowing the context they work in can further provide a richer understanding to the problem space, and applying skills for contextual inquiry can help make the most of such visits.

5.2.3 Anticipate Misinterpretations

Misinterpretations between medical experts and technology experts on the team are almost guaranteed to happen. After all, we all come from different disciplines, use different terminology, have different understandings of what is common knowledge, and often bring in misconceptions about each other's disciplines. While one capability in another domain may surprise is, we also often have unrealistic expectations of what is possible (e.g., from misconceptions popularized or perpetuated on television such as human-like, all knowing artificial intelligence or elite hackers who can build anything in hours).

I found that regular meetings and discussions with the entire team has repeatedly helped my team and I recognize and address misinterpretations before they cost my team and I too much. We would meet nearly every week to check in, give each other updates, and discuss any ideas so that we would be on the same page and check in with each other to make sure we're headed in the right direction.

These kinds of check-ins and updates helped shape BiliCam. While brainstorming approaches for BiliCam, the technical side of the team thought about how jaundice would change the skin color over time and started to conclude that the approach would involve taking pictures of the same baby over multiple days to observe changes in visible jaundice and whether it was trending upwards or downwards. The idea was that app wouldn't need to estimate

bilirubin levels – it would only estimate whether the jaundice was changing (a potentially easier or simpler task). However, when we discuss these ideas with the pediatricians on the team, we quickly realized that we nearly steered the project in the wrong direction. When it comes to assessing risk from newborn jaundice, it's not about which direction jaundice is trending so much as what the absolute level of jaundice at any given moment. Even if future research investigates whether trends are helpful for predicting upcoming jaundice levels, these trends would still need to be based on bilirubin concentrations; jaundice levels can trend upwards without reaching dangerous levels. By having this scheduled check-in, we were able to correct the course of this project within only a few days of coming up with the wrong idea – well before we started to develop anything.

Discussions between the medical and technical experts on the team for CHAMP also directed what projects to focus on. Between these discussions, the technical side of the team became excited about multiple ideas they came up with that were technically tractable and interesting to build. Through discussions with the medical side of the team, however, they uncovered misconceptions about how medically useful these ideas were. For example, one idea involved applying computer vision to calculate how much a baby moved around in CHAMP videos. The intuition was that healthy babies were more likely to wiggle their hands and feet, whereas babies who were sick or struggling to breathe would remain relatively still. While that may be the case, many healthy babies are also resting in these videos and hence, do not move much, and unhealthy babies could also be crying and screaming, in which they may be moving. Although I haven't yet evaluated the data for certainty, I realized there was a low chance for whether the amount of movement in videos correlates with a CHAMP patient's health status. Even if it were to correlate, there are many more factors that have more predictive power based on the nurses' experience and intuition. By iteratively checking in and trying to all be on the same page, I prevented my team and I from investing days or weeks of work in what turned out to be unhelpful or high-risk directions.

Another benefit to regular discussions within the team can include simplifying or streamlining an approach. Again, because we bring different backgrounds and training to the team, it's possible for one part of the team to come up with an approach that is needlessly complex from the point of view of the other. By repeatedly synchronizing and taking efforts to understand each other's ideas, we can recognize unnecessary complexities and simplify our approach.

A supplemental approach to these frequent check-ins is to write up the project's background and proposal and share it within the team. The process of writing such a document forces people to articulate their thoughts in more a more precise and complete way than they may in conversations. In turn, it can be easier to recognize errors and misconceptions when reviewing a document that spells them out explicitly. BiliCam, for instance, benefited from this

process of writing such documentation and involving everyone on the team for reviews, edits, and iterations. Different members of the team caught different conceptual errors that did not come up in our discussions.

Committing regular time to meet regularly and write documents about the project may be time-consuming, but basing work on misconceptions can cost significantly more time. I believe these measures are worthwhile. After all, it's easier to change blueprints than houses.

5.2.4 Formulate Studies Intentionally

Running studies and collecting data is difficult. BiliCam's formal local study, for example, lasted for 9 months and cost upwards of $1,000 per sample if you divide the grant by the number of participants, not to mention the team's coordinated efforts and many months of preparation. Re-running a study because I failed to collect some piece of data would be even more difficult. As a result, I believe in designing such studies carefully and intentionally. That process includes identifying what external variables to control, what data to capture, and what sources to capture from.

As with any scientific experiment, it's important to identify and control or record external variables. When developing a study for a proof of concept, I believe in trying to control for as many external variables as possible even if the concept should ultimately work within a wider set of conditions. BiliCam, for example, should ideally work on multiple types of phones. However, should the results of the proof of concept study fail, it may be impossible to tease apart whether the failure was due to problems in the concept itself or from not controlling the type of phone in that limited dataset. For this reason, I constrained as many external variables as possible in BiliCam (as described in subsubsection 3.3.1.2 and subsection 3.3.2). It's easier to set constraints in the beginning to test the concept, followed by loosening constraints in the future should the concept prove promising.

I also believe in collecting as many potentially relevant pieces data as possible during a study. It's easier to discard data that turns out to be unnecessary than it is to run another study just to collect it. Try to foresee or think laterally of any or all types of data you may need. For example, in BiliCam I decided to capture both still images and videos, with and without flash, from different distances, with multiple colors on the card, in multiple sets, and so forth as discussed in subsection 3.3.1.

In addition to controlling for external variables and collecting a breadth of data types, considering internal variables is also very important. Identifying, collecting from, and documenting the breadth of data sources impacts the study's efficacy. Think about what demographic dimensions are potentially relevant to a project. Skin color is one example in BiliCam, which was one of the driving factors for developing a nation-wide study with geographically distant study

sites and documentation and analysis of different reported racial groups. Some potential demographics to consider in study designs can include race, socio-economic status, sex, gender, age, geographic location, educational background, and many other factors. Failing to collect from a wide demographic breadth or identify gaps in breadth can lead to false claims – for example, many psychological studies that recruited all their participants from university settings have been debunked [54]. Failure to control for internal variables during analysis can also lead to false claims, much like how the famous marshmallow experiment [118] failed to control for socio-economic status and was later debunked [142].

Another dimension to data collection is not just what data to collect, but also the process for collecting it. Failures in the process lead to failures in collecting the right data in the first place (as illustrated by the anecdote in subsubsection 3.3.2.3, in which usability problems in BiliCam's data collection app lead to losing study participants). One of the most powerful tools for perfecting the data collection process is to run pilot studies (see subsection 5.2.5). HCI skills to understand the study site and context also helps shape or improve the data collection process – everywhere from user experience for data collection apps (see subsection 5.2.6) to identifying key factors to consider through interviews and contextual inquiry both before and throughout a study.

5.2.5 Pilot Studies

Preceding formal studies with a pilot study played an important role in BiliCam's success. It enabled my team and I to iron out important problems in the study (such as bugs in the data collection app or the challenges in image quality described in subsubsection 3.3.2.2 and subsection 3.4.3). The results of the pilot study also justified the expenses of expanding to a formal local study, both to ourselves as well as to grant sources.

Hence, I recommend running a pilot study on a small set of participants before any large scale studies. A pilot is important to justify scaling the project and prepare the process for any larger studies. They help discover and address unexpected kinks in the study process (which are almost guaranteed to happen!).

Furthermore, when running a pilot study, I recommend conducting it in a nearby location and limiting personnel involvement to only a few people whenever possible. Proximity and a small team size enables rapid feedback and more nimble changes during the study. As the pilot study is the best opportunity to test the study design, the ability to make changes quickly allows for testing updated versions of the study as well.

For example, different parts of the BiliCam team could quickly consult with each other because I only had a few data collectors and the data collection site was within walking distance of the computer science lab. When I discovered problems in image quality, a member of the technical team could meet with the data collectors to explain image quality

needs and iterate with them on a better quality feedback process before launching into the formal local study. Similarly, it was easier for data collectors to tell the technical team about bugs in the data collection app and give feedback on how well they were fixed when the group was small enough for everyone to be comfortable with talking to each other. After all, getting feedback from and pushing major app changes to larger groups of data collectors spread across the nation is more challenging and I would rather get it right before working at that scale.

5.2.6 Pay Attention to Usability

The usability of both any study or final product is important. It can make or break whether people would use a device or system correctly, let alone whether they would use it at all – regardless of how accurate an underlying algorithm or technology is. Usability problems in the BiliCam data collection app, for example, resulted in losing study participants and cost my team and I a nontrivial amount of money, effort, time, and stress (described in more detail in subsubsection 3.3.2.3).

