BiliCam: Using Mobile Phones to Monitor Newborn Jaundice

Lilian de Greef[1], Mayank Goel[1], Min Joon Seo[1], Eric C. Larson[2], James W. Stout MD MPH[3], James A. Taylor MD[3], Shwetak N. Patel[1]

[1]Computer Science & Engineering, DUB Group
[3]Department of Pediatrics
University of Washington
Seattle, WA 98195

[2]Computer Science & Engineering
Southern Methodist University
Dallas, TX 75205

{ldegreef, mayankg, minjoon, shwetak, jstout, uncjat}@uw.edu, eclarson@lyle.smu.edu

ABSTRACT

Health sensing through smartphones has received considerable attention in recent years because of the devices' ubiquity and promise to lower the barrier for tracking medical conditions. In this paper, we focus on using 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, we present BiliCam, a low-cost system that uses smartphone cameras to assess newborn jaundice. We evaluated BiliCam on 100 newborns, yielding a 0.85 rank order correlation with the gold standard blood test. We also discuss usability challenges and design solutions to make the system practical.

Author Keywords

Health sensing, mobile phones, neonatal jaundice, bilirubin, image processing.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

A number of smartphone-based medical devices are becoming increasingly common for fitness [13], heart rate monitoring [30], pulmonology [18,19], point-of-care diagnostics [14,33], and various other health applications [1]. They demonstrate how the capabilities and ubiquity of modern smartphones make them excellent candidates for clinical and health platforms, despite their inherent sensing limitations. In this paper, we present the design and critical evaluation of assessing newborn jaundice using a smartphone in a study with 100 newborns. Our prototype, BiliCam, is a smartphone-based medical device that uses the embedded camera and a paper-based color calibration card to monitor newborn jaundice.

Figure 1. Parents or medical practitioners can monitor a newborn's jaundice with their smartphones through BiliCam.

Jaundice is defined as the yellow discoloration of the skin caused by excess bilirubin, a chemical byproduct of recycling old blood cells. It is one of the most common physiological conditions among newborns: an estimated 84% of newborns develop jaundice [5]. A moderate level of bilirubin is normal in healthy newborns. However, if not treated, extreme levels can be fatal or cause devastating and irreversible brain damage. 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

UbiComp '14, September 13 – 17, 2014, Seattle, WA, USA

ACM 978-1-4503-2968-2/14/09···$15.00.
http://dx.doi.org/10.1145/2632048.2632076

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 [26]. The importance of monitoring newborn jaundice at home under these conditions creates the need for an accessible screening system such as BiliCam.

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.

Our research team includes experienced pediatricians who have helped to design and evaluate BiliCam. We conducted a clinical study to validate our smartphone-based approach with 100 newborn participants at the University of Washington Medical Center. We collected photographs of newborns with BiliCam within two hours of measuring total serum bilirubin levels (TSB), the medical gold standard. BiliCam compares to the TSB with a rank order correlation of 0.85 and a mean error of 2.0 mg/dl. We also compare BiliCam's results to that of a transcutaneous bilirubinometer (TcB), another non-invasive, although costly, technique that uses structured light to screen for high levels of bilirubin. We conclude that BiliCam cannot yet replace the TcB, but offers distinct cost and accessibility advantages that make it appropriate for screening newborns from home.

We describe the image processing and machine-learning techniques used to infer bilirubin and show its robustness with a range of skin colors and lighting conditions. In tandem with several pilot deployments, these studies also informed important user interface considerations and future designs of BiliCam.

EXPLANATION OF JAUNDICE AND TREATMENT

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 medical conditions in newborns; up to 84% of them develop jaundice during their first week of life [5]. This temporary excess of bilirubin is usually harmless. However, highly elevated concentrations of bilirubin in newborns are neurotoxic and can irreversibly damage the brain. 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 [26].

Medically Accepted Methods to Measure Bilirubin

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 [26]. One such graph is the Bhutani nomogram [4], like the one shown in Figure 2, 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) [26].