A good design for usability can also improve workflow and the quality of collected data. BiliCam benefited from many iterations on that front, as discussed in subsubsection 3.3.2.3, subsubsection 3.5.1.4, and subsubsection 3.6.1.4. The improvements in data quality from BiliCam's automatic photo capture user interface likely had dramatic implications for the effectiveness of its machine learning model.

User interfaces also present an opportunity to incorporate artificial constraints that make machine learning, experimental controls, or other technical aspects of a project more straight forward or tractable. In BiliCam, I took advantage of this concept to introduce many constraints – both for controlling external variables like the positioning of the phone as well as for easier image segmentation in the algorithm. Segmenting a generic photograph of an infant to extract standardized card colors and specific skin patches is significantly more difficult. Such a method would require infant detection and recognition, identifying which parts of the infant is the sternum, and evaluating which part of the sternum offers the best data without occlusions, shadows, or other such issues, all which high accuracy. The combination of the card and user interface design shortcuts these challenges. Furthermore, these constraints enabled image segmentation to run on the phone and be fast and reliable enough for real-time automatic image capture for higher image quality and better app usability (at a time when deep neural networks could not yet run on a phone too) (more details in subsubsection 3.6.1.4).

Especially during pilot studies, establish a frequent feedback loop. Regularly ask people using the app about problems and make it easy for them to report problems. This kind of communication or infrastructure increases the

likelihood that the technical team can find out about and address problems quickly, both for logistical and emotional reasons; people may be scared to criticize the app or may misattribute problems to their own tech-savviness instead of the user interface.

5.2.7 Coordinate Both People and Agendas

One challenge that I did not anticipate during my work was how despite our best intentions, different parts of the team had invisible, unspoken differences in priorities. They lead to conflicting expectations, agendas, and timelines which at one point endangered the project; the expectations from the commercialization part of the BiliCam team began to make decisions for and rush the research and app development. We eventually unpacked what was going on and reestablished as a team that the project would fail if we don't do the research component correctly. To address this challenge, we decided to explicitly discuss our agendas and expectations during our regular meetings and elect somebody who is not invested in one specific agenda (i.e., the technical, medical, and commercialization agendas) to oversee our timeline and priorities.

5.2.8 Governance of Ethics, Privacy, and Consent

The governance of ethics, privacy, and consent is very important in technical and medical work. If we want to build technology to enable positive impact, it is paramount that we address the different dimensions to ethics, privacy, and consent in our work.

Past failures have generated or perpetuated real-world problems. For example, racial biases datasets used to train machine learning models for rating how likely someone will be a criminal re-offender has caused courts to amplify faults of our society's racial biases. An evaluation of the model's predictions against real-world outcomes of the people it evaluated revealed that he model often predicted black defendants to have a higher risk than they actually were, and predicted white defendants to have a lower risk than they actually were [6]. Unfortunately, because courts use these models in their judicial decision process, these ethical mistakes have (in my opinion, unacceptable) real-world impacts.

Working with an ethics committee, such as an institutional review board (IRB) can help navigate governance when conducting studies – a requirement from many publication venues. While not complete, additional strategies can help with thinking about some other dimensions to ethics, privacy, and consent: applying value sensitive design, incorporating diversity into the team and/or getting feedback about the project from a diverse body of people, and staying up to date on the governance or problems of similar projects or technologies.

Chapter 6

Conclusion

In this thesis, I present evidence from two projects to support my statement: that the integration of machine learning and human-centered design can forge consumer devices into tools that can help monitor infants for medical conditions in outpatient settings. These two projects span opposite ends of the spectrum in disease complexity and prevalence, which supports the generalizability of this approach.

Within the scope of this dissertation, BiliCam demonstrates a proof of concept for using smartphones to help monitor newborns for dangerous levels of jaundice in outpatient settings. It requires a fusion of machine learning and human centered design; machine learning enables it to compute bilirubin estimates, and human-centered design plays a critical role in BiliCam's foundational work as well as BiliCam's development for both usability and the ability to capture quality data. Through the process of carrying BiliCam from conceptualization to commercialization, a number of insights emerged that already influence similar research projects and informed the FDA's development of SaMD (as mentioned in subsection 3.7.3), and could assist future work in this space of applying technology for health applications (as discussed in chapter 5).

Similarly, the exploration of AI opportunities for CHAMP in this dissertation also offers a proof of concept on how machine learning, when coupled with human-centered design, can help with home-monitoring of infants with single ventricle heart disease. Contributions from this early stage work and design ethnography include concretely defined, actionable goals for machine learning; early data analysis; formulation of methods and infrastructure to implement these goals; initial algorithm investigation; and a first-ever empirical evaluation of current automated systems in CHAMP. Like BiliCam, this work on CHAMP can serve as a springboard for further research in this domain (as discussed in subsection 4.5.3) and brings insights for working on health-related technology in general (see chapter 5).

In addition to improving health outcomes and quality of life for infants monitored by BiliCam or CHAMP, reducing

the burden of these diseases also has the potential to free resources at the hospital to better care for other patients. Although single ventricle heart disease is rare, treating related emergencies is expensive and resource-intensive; the ability to free or redistribute these resources is not insignificant. Monitoring needs for newborn jaundice, on the other hand, is so pervasive that the accumulated resources necessary to examine every patient is also not insignificant.

This work focused on improving the accessibility of healthcare outside of hospitals through devices already available in these settings and integrating a human-centered approach at every step. However, many of the same overarching strategies apply to other domains. Although work in this thesis targets health care for infants, these techniques can also apply to health applications for older children or adults. While this work focuses on healthcare, its approaches generalize to other aspects to quality of life, such as education or the ability to conduct everyday activities. When integrating machine learning with human centered design into the world's increasingly ubiquitous devices, my underlying goal is to improve equity for quality of life.

In closing, my hope is that we, as technologists, will angle our work to improve equity and access to better quality of life, and do the necessary non-technical work to assess whether our approaches can be effective.

Appendix A

BiliCam Data Collection App's Instructions and Tips

BiliCam's data collection app included a set of instructions and tips. The version from the nation-wide study are included here as they can showcase some of the challenges in taking BiliCam photos.

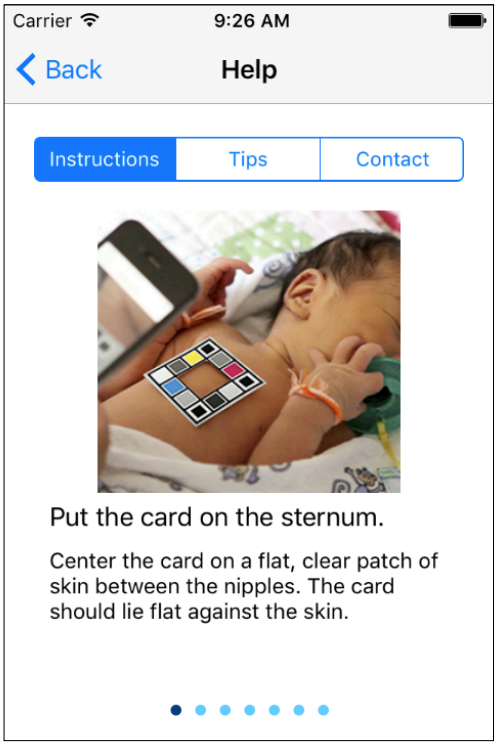


FIGURE A.1: UI for viewing BiliCam data collection app's instructions and tips.