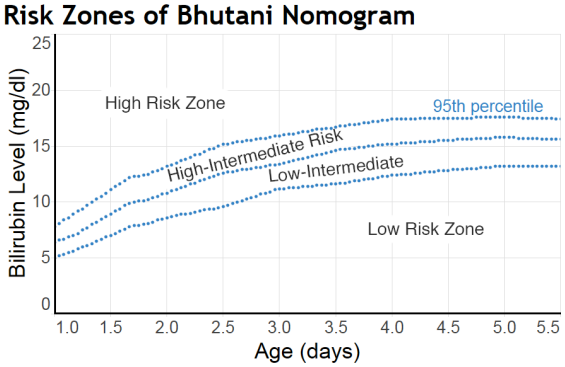

Figure 2. A Bhutani nomogram used to assess a newborn's risk based on bilirubin level and age, generated with risk zone boundaries from BiliTool™ [23,24].

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 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 absorbance 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 [26]. 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, we 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 assessement. 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 [26]. 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 [26]. Icterometers are specialized plexiglass rulers marked with different tones of yellow to reference when pressed against a newborn's skin [2]. Clinical guidelines explicitly advocate against using icterometers [26].

A number of studies investigate other 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 [26].

Although there are brief proposals currently posted online about using mobile phone cameras to measure neonatal jaundice [3,28], they provide little detail and do not mention any clinical studies. We believe this is the first work to describe, in detail, an approach for using mobile phones to screen neonatal jaundice. Furthermore, our methods are evaluated in a clinical setting and compared to the TSB and TcB measures of bilirubin.

RELATED WORK

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 relatively new space inspired and motivated the design of BiliCam. We divide our related work into two categories: smartphone-based health sensing and visual- or camera-based health sensing.

Smartphone-Based Health Sensing

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 [13]. 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 [30]. 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 [16]. 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. Our previous work uses the it to measure lung function (spirometry) in order to detect and monitor chronic lung conditions, and has achieved results akin to a clinical spirometer [18]. 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 [19]. Similarly, Chen *et al*. used the microphone to continually monitor nasal conditions, such as sneezing and runny nose [12].

Vision- or Camera-Based Health Sensing

Like BiliCam, a number of recent explorations of health applications are vision based. The most similar work comes from a medical group in Thailand who investigated the feasibility of screening newborn jaundice with a camera-based system similar to BiliCam. They found a correlation between color values from images of newborns, taken with a digital camera, and corresponding bilirubin levels. To do so, they manually inspected and adjusted each image in Photoshop [20]. 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 [21,22]. Researchers have also investigated assisting rehabilitative physical therapy using a depth camera [11] or with infra-red cameras in a touch-screen table [7]. 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 [14,33]. 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 [27]. Also examining the eye using a phone camera, Bourouis *et al*. developed a method of detecting retinal cancer [8]. Other active areas of research with computer vision include recognizing skin cancer [37] and tracking chronic foot ulcers from diabetes [38].

DATA COLLECTION

To evaluate and inform the design of BiliCam, we 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. We collected images within two hours of the TSB blood draw to ensure that bilirubin measures were as accurate as possible.

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. We 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 [34]. We also noted which blood samples were effected by hemolysis, a condition that affects the accuracy of TSB readings [10].

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 1. Demographic information for participants. Note that participants may report multiple ethnicities.

Medical professionals collected all of the images on iPhone 4S smartphones using a custom data collection app and the built-in camera. We 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 we ran with 40 newborn participants. The pilot study data is not included in our evaluation of BiliCam due to significant differences in study procedure.

Data Collection Timeline

We 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. We 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, we 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 we expect even lighting. We 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.

Data Collection Application

Figure 3 shows a screenshot of our 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, and one or more sets of photographs and videos. For taking photographs, it first instructs the placement of the color calibration card on the newborn (Figure 3B) and prompts

Figure 3. 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.