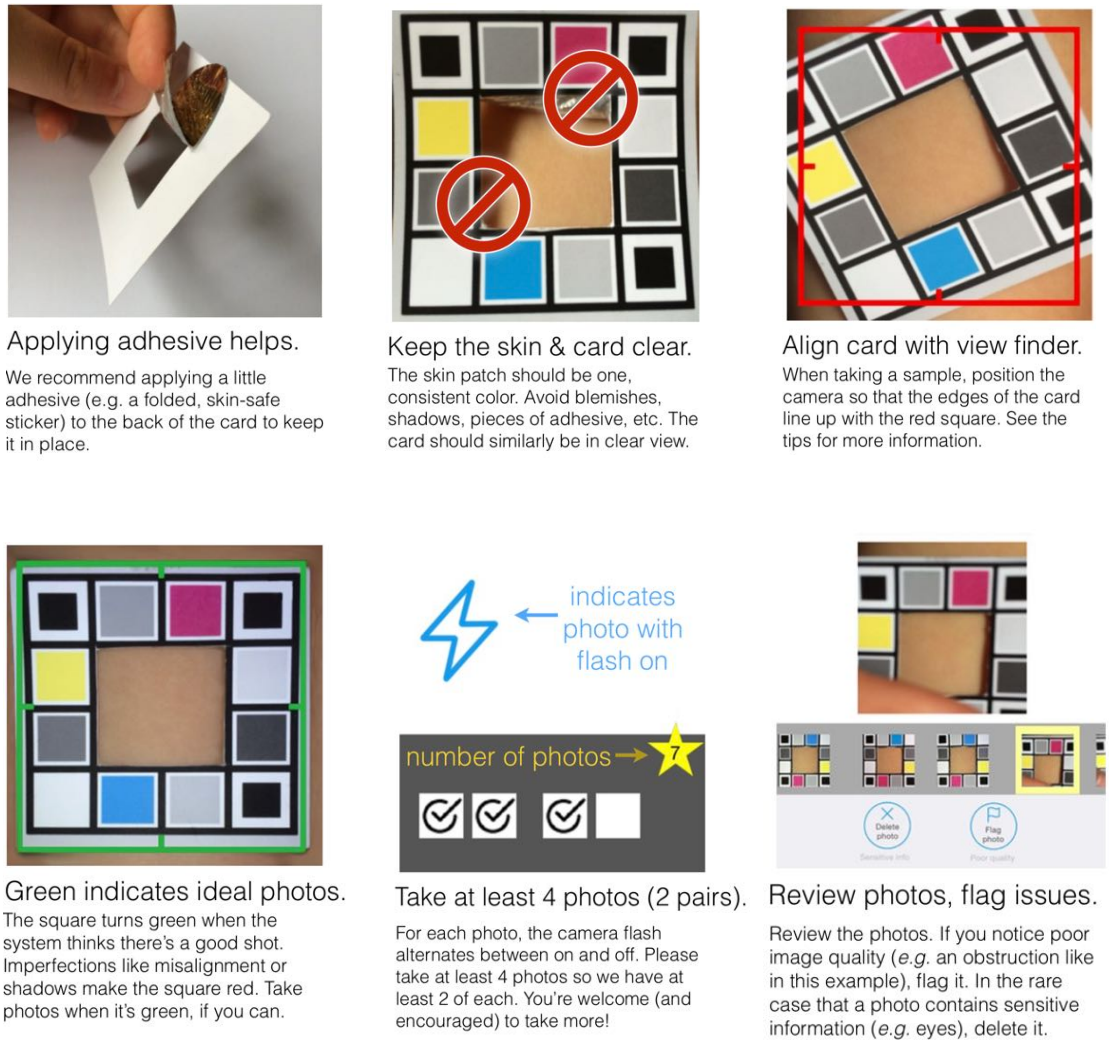


FIGURE A.2: A set of instructions visible in the BiliCam data collection app.

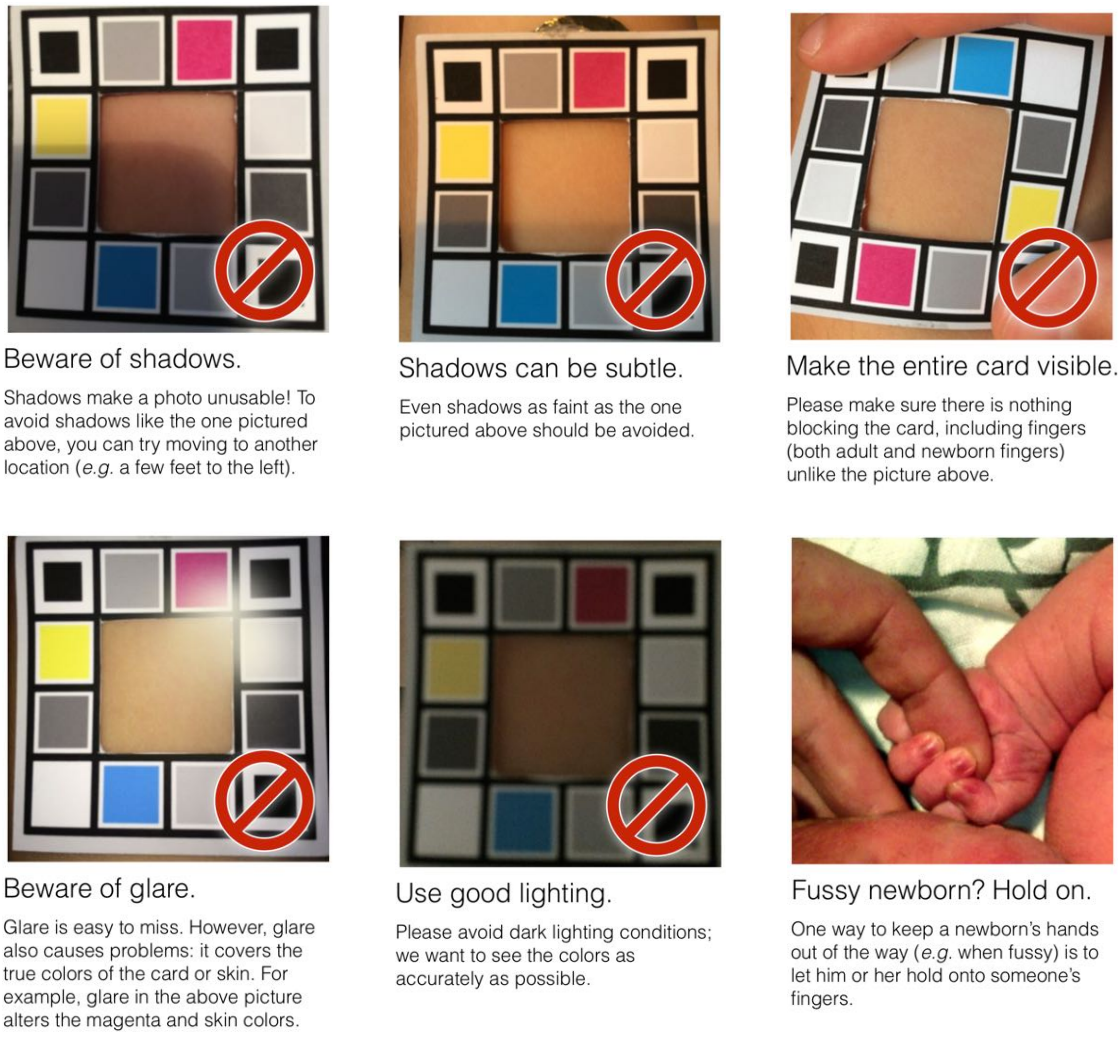


FIGURE A.3: A set of tips visible in the BiliCam data collection app.

Acknowledgements

I would like to express my gratitude to all who supported and assisted me and the work encapsulated in this thesis.

I would first like to thank my advisor, Dr. Shwetak Patel, who taught me new ways to approach, think about, and ask questions about problems. I would not be where I am today without his unbelievable skills in offering advice, as well as his instrumental role in supporting me as both a researcher and an individual.

I also want to thank Dr. Gaetano Borriello – a co-advisor and mentor who continues to inspire me to this day, in both his research vision as well as his warm and generous spirit. Gaetano, you are dearly missed.

I'm grateful for the members of my thesis committe: Dr. Richard Anderson and Dr. James Fogarty on my reading committee, and Dr. Beth Kolko, my GSR. Your input and advice both during my PhD and writing my thesis have played a valuable role in shaping this document.