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 3C). These cues constrain the distance of the camera from the newborn and comfortably fit the card and newborn's sternum within the image. Our 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 problems arose. 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.

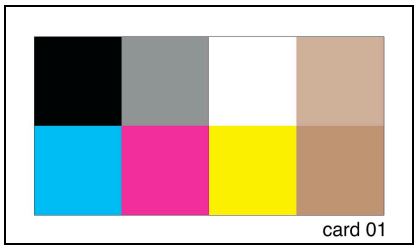

Figure 4: Design of the color calibration card used in our studies.

We 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% grey, white, cyan, magenta, yellow, and two skin tones (see Figure 4). 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. We used a new card for each sample to prevent spreading disease. When taking a sample, we 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.

ALGORITHM

We 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. Our 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. We explain each stage in turn and show a method outline in Figure 6. As an overview:

- color balancing is carried out using the calibration cards captured in each image
- for each skin patch, we estimate the mean red, green, and blue values, and the gradients of colors in the patch
- we 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

Color Balancing

We derive our features from the observed skin color, which can vary in different lighting conditions. To mitigate some of the effects of different lighting, we compute normalized red, green, and blue values. We 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, we include our color calibration card in each image for further color balancing. In order to use it, the system first needs to identify the location of the card and each of its color patches.

Image Segmentation

Although full automation and segmentation is not the focus of this proof of concept for BiliCam, we developed an algorithm to segment the card which we used in automatic image quality feedback when collecting data. We segmented skin patches by hand to reduce confounds.

The data collection UI constrained the card to a specific region on the image (Figure 3C). 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 5 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, we 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 [32].

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.

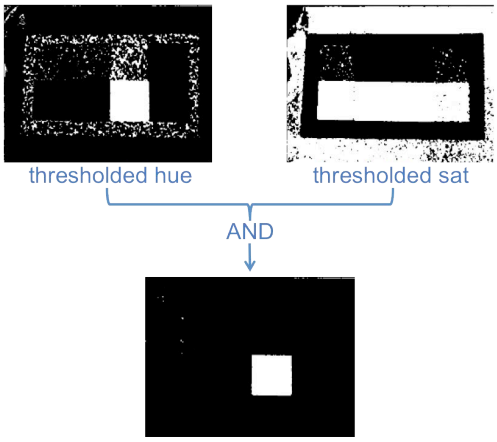

Figure 5. Segmentation of the yellow patch from a color calibration card.

White Balancing

We 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 [36].

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. We calculate their mean values for each color channel, resulting in 9 features.

In addition to color transformations, we 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. We 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). We also use principle 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.

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 6 shows a flowchart of our machine learning process.

We use an ensemble of five different regression algorithms. Each regression is discussed in turn. Most regressions are carried out using the *scikit-learn toolkit* [29] in Python. In order to avoid overfitting, we use 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.

k-Nearest Neighbor

The first regression algorithm is an encapsulated k-Nearest Neighbor regression (k = 7) [17]. Intuitively, this regression takes a more "local" estimate of the bilirubin level based upon training points that have similar feature values. In this regression, we 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

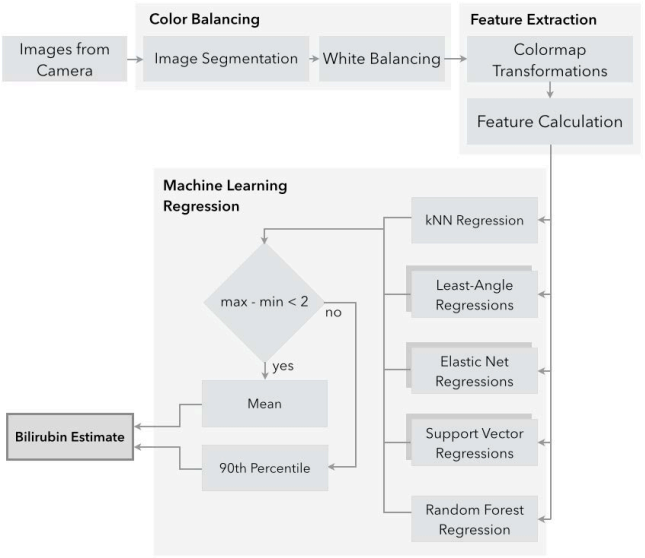

Figure 6. Flowchart of our algorithm.

neighbors are the first two components of the PCA transformation. We 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.