I am indebted to every person who was involved with BiliCam. This project would not have been possible with the joint effort of many people from many different backgrounds. Thank you to the research team: pediatricians Dr. James A. Taylor MD and Dr. James W. Stout MD MPH, who contributed medical expertise; Dr. Mayank Goel and Dr. Eric C. Larson, former labmates who contributed additonal technical experience and are now professors; and Dr. Shwetak N. Patel who contributed technical, logistical, and entreprenurial experience. Thank you to everyone who took part in data collection: the pediatricians on the research team; Barbara, Sue, and Tatiana who not only collected data for local BiliCam studies but also responded rapidly to and gave invaluable feedback during early studies; Vickie L. Baer, RN, and Robert D. Christensen, MD, for conducting the study at McKay-Dee Hospital in Ogden, Utah. Drs Chung, Koduri, McMahon, Dickerson, and Simpson for assisting the development or analysis of the nation-wide study and supervising data collection at many of the sites; the many people who collected data at these multiple sites for our nation-wide study; all the newborns for donating their time for BiliCam, TcB, and TSB capture, and tolerated having their heels sliced open and sqeezed for blood samples (even if it wasn't their decision); all the families who consented to the study and especially families who went out of their way to return for follow-up visits; and lab workers who processed these blood samples into TSB measurements. Thank you, former MCHI+D student Irene Ma and Gwenyth Hardiman, for

your in-depth analysis and recommendaton on usability for the BiliCam data collection app in preparation for the nation-wide study. Thank you, Elliot Saba, for supporting and managing the server to host both collected BiliCam data and environments for computationally-heavy data analysis. Thank you, members of UW CoMotion for your invaluable expertise and efforts to help transition BiliCam into commercialization. Thank you, Min Joon Seo, for working with me in the early exploration for BiliCam's machine learning model. Thank you, Alex Mariakakis for your assistance with some of your code contributions to the BiliCam app or analysis process. Thank you, Mohit Jain, for working with me to explore alternative database methods for storing BiliCam data. Thank you, Tien-jui Lee and Jake Garrison, for your suggestions for developing models and your continued work on BiliCam in its commercial setting. Thank you, BSquare, for you assistance in building the data collection app for the nation-wide study. I also want to acknowledge the valuable suggestions and input from Dr. Alan Borning and students of UW's 2013 graduate-level HCI course for BiliCam's usability and value sensitive design.

I also have a lot of gratitude for everyone who was involved in the CHAMP project. I sincerely appreciate all of the medical experts from Children's Mercy Hospital and Seattle Children's Hospital who worked with me – I am in awe of what you do for single ventricle heart patients, and am grateful for all the time and energy you volunteered to help me unpack this space. A special shout out to nurse Lori Erickson, who spent many hours with me over Skype to answer questions and share her wealth of experience and knowledge for single ventricle heart disease and CHAMP. Also a shout out to Dr. Matthew Files and nurse Kendra Waldburger from Seattle Children's Hospital for inviting me to shadow their work with single ventricle heart patients and the CHAMP system in addition to my focus groups an interviews. Thank you to the Health AI team at Microsoft Research Next for inviting me to work on this project – especially my manager Jessica Lundin, from whom I received advice and support throughout the project, and my colleague Naoto Usuyama, from whom I've had a number of technical discussions and code contributions. Additional thanks to the many researchers, engineers, and staff at Microsoft Research Next who offered advice, logistical support, and themselves as sounding boards. Also thank you to all the newborns and families to participated in CHAMP, and the diligent record keeping by CHAMP's medical teams.

None of this work would have been possible without the funding from generous donors. Thank you to the Coulter Foundation, the National Insitutes of Health (NIH), National Science Foundation Graduate Research Fellowship Program (NSF), Microsoft Research PhD Fellowhip Program (MSR), the Marilyn Fries Endowed Regental Fellowship, the Microsoft Research Graduate Women's Scholarship, the University of Washington Three-Sixty Fellowship Fund, and Microsoft Research's research internship program for supporting both me and my work throughout my PhD

career.

In addition to people who contributed directly to my research projects, I am also incredibly grateful for everyone who supported me during my PhD in general. Thank you to my colleagues at the Ubiquitous Computing Lab (ubicomp lab) and the Information and Communication Technologies for Development Lab (ICTD lab), both past and present, at the University of Washington. Thank you to the many professors at the University of Washington who, in addition to those on my thesis committee, offered insightful suggestions, feedback, and alternate perspectives or lenses through which to think about my work. I also want to thank my academic network outside of UW, including Dr. Gregory Abowd and the community of our "Abowd Family Tree" (i.e., his academic descendants). Thank you to the various mentors who have offered me great advice, be it at UW, at internships, my alma maters, and outside of school or work. Many thanks to all my friends who helped me either directly with my work or in balancing my life outside of work, be it to play Dungeons and Dragons, ultimate frisbee, chamber music, board games, or many of our other activities. Special thanks to my partner Dominik Moritz, who has been a tremendous support over the last few years of my PhD (and not just because he makes the awesome data-visualization tool I've used in many parts of this document).

Bibliography

[1] Emmanuel Agu, Peder Pedersen, Diane Strong, Bengisu Tulu, Qian He, Lei Wang, and Yejin Li. "The smartphone as a medical device: Assessing enablers, benefits and challenges". In: *Internet-of-Things Networking and Control (IoT-NC), 2013 IEEE International Workshop of*. IEEE. 2013, pp. 48–52. DOI: 10.1109/IoT-NC.2013.6694053.

[2] İpek Akman, Çiğdem Arikan, Hülya Bilgen, Sibel Kalaça, and Eren Özek. "Transcutaneous measurement of bilirubin by icterometer during phototherapy on a bilibed". In: *Turkish Journal of Medical Sciences* 32.2 (2002), pp. 165–168. DOI: http://journals.tubitak.gov.tr/medical/abstract.htm?id=5393.

[3] Mari Akre, Marsha Finkelstein, Mary Erickson, Meixia Liu, Laurel Vanderbilt, and Glenn Billman. "Sensitivity of the pediatric early warning score to identify patient deterioration". In: *Pediatrics* (2010), peds–2009. DOI: 10.1542/peds.2009-0338.

[4] Ruben Amarasingham, Billy J Moore, Ying P Tabak, Mark H Drazner, Christopher A Clark, Song Zhang, W Gary Reed, Timothy S Swanson, Ying Ma, and Ethan A Halm. "An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data". In: *Medical care* (2010), pp. 981–988.

[5] Ruben Amarasingham, Parag C Patel, Kathleen Toto, Lauren L Nelson, Timothy S Swanson, Billy J Moore, Bin Xie, Song Zhang, Kristin S Alvarez, Ying Ma, et al. "Allocating scarce resources in real-time to reduce heart failure readmissions: a prospective, controlled study". In: *BMJ Qual Saf* (2013), bmjqs–2013. DOI: 10.1136/bmjqs-2013-001901.

[6] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks." In: *ProPublica* (Mar. 23, 2016).

[7] Virginia Apgar. "The newborn (Apgar) scoring system". In: *Pediatr Clin North Am* 13.3 (1966), pp. 645–50.

[8] *Baby Connect*. Seacloud Software. 2009. URL: https://www.babyconnect.com/ (visited on 11/06/2018).

[9] *Baby Products with SIDS Prevention Claims*. U.S. Food & Drug Administration. Oct. 3, 2018. URL: https://www.fda.gov/medicaldevices/productsandmedicalprocedures/sidspreventionclaims/default.htm (visited on 11/20/2018).

[10] Richard G Baraniuk. "Compressive sensing". In: *IEEE signal processing magazine* 24.4 (2007).

[11] David W Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. "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. DOI: 10.1377/hlthaff.2014.0041.

[12] *bempu*. Bempu. 2018. URL: https://bempu.com/ (visited on 11/06/2018).

[13] Vinod K Bhutani, Glenn R Gourley, Saul Adler, Bill Kreamer, Chris Dalin, and Lois H Johnson. "Noninvasive measurement of total serum bilirubin in a multiracial predischarge newborn population to assess the risk of severe hyperbilirubinemia". In: *Pediatrics* 106.2 (2000), e17–e17. DOI: 10.1542/peds.106.2.e17.

[14] 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. DOI: 10.1542/peds.103.1.6.

[15] Vinod K Bhutani, Ann R Stark, Laura C Lazzeroni, Ronald Poland, Glenn R Gourley, Steve Kazmierczak, Linda Meloy, Anthony E Burgos, Judith Y Hall, David K Stevenson, et al. "Predischarge screening for severe neonatal hyperbilirubinemia identifies infants who need phototherapy". In: *The Journal of pediatrics* 162.3 (2013), pp. 477–482. DOI: 10.1016/j.jpeds.2012.08.022.

[16] Jeffrey P. Bigham. *The Coming AI Autumn*. Apr. 2016. URL: http://jeffreybigham.com/blog/2019/the-coming-ai-autumnn.html (visited on 07/07/2019).

[17] *bili-hut™*. Little Sparrows Technology. URL: https://little-sparrows-tech.com/ (visited on 10/30/2018).

[18] *BiliTool™*. BiliTool, Inc. 2016. URL: http://bilitool.org/ (visited on 10/03/2018).

[19] Michael Bingler, Lori A Erickson, Kimberly J Reid, Brian Lee, James O'Brien, Johnathan Apperson, Kathy Goggin, and Girish Shirali. "Interstage Outcomes in Infants With Single Ventricle Heart Disease Comparing Home Monitoring Technology to Three-Ring Binder Documentation: A Randomized Crossover Study". In: *World Journal for Pediatric and Congenital Heart Surgery* 9.3 (2018), pp. 305–314. DOI: 10.1177/2150135118762401.

[20] Christopher P Bonafide, David T Jamison, and Elizabeth E Foglia. "The emerging market of smartphone-integrated infant physiologic monitors". In: *Jama* 317.4 (2017), pp. 353–354. DOI: 10.1001/jama.2016.19137.

[21] Nem-Yun Boo and Shareena Ishak. "Prediction of severe hyperbilirubinaemia using the Bilicheck transcutaneous bilirubinometer". In: *Journal of paediatrics and child health* 43.4 (2007), pp. 297–302. DOI: 10.1111/j.1440-1754.2007.01062.x.

[22] Cati Boulanger, Adam Boulanger, Lilian de Greef, Andy Kearney, Kiley Sobel, Russell Transue, Z Sweedyk, Paul H Dietz, and Steven Bathiche. "Stroke rehabilitation with a sensing surface". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2013, pp. 1243–1246. DOI: 10.1145/2470654.2466160.

[23] Abderrahim Bourouis, Mohammed Feham, M Alamgir Hossain, and Li Zhang. "An intelligent mobile based decision support system for retinal disease diagnosis". In: *Decision Support Systems* 59 (2014), pp. 341–350. DOI: 10.1016/j.dss.2014.01.005.

[24] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[25] *Brilliance: world-class jaundice treatment at an affordable price*. D-Rev. URL: http://d-rev.org/projects/newborn-health/.

[26] Paola Brunori, Piergiorgio Masi, Luigi Faggiani, Luciano Villani, Michele Tronchin, Claudio Galli, Clarissa Laube, Antonella Leoni, Maila Demi, and Antonio La Gioia. "Evaluation of bilirubin concentration in hemolysed samples, is it really impossible? The altitude-curve cartography approach to interfered assays". In: *Clinica Chimica Acta* 412.9-10 (2011), pp. 774–777. DOI: 10.1016/j.cca.2011.01.010.

[27] Bryan Burke, James Robbins, Charlotte Hobbs, et al. "American Academy of Pediatrics Subcommittee on Hyperbilirubinemia Management of hyperbilirubinemia in the newborn infant 35 or more weeks of gestation". In: *Pediatrics* 114.1 (2004), pp. 297–316.

[28] A Callahan. "New type of baby monitors offers "peace of mind" but may deliver just the opposite". In: *Washington Post* (2017).

[29] Yao-Jen Chang, Shu-Fang Chen, and Jun-Da Huang. "A Kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities". In: *Research in developmental disabilities* 32.6 (2011), pp. 2566–2570. DOI: 10.1016/j.ridd.2011.07.002.

[30] Nan-Chen Chen, Kuo-Cheng Wang, and Hao-Hua Chu. "Listen-to-nose: a low-cost system to record nasal symptoms in daily life". In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM. 2012, pp. 590–591.

[31] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[32] Ajay Kumar Chowdhary, Sudipta Dutta, and Rabindranath Ghosh. "Neonatal Jaundice Detection using Colour Detection Method". In: *International Advanced Research Journal in Science, Engineering and Technology* 4.7 (2017). DOI: 10.17148/IARJSET.2017.4733.

[33] Carolyn M Clancy. "Commentary: reducing hospital readmissions: aligning financial and quality incentives". In: *American Journal of Medical Quality* 27.5 (2012), pp. 441–443. DOI: 10.1177/1062860612452371.

[34] Sunny Consolvo, Predrag Klasnja, David W McDonald, Daniel Avrahami, Jon Froehlich, Louis LeGrand, Ryan Libby, Keith Mosher, and James A Landay. "Flowers or a robot army?: encouraging awareness & activity with personal, mobile displays". In: *Proceedings of the 10th international conference on Ubiquitous computing*. ACM. 2008, pp. 54–63. DOI: 10.1145/1409635.1409644.

[35] Ashlesha Datar and Neeraj Sood. "Impact of postpartum hospital-stay legislation on newborn length of stay, readmission, and mortality in California". In: *Pediatrics* 118.1 (2006), pp. 63–72. DOI: 10.1542/peds.2005-3044.

[36] Daniele De Luca, Gregory L Jackson, Ascanio Tridente, Virgilio P Carnielli, and William D Engle. "Transcutaneous bilirubin nomograms: a systematic review of population differences and analysis of bilirubin kinetics". In: *Archives of pediatrics & adolescent medicine* 163.11 (2009), pp. 1054–1059. DOI: 10.1001/archpediatrics.2009.187.

[37] Nicola Dell and Gaetano Borriello. "Mobile tools for point-of-care diagnostics in the developing world". In: *Proceedings of the 3rd ACM Symposium on Computing for Development*. ACM. 2013, p. 9. DOI: 10.1145/2442882.2442894.

[38] Alexander P Dumont, Brandon Harrison, Zachary T McCormick, Nishant Ganesh Kumar, and Chetan A Patil. "Development of mobile phone based transcutaneous billirubinometry". In: *Optics and Biophotonics in Low-Resource Settings III*. Vol. 10055. International Society for Optics and Photonics. 2017, 100550T. DOI: 10.1117/12.2257428.

[39] F Ebbesen, LM Rasmussen, and PD Wimberley. "A new transcutaneous bilirubinometer, BiliCheck, used in the neonatal intensive care unit and the maternity ward". In: *Acta Paediatrica* 91.2 (2002), pp. 203–211. DOI: 10.1111/j.1651-2227.2002.tb01696.x.

[40] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. "Least angle regression". In: *The Annals of statistics* 32.2 (2004), pp. 407–499. DOI: 10.1214/009053604000000067.

[41] William D Engle, Gregory L Jackson, Dorothy Sendelbach, Denise Manning, and William H Frawley. "Assessment of a transcutaneous device in the evaluation of neonatal hyperbilirubinemia in a primarily Hispanic population". In: *Pediatrics* 110.1 (2002), pp. 61–67. DOI: 10.1542/peds.110.1.61.

[42] William D Engle, Gregory L Jackson, Elizabeth K Stehel, Dorothy M Sendelbach, and M Denise Manning. "Evaluation of a transcutaneous jaundice meter following hospital discharge in term and near-term neonates". In: *Journal of perinatology* 25.7 (2005), p. 486. DOI: 10.1038/sj.jp.7211333.

[43] Gabriel J Escobar, Karen M Puopolo, Soora Wi, Benjamin J Turk, Michael W Kuzniewicz, Eileen M Walsh, Thomas B Newman, John Zupancic, Ellice Lieberman, and David Draper. "Stratification of risk of early-onset sepsis in newborns $\geq$ 34 weeks' gestation". In: *Pediatrics* 133.1 (2014), pp. 30–36. DOI: 10.1542/peds.2013-1689.

[44] G Duncan Finlay, Michael J Rothman, and Robert A Smith. "Measuring the modified early warning score and the Rothman index: advantages of utilizing the electronic medical record in an early warning system". In: *Journal of hospital medicine* 9.2 (2014), pp. 116–119. DOI: 10.1002/jhm.2132.

[45] Mieczyslaw Finster and Margaret Wood. "The Apgar score has survived the test of time". In: *The Journal of the American Society of Anesthesiologists* 102.4 (2005), pp. 855–857.

[46] *Firefly Phototherapy*. LifeKit by MTTS. URL: http://www.mtts-asia.com/firefly-phototherapy/ (visited on 11/03/2018).

[47] Sotirios Fouzas, Lito Mantagou, Eleni Skylogianni, Stefanos Mantagos, and Anastasia Varvarigou. "Transcutaneous bilirubin levels for the first 120 postnatal hours in healthy neonates". In: *Pediatrics* 125.1 (2010), e52–e57. DOI: 10.1542/peds.2009-0403.

[48] 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.

[49] Yoav Freund and Robert E Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of computer and system sciences* 55.1 (1997), pp. 119–139. DOI: 10.1006/jcss.1997.1504.

[50] NS Ghanayem, GM Hoffman, KA Mussatto, JR Cava, PC Frommelt, NA Rudd, MM Steltzer, SM Bevandic, SJ Frisbee, RDB Jaquiss, et al. "Home surveillance program prevents interstage mortality after the Norwood procedure". In: *The Journal of thoracic and cardiovascular surgery* 126.5 (2003), pp. 1367–1375. DOI: 10.1016/S0022-5223(03)00071-0.

[51] Lilian de Greef, Mayank Goel, Min Joon Seo, Eric C Larson, James W Stout, James A Taylor, and Shwetak N Patel. "Bilicam: using mobile phones to monitor newborn jaundice". In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM. 2014, pp. 331–342. DOI: 10.1145/2632048.2632076.