LARS

The second set of regressions uses least angle regression (LARS) [15]. 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. This prediction is subtracted from the output to obtain the residuals. Then, the algorithm attempts to find another feature with roughly the same correlation to the 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, we train two LARS regressions: one with scaled features and another with PCA components.

LARS-Lasso Elastic Net

The third regression uses the elastic net algorithm [39]. 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, we train two regressions, one with scaled features and another with PCA components.

SVR

All the regressions up to this stage were linear regressions. In order to capture the possible non-linear relationship, we employ two support vector regressions [35]. 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. We train two SVRs: the first uses a linear kernel and the second uses a nonlinear sigmoidal basis function.

Random Forest

The last algorithm uses random forest regression [9] 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 than the regressions discussed up to this point. The random forest uses scaled features only.

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.

RESULTS

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

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 underreport 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, we only focus on the performance of this ensemble for the rest of the paper.

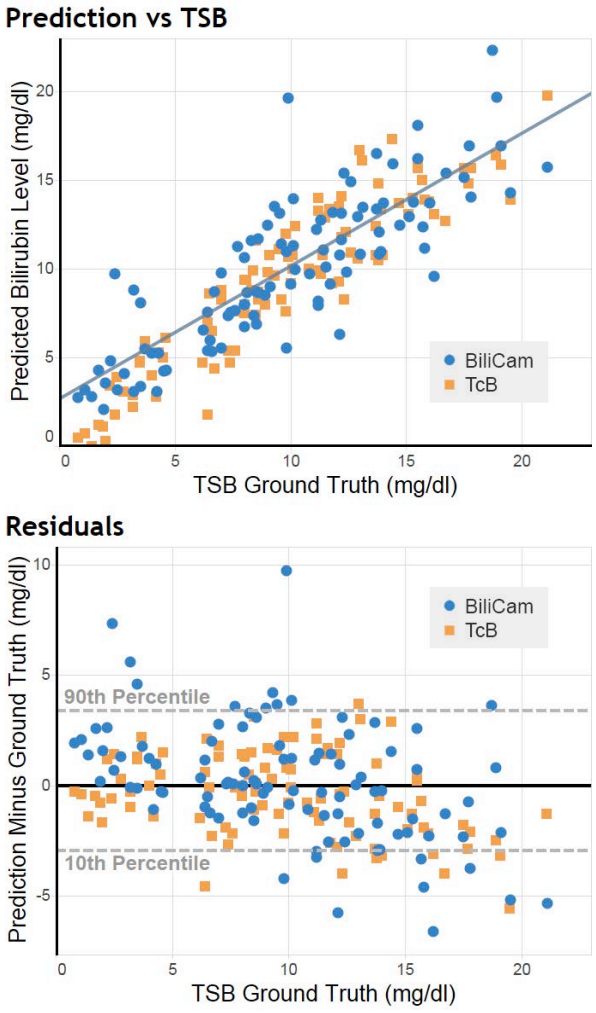

Figure 7. (Top) A comparison between predicted bilirubin levels and TSB. Blue circles represent predictions from BiliCam, orange squares predictions from the TcB. (Bottom) A Bland-Altman plot showing residuals of predicted bilirubin levels against the TSB.

Explanation: Figure 7 shows a scatter plot of BiliCam estimates (circles), calculated through leave-one-out cross-validation, compared to the TSB. It also shows a modified Bland-Altman plot [6], where residuals (BiliCam – TSB) are plotted against the TSB. For comparison, each plot also contains predictions from the TcB (squares). **Results**: Our 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. We 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 our 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 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.