[52] David Anthony Green, Amod Kumar, and Rajesh Khanna. "Neonatal hypothermia detection by ThermoSpot in Indian urban slum dwellings". In: *Archives of Disease in Childhood-Fetal and Neonatal Edition* 91.2 (2006), F96–F98. DOI: 10.1136/adc.2005.078410.

[53] Maya R Gupta, Eric K Garcia, and Erika Chin. "Adaptive local linear regression with application to printer color management". In: *IEEE Transactions on Image Processing* 17.6 (2008), pp. 936–945. DOI: 10.1109/TIP.2008.922429.

[54] Paul HP Hanel and Katia C Vione. "Do student samples provide an accurate estimate of the general public?" In: *PloS one* 11.12 (2016), e0168354.

[55] RE Hannemann, DP DeWitt, and JF Wiechel. "Neonatal serum bilirubin from skin reflectance". In: *Pediatric research* 12.3 (1978), p. 207. DOI: 10.1203/00006450-197803000-00009.

[56] Trevor Hastie and Robert Tibshirani. "Generalized additive models: some applications". In: *Journal of the American Statistical Association* 82.398 (1987), pp. 371–386.

[57] Gillian R Hayes, Karen G Cheng, Sen H Hirano, Karen P Tang, Marni S Nagel, and Dianne E Baker. "Estrellita: a mobile capture and access tool for the support of preterm infants and their caregivers". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 21.3 (2014), p. 19. DOI: 10.1145/2617574.

[58] Javier Hernández-Andrés, Raymond L Lee, and Javier Romero. "Calculating correlated color temperatures across the entire gamut of daylight and skylight chromaticities". In: *Applied optics* 38.27 (1999), pp. 5703–5709. DOI: 10.1364/AO.38.005703.

[59] Arthur E Hoerl and Robert W Kennard. "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1 (1970), pp. 55–67. DOI: 10.1080/00401706.1970.10488634.

[60] Stephen F Jencks, Mark V Williams, and Eric A Coleman. "Rehospitalizations among patients in the Medicare fee-for-service program". In: *New England Journal of Medicine* 360.14 (2009), pp. 1418–1428. DOI: 10.1056/NEJMsa0803563.

[61] Elizabeth Johansen. "Making Human Factors Affordable for Medical Device and Global Health Startups". In: *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*. Vol. 7. 1. SAGE Publications Sage India: New Delhi, India. 2018, pp. 140–147. DOI: 10.1177/2327857918071036.

[62] David King. "Marketing wearable home baby monitors: real peace of mind?" In: *Bmj* 349 (2014), g6639. DOI: 10.1136/bmj.g6639.

[63] Robert P Kocher and Eli Y Adashi. "Hospital readmissions and the Affordable Care Act: paying for coordinated quality care". In: *Jama* 306.16 (2011), pp. 1794–1795. DOI: 10.1001/jama.2011.1561.

[64] William H Kruskal and W Allen Wallis. "Use of ranks in one-criterion variance analysis". In: *Journal of the American statistical Association* 47.260 (1952), pp. 583–621. DOI: 10.1080/01621459.1952.10483441.

[65] V Kumar, JC Shearer, A Kumar, and GL Darmstadt. "Neonatal hypothermia in low resource settings: a review". In: *Journal of Perinatology* 29.6 (2009), p. 401. DOI: 10.1038/jp.2008.233.

[66] Michael W Kuzniewicz, Gabriel J Escobar, and Thomas B Newman. "Impact of universal bilirubin screening on severe hyperbilirubinemia and phototherapy use". In: *Pediatrics* 124.4 (2009), pp. 1031–1039. DOI: 10.1542/peds.2008-2980.

[67] Eric C Larson, Mayank Goel, Gaetano Boriello, Sonya Heltshe, Margaret Rosenfeld, and Shwetak N Patel. "SpiroSmart: using a microphone to measure lung function on a mobile phone". In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM. 2012, pp. 280–289. DOI: 10.1145/2370216.2370261.

[68] Eric C Larson, TienJui Lee, Sean Liu, Margaret Rosenfeld, and Shwetak N Patel. "Accurate and privacy preserving cough sensing using a low-cost microphone". In: *Proceedings of the 13th international conference on Ubiquitous computing*. ACM. 2011, pp. 375–384. DOI: 10.1145/2030112.2030163.

[69] 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.

[70] Hoshik Lee, Craig G Rusin, Douglas E Lake, Matthew T Clark, Lauren Guin, Terri J Smoot, Alix O Paget-Brown, Brooke D Vergales, John Kattwinkel, J Randall Moorman, et al. "A new algorithm for detecting central apnea in neonates". In: *Physiological measurement* 33.1 (2011), p. 1. DOI: 10.1088/0967-3334/33/1/1.

[71] Jinseok Lee, Bersain A Reyes, David D McManus, Oscar Maitas, and Ki H Chon. "Atrial fibrillation detection using an iPhone 4S". In: *IEEE Transactions on Biomedical Engineering* 60.1 (2013), pp. 203–206. DOI: 10.1109/TBME.2012.2208112.

[72] Jinseok Lee, Bersain A Reyes, David D McManus, Oscar Mathias, and Ki H Chon. "Atrial fibrillation detection using a smart phone". In: *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE. 2012, pp. 1177–1180. DOI: 10.1109/EMBC.2012.6346146.

[73] Terence S Leung, Karan Kapur, Ashley Guilliam, Jade Okell, Bee Lim, Lindsay W MacDonald, and Judith Meek. "Screening neonatal jaundice based on the sclera color of the eye using digital photography". In: *Biomedical optics express* 6.11 (2015), pp. 4529–4538. DOI: 10.1364/BOE.6.004529.

[74] Shai Linn, Stephen C Schoenbaum, Richard R Monson, Bernard Rosner, Phillip G Stubblefield, and Kenneth J Ryan. "Epidemiology of neonatal hyperbilirubinemia". In: *Pediatrics* 75.4 (1985), pp. 770–774.

[75] Christopher Longhurst, Stuart Turner, and Anthony E Burgos. "Development of a Web-based decision support tool to increase use of neonatal hyperbilirubinemia guidelines". In: *The Joint Commission Journal on Quality and Patient Safety* 35.5 (2009), pp. 256–262. DOI: 10.1016/S1553-7250(09)35035-7.

[76] David JC MacKay. "Bayesian interpolation". In: *Neural computation* 4.3 (1992), pp. 415–447. DOI: 10.1162/neco.1992.4.3.415.

[77] M Jeffrey Maisels. "Managing the jaundiced newborn: a persistent challenge". In: *CMAJ* 187.5 (2015), pp. 335–343. DOI: 10.1503/cmaj.122117.

[78] M Jeffrey Maisels. "Transcutaneous bilirubin measurement: does it work in the real world?" In: *Pediatrics* 135.2 (2015), pp. 364–366. DOI: 10.1542/peds.2014-3472.

[79] M Jeffrey Maisels, Vinod K Bhutani, Debra Bogen, Thomas B Newman, Ann R Stark, and Jon F Watchko. "Hyperbilirubinemia in the newborn infant $\geq$ 35 weeks' gestation: an update with clarifications". In: *Pediatrics* 124.4 (2009), pp. 1193–1198. DOI: 10.1542/peds.2009-0329.

[80] M Jeffrey Maisels, Vinod K Bhutani, Debra Bogen, Thomas B Newman, Ann R Stark, and Jon F Watchko. "Hyperbilirubinemia in the newborn infant geq 35 weeks' gestation: an update with clarifications". In: *Pediatrics* 124.4 (2009), pp. 1193–1198. DOI: 10.1542/peds.2009-0329.

[81] M Jeffrey Maisels, Enrique M Ostrea, Suzanne Touch, Sarah E Clune, Eugene Cepeda, Elizabeth Kring, Karin Gracey, Cheryl Jackson, Deborah Talbot, and Raywin Huang. "Evaluation of a new transcutaneous bilirubinometer". In: *Pediatrics* 113.6 (2004), pp. 1628–1635. DOI: 10.1542/peds.113.6.1628.

[82] MJ Maisels, WD Engle, S Wainer, GL Jackson, S McManus, and F Artinian. "Transcutaneous bilirubin levels in an outpatient and office population". In: *Journal of Perinatology* 31.9 (2011), p. 621. DOI: 10.1038/jp.2011.5.

[83] Alex Mariakakis, Megan A Banks, Lauren Phillipi, Lei Yu, James Taylor, and Shwetak N Patel. "Biliscreen: smartphone-based scleral jaundice monitoring for liver and pancreatic disorders". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.2 (2017), p. 20. DOI: 10.1145/3090085.

[84] Mary C McLellan and Jean A Connor. "The cardiac children's hospital early warning score (C-CHEWS)". In: *Journal of pediatric nursing* 28.2 (2013), pp. 171–178. DOI: 10.1016/j.pedn.2012.07.009.

[85] *Mimo*. Rest Devices, Inc. 2018. URL: https://www.mimobaby.com/ (visited on 11/06/2018).

[86] Thomas B Mole, Neil Kennedy, Noel Ndoya, and Alan Emond. "ThermoSpots to detect hypothermia in children with severe acute malnutrition". In: *PLoS One* 7.9 (2012), e45823. DOI: 10.1371/journal.pone.0045823.

[87] Alan Monaghan. "Detecting and managing deterioration in children". In: *Paediatric nursing* 17.1 (2005), p. 32.

[88] Rachel Y Moon, Task Force on Sudden Infant Death Syndrome, et al. "SIDS and other sleep-related infant deaths: evidence base for 2016 updated recommendations for a safe infant sleeping environment". In: *Pediatrics* (2016), e20162940. DOI: 10.1542/peds.2016-2940.

[89] Sarah B Munkholm, Tobias Krøgholt, Finn Ebbesen, Pal B Szecsi, and Søren R Kristensen. "The smartphone camera as a potential method for transcutaneous bilirubin measurement". In: *PloS one* 13.6 (2018), e0197938. DOI: 10.1371/journal.pone.0197938.

[90] Travis B Murdoch and Allan S Detsky. "The inevitable application of big data to health care". In: *Jama* 309.13 (2013), pp. 1351–1352. DOI: 10.1001/jama.2013.393.

[91] Patrick W O'Leary. "Prevalence, clinical presentation and natural history of patients with single ventricle". In: *Progress in Pediatric Cardiology* 16.1 (2002), pp. 31–38. DOI: 10.1016/S1058-9813(02)00042-5.

[92] Felix Outlaw, Judith Meek, Lindsay W MacDonald, and Terence S Leung. "Screening for neonatal jaundice with a smartphone". In: *Proceedings of the 2017 International Conference on Digital Health*. ACM. 2017, pp. 241–242. DOI: 10.1145/3079452.3079488.

[93] *Owlet*. Owlet. 2018. URL: https://owletcare.com/ (visited on 11/06/2018).

[94] Vitor F Pamplona, Ankit Mohan, Manuel M Oliveira, and Ramesh Raskar. "NETRA: interactive display for estimating refractive errors and focal range". In: *ACM transactions on graphics (TOG)*. Vol. 29. 4. ACM. 2010, p. 77. DOI: 10.1145/1778765.1778814.

[95] Perzen Patel. *ClikJaundice: Using the phone to monitor jaundice in newborns*. June 12, 2013. URL: http://www.thealternative.in/business/clickjaundice-using-the-phone-to-prevent-jaundice-in-newborns/ (visited on 10/25/2018).

[96] *Pediatric Medical Devices*. U.S. Food & Drug Administration. Oct. 3, 2018. URL: https://www.fda.gov/medicaldevices/productsandmedicalprocedures/ucm135104.htm (visited on 11/20/2018).

[97] American Academy of Pediatrics et al. "Apnea, sudden infant death syndrome, and home monitoring". In: *Pediatrics* 111 (2003), pp. 914–917. DOI: 10.1542/peds.111.4.914.

[98] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. "Scikit-learn: Machine learning in Python". In: *Journal of Machine Learning Research* 12.Oct (2011), pp. 2825–2830.

[99] Ranjan Kumar Pejaver, R Nisarga, and B Gowda. "Temperature monitoring in newborns using thermospot". In: *The Indian Journal of Pediatrics* 71.9 (2004), pp. 795–796. DOI: 10.1007/BF02730715.

[100] Ming-Zher Poh, Kyunghee Kim, Andrew D Goessling, Nicholas C Swenson, and Rosalind W Picard. "Heartphones: Sensor earphones and mobile application for non-obtrusive health monitoring". In: *Wearable Computers, 2009. ISWC'09. International Symposium on*. IEEE. 2009, pp. 153–154. DOI: 10.1109/ISWC.2009.35.

[101] Ronald L Poland, Carol Hartenberger, Helen McHenry, and Andrew Hsi. "Comparison of skin sites for estimating serum total bilirubin in in-patients and out-patients: chest is superior to brow". In: *Journal of perinatology* 24.9 (2004), p. 541. DOI: 10.1038/sj.jp.7211141.

[102] Urs Ramer. "An iterative procedure for the polygonal approximation of plane curves". In: *Computer graphics and image processing* 1.3 (1972), pp. 244–256. DOI: 10.1016/S0146-664X(72)80017-0.

[103] Janet Rennie, Shona Burman-Roy, and M Stephen Murphy. "Neonatal jaundice: summary of NICE guidance". In: *Bmj* 340 (2010), p. c2409. DOI: 10.1136/bmj.c2409.

[104] Michael J Rothman, Steven I Rothman, and Joseph Beals IV. "Development and validation of a continuous measure of patient condition using the Electronic Medical Record". In: *Journal of biomedical informatics* 46.5 (2013), pp. 837–848. DOI: 10.1016/j.jbi.2013.06.011.

[105] Steven I Rothman, Michael J Rothman, and Alan B Solinger. "Placing clinical variables on a common linear scale of empirically based risk as a step towards construction of a general patient acuity score from the electronic health record: a modelling study". In: *BMJ open* 3.5 (2013), e002367. DOI: 10.1136/bmjopen-2012-002367.

[106] Firmino F Rubaltelli, Glenn R Gourley, Norbert Loskamp, Neena Modi, Matthias Roth-Kleiner, Alfred Sender, and Paul Vert. "Transcutaneous bilirubin measurement: a multicenter evaluation of a new device". In: *Pediatrics* 107.6 (2001), pp. 1264–1271. DOI: 10.1542/peds.107.6.1264.

[107] Nancy A Rudd, Michele A Frommelt, James S Tweddell, David A Hehir, Kathleen A Mussatto, Katherine D Frontier, Julie A Slicker, Peter J Bartz, and Nancy S Ghanayem. "Improving interstage survival after Norwood operation: outcomes from 10 years of home monitoring". In: *The Journal of thoracic and cardiovascular surgery* 148.4 (2014), pp. 1540–1547. DOI: 10.1016/j.jtcvs.2014.02.038.

[108] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. "Learning representations by back-propagating errors". In: *Cognitive modeling* 5.3 (1988), p. 1.

[109] Craig G Rusin, Sebastian I Acosta, Lara S Shekerdemian, Eric L Vu, Aarti C Bavare, Risa B Myers, Lance W Patterson, Ken M Brady, and Daniel J Penny. "Prediction of imminent, severe deterioration of children with parallel circulations using real-time processing of physiologic data". In: *The Journal of thoracic and cardiovascular surgery* 152.1 (2016), pp. 171–177. DOI: 10.1016/j.jtcvs.2016.03.083.

[110] S Samanta, M Tan, C Kissack, S Nayak, R Chittick, and CW Yoxall. "The value of Bilicheck as a screening tool for neonatal jaundice in term and near-term babies". In: *Acta Paediatrica* 93.11 (2004), pp. 1486–1490. DOI: 10.1111/j.1651-2227.2004.tb02634.x.

[111] Suwimol Sanpavat and Issarang Nuchprayoon. "Comparison of two transcutaneous bilirubinometers-Minolta Airshields Jaundice Meter JM 103 and SpectRx bilicheck-in Thai neotates". In: *Southeast Asian Journal of Tropical Medicine and Public Health* 36.6 (2005), p. 1533.

[112] Suwimol Sanpavat and Issarang Nuchprayoon. "Noninvasive transcutaneous bilirubin as a screening test to identify the need for serum bilirubin assessment". In: *Journal - Medical Association of Thailand* 87.10 (2004), pp. 1193–1198.

[113] Suchi Saria, Daphne Koller, and Anna Penn. "Learning individual and population level traits from clinical temporal data". In: *Proceedings of Neural Information Processing Systems*. Citeseer. 2010, pp. 1–9.