We 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. We need more data to characterize the existence of outliers.

Predicting Newborn Risk

To evaluate how well BiliCam assesses a newborn's risk from jaundice, we 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* [4]. Our classifications are based on the predicted bilirubin levels and participant age at the time of sample, using risk zone boundaries defined by BiliTool™ [23,24].

Explanation: Figure 8 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 8's legend). Gray circles denote correct risk zone classification. **Result**: 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 [25]. To compare the effectiveness of BiliCam as such a screening tool, we 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*

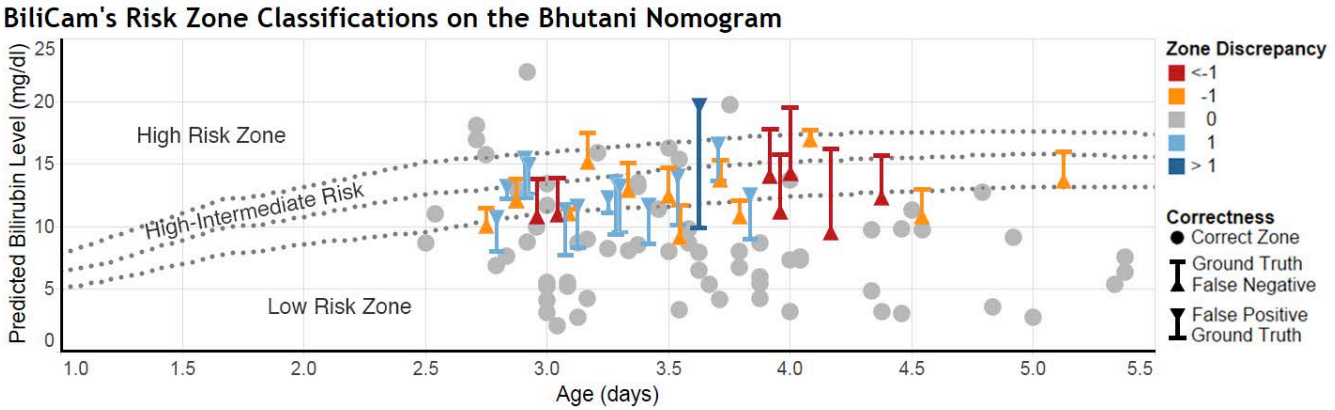

Figure 8. Risk zone classification of BiliCam using Bhutani nomogram. Colored lines map incorrectly classified points to the ground truth classification and bilirubin level.

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 our dataset. These results indicate that BiliCam could have a very similar utility to TcB as a screening tool, with the advantage of greater accessibility.

DISCUSSION

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 we can fully characterize BiliCam as an in-home jaundice screening tool.

Limitations

Our current data collection was 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. We 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 we 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 our 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. To this end, we are currently planning several national and international clinical studies for BiliCam.

Additionally, the quality of images has a major bearing on the output of BiliCam. We 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 our dataset.

Sample 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 we segmented images for patches of skin on both the forehead and sternum, we ultimately focused solely on the sternum. There are several possible explanations for why the sternum yielded better results. We 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 [31].

Given the quality of our initial results from using just the still images, we focused on those and have not investigated the videos in our 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 we discuss further in our future work.

Foreseeable 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 [34]. Because of the seriousness of these complications, the social return on investment for this technology in resource poor areas is considerably elevated.