[114] Suchi Saria, Anand K Rajani, Jeffrey Gould, Daphne Koller, and Anna A Penn. "Integration of early physiological responses predicts later illness severity in preterm infants". In: *Science translational medicine* 2.48 (2010), 48ra65–48ra65. DOI: 10.1126/scitranslmed.3001304.

[115] Gaurav Sharma, Wencheng Wu, and Edul N Dalal. "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations". In: *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur* 30.1 (2005), pp. 21–30. DOI: 10.1002/col.20070.

[116] 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.

[117] Girish Shirali, Lori Erickson, Jonathan Apperson, Kathy Goggin, David Williams, Kimberly Reid, Andrea Bradley-Ewing, Dawn Tucker, Michael Bingler, John Spertus, et al. "Harnessing teams and technology to improve outcomes in infants with single ventricle". In: *Circulation: Cardiovascular Quality and Outcomes* 9.3 (2016), pp. 303–311. DOI: 10.1161/CIRCOUTCOMES.115.002452.

[118] Yuichi Shoda, Walter Mischel, and Philip K Peake. "Predicting adolescent cognitive and self-regulatory competencies from preschool delay of gratification: Identifying diagnostic conditions." In: *Developmental psychology* 26.6 (1990), p. 978.

[119] Ruchika Singla and Surender Singh. "A framework for detection of jaundice in new born babies using homomorphic filtering based image processing". In: *Inventive Computation Technologies (ICICT), International Conference on*. Vol. 3. IEEE. 2016, pp. 1–5. DOI: 10.1109/INVENTIVE.2016.7830209.

[120] Tina M Slusher, Ishaya A Angyo, Fidela Bode-Thomas, Francis Akor, Sunday D Pam, Adedotun A Adetunji, Donald W McLaren, Ronald J Wong, Hendrik J Vreman, and David K Stevenson. "Transcutaneous bilirubin measurements and serum total bilirubin levels in indigenous African infants". In: *Pediatrics* 113.6 (2004), pp. 1636–1641. DOI: 10.1542/peds.113.6.1636.

[121] Tina M Slusher, Alvin Zipursky, and Vinod K Bhutani. "A global need for affordable neonatal jaundice technologies". In: *Seminars in perinatology*. Vol. 35. 3. Elsevier. 2011, pp. 185–191. DOI: 10.1053/j.semperi.2011.02.014.

[122] Alex J Smola and Bernhard Schölkopf. "A tutorial on support vector regression". In: *Statistics and computing* 14.3 (2004), pp. 199–222. DOI: 10.1023/B:STCO.0000035301.49549.88.

[123] *snuza*. Snuza International. 2017. URL: https://www.snuza.com/ (visited on 11/06/2018).

[124] Eileen K Steinberger, Charlotte Ferencz, and Christopher A Loffredo. "Infants with single ventricle: A population-based epidemiological study". In: *Teratology* 65.3 (2002), pp. 106–115. DOI: 10.1002/tera.10017.

[125] Amr T Sufian, Gordon R Jones, Hameed M Shabeer, Ezzaldeen Y Elzagzoug, and Joseph W Spencer. "Chromatic techniques for in vivo monitoring jaundice in neonate tissues". In: *Physiological measurement* 39.9 (2018), p. 095004. DOI: 10.1088/1361-6579/aadbdb.

[126] Satoshi Suzuki and Keiichi Abe. "Topological structural analysis of digitized binary images by border following". In: *Computer vision, graphics, and image processing* 30.1 (1985), pp. 32–46. DOI: 10.1016/0734-189X(85)90016-7.

[127] Vasanthan Tanigasalam, B Vishnu Bhat, B Adhisivam, Bharathi Balachander, and Harichandra Kumar. "Hypothermia detection in low birth weight neonates using a novel bracelet device". In: *The Journal of Maternal-Fetal & Neonatal Medicine* (2018), pp. 1–4. DOI: 10.1080/14767058.2018.1443072.

[128] James A Taylor, Anthony E Burgos, Valerie Flaherman, Esther K Chung, Elizabeth A Simpson, Neera K Goyal, Isabelle Von Kohorn, Niramol Dhepyasuwan, BORN Investigators, et al. "Utility of decision rules for transcutaneous bilirubin measurements". In: *Pediatrics* 137.5 (2016), e20153032. DOI: 10.1542/peds.2015-3032.

[129] James A Taylor, Anthony E Burgos, Valerie Flaherman, Esther K Chung, Elizabeth A Simpson, Neera K Goyal, Isabelle Von Kohorn, Nui Dhepyasuwan, et al. "Discrepancies between transcutaneous and serum bilirubin measurements". In: *Pediatrics* 135.2 (2015), pp. 224–231. DOI: 10.1542/peds.2014-1919.

[130] James A Taylor, Shewtak N Patel Patel, James W Stout, Lilian de Greef, Mayank Goel, and Eric C Larson. "Estimating bilirubin levels". Pat. WO2014172033A1. Oct. 23, 2014.

[131] James A Taylor, Shewtak N Patel Patel, James W Stout, Lilian de Greef, Mayank Goel, and Eric C Larson. "Systems, devices, and methods for estimating bilirubin levels". Pat. US20150359459A1. May 14, 2019.

[132] James A Taylor, James W Stout, Lilian de Greef, Mayank Goel, Shwetak Patel, Esther K Chung, Aruna Koduri, Shawn McMahon, Jane Dickerson, Elizabeth A Simpson, et al. "Use of a smartphone app to assess neonatal jaundice". In: *Pediatrics* (2017), e20170312. DOI: 10.18203/2349-3291.ijcp20175928.

[133] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

[134] Kentaro Toyama. *Geek Heresy: Rescuing Social Change from the Cult of Technology*. PublicAffairs, 2015.

[135] Kentaro Toyama. "Technology as amplifier in international development". In: *Proceedings of the 2011 iConference*. ACM. 2011, pp. 75–82.

[136] *Trixie Tracker™*. Trixie Telemetry LLC. 2014. URL: https://www.trixietracker.com/ (visited on 11/06/2018).

[137] Gunnar Vartdal. "Development of a Smartphone-based diagnostic Tool for Jaundice". MA thesis. NTNU, 2014.

[138] JA Stephen Viggiano. "Comparison of the accuracy of different white-balancing options as quantified by their color constancy". In: *Sensors and Camera Systems for Scientific, Industrial, and Digital Photography Applications V*. Vol. 5301. International Society for Optics and Photonics. 2004, pp. 323–334. DOI: 10.1117/12.524922.

[139] Tarun Wadhawan, Ning Situ, Hu Rui, Keith Lancaster, Xiaojing Yuan, and George Zouridakis. "Implementation of the 7-point checklist for melanoma detection on smart handheld devices". In: *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE. 2011, pp. 3180–3183. DOI: 10.1109/IEMBS.2011.6090866.

[140] Edward Jay Wang, William Li, Doug Hawkins, Terry Gernsheimer, Colette Norby-Slycord, and Shwetak N Patel. "HemaApp: noninvasive blood screening of hemoglobin using smartphone cameras". In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM. 2016, pp. 593–604.

[141] Lei Wang, Peder C Pedersen, Diane Strong, Bengisu Tulu, and Emmanuel Agu. "Wound image analysis system for diabetics". In: *Medical Imaging 2013: Image Processing*. Vol. 8669. International Society for Optics and Photonics. 2013, p. 866924. DOI: 10.1117/12.2004762.

[142] Tyler W Watts, Greg J Duncan, and Haonan Quan. "Revisiting the marshmallow test: A conceptual replication investigating links between early delay of gratification and later outcomes". In: *Psychological science* 29.7 (2018), pp. 1159–1177.

[143] Stephen Wolf. *Color correction matrix for digital still and video imaging systems*. National Telecommunications and Information Administration Washington, DC, 2003.

[144] National Collaborating Centre for Women's, Children's Health (UK, et al. "Neonatal jaundice". In: (2010).

[145] CD Coda Zabetta, IF Iskander, C Greco, C Bellarosa, S Demarini, C Tiribelli, and RP Wennberg. "Bilistick: a low-cost point-of-care system to measure total plasma bilirubin". In: *Neonatology* 103.3 (2013), pp. 177–181. DOI: 10.1159/000345425.

[146] Kiyana Zolfaghar, Naren Meadem, Ankur Teredesai, Senjuti Basu Roy, Si-Chi Chin, and Brian Muckian. "Big data solutions for predicting risk-of-readmission for congestive heart failure patients". In: *Big Data, 2013 IEEE International Conference on*. IEEE. 2013, pp. 64–71. DOI: 10.1109/BigData.2013.6691760.

[147] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x.