Depending on the context, we expect parents, clinicians, or community health workers to use BiliCam. For newborns delivered in hospitals, families could receive a card, with the color calibration target printed on one side and instructions to download the system on the other side, before discharge from the hospital. Parents would then use BiliCam to screen their newborns in the comfort of their own homes, potentially saving inaccuracies, anxiety, unnecessary clinic visits, and blood draws when compared to the current method of visual assessment. Moreover, BiliCam offers them a method of continuous screening and earlier identification of newborns in need of treatment. At the initial outpatient visits for newborns, clinicians could use BiliCam to assess the severity of jaundice and determine if testing with TSB is warranted or unnecessary. In areas where visiting nurses are common, they could bring a smartphone and color calibration cards as they visit families to more objectively assess neonatal jaundice than their current, visual method. Many low-resource areas employ community health workers (CHWs), who travel from one village to another to provide limited health care. Having them carry and use this screening tool could help reduce the frequency of kernicterus and its consequences where it is most prevalent.

Future Work

Future work will focus primarily on further data collection to reduce the system's limitations. In addition to acquiring more data points and increasing the diversity of our samples, we would also like to introduce the following improvements for future clinical studies.

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 our 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 we 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 we 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 our results are using still photographs, one improvement could be to take a series of clearly punctuated still images instead of a video. We 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. We believe a server-connected version is otherwise preferred.

CONCLUSION

In conclusion, we presented BiliCam, a non-invasive smartphone-based system to monitor newborn jaundice using the built-in camera. Our initial evaluations imply that although BiliCam cannot yet replace it, BiliCam could become an effective screening tool comparable to a TcB. Unlike current screening techniques, it also offers distinct cost and accessibility advantages that make it appropriate for screening newborns in the comforts of their own homes.

ACKNOWLEDGEMENTS

We heartily thank our data collectors: Barbara Baker, Susan Sargent, and Tatiana Gellein. This research was funded in part by Coulter Foundation and an NSF Graduate Research Fellowship.

REFERENCES

1. Agu, E., Pedersen, P., Strong, D., et al. The smartphone as a medical device: Assessing enablers, benefits and challenges. *IEEE International Workshop of Internet-of-Things Networking and Control (IoT-NC)*, (2013), 76–80.
2. Akman, I., Arika, Ç., Bilgen, H., KalaCa, S., and Özek, E. Transcutaneous Measurement of Bilirubin by Icterometer During Phototherapy on a Bilibed. *Turkish Journal of Medical Sciences 32*, June 2000 (2002), 165–168.
3. Baker, C., Fontela, G., Jones, P., Lynch, B., Sypher, S., and Patil, C. BME 272 NCIIA Project Proposal Neonatal Jaundice. 2012.
4. Bhutani, V., Johnson, L., and Sivieri, E. Predictive ability of a predischarge hour-specific serum bilirubin for subsequent significant hyperbilirubinemia in healthy term and near-term newborns. *Pediatrics 103*, 1 (1999).
5. Bhutani, V.K., Stark, A.R., Lazzeroni, L.C., et al. Predischarge Screening for Severe Neonatal Hyperbilirubinemia Identifies Infants Who Need Phototherapy. *The Journal of pediatrics 162*, 3 (2013), 477–482.e1.
6. Bland, J.M. and Altman, D.G. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *The Lancet*, (1986), 307–310.
7. Boulanger, C., Boulanger, A., de Greef, L., et al. Stroke rehabilitation with a sensing surface. *ACM CHI Conference on Human Factors in Computing Systems*, (2013).
8. Bourouis, a., Feham, M., Hossain, M. a., and Zhang, L. An intelligent mobile based decision support system for retinal disease diagnosis. *Decision Support Systems 59*, (2014), 341–350.
9. Breiman, L.E.O. Random Forests. *Machine Learning 45*, 1 (2001), 5–32.
10. Brunori, P., Masi, P., Faggiani, L., et al. Evaluation of bilirubin concentration in hemolysed samples, is it really impossible? The altitude-curve cartography approach to interfered assays. *Clinica chimica acta; international journal of clinical chemistry 412*, 9-10 (2011), 774–7.
11. Chang, Y.-J., Chen, S.-F., and Huang, J.-D. A Kinect-based system for physical rehabilitation: a pilot study for young adults with motor disabilities. *Research in developmental disabilities 32*, 6 (2011), 2566–70.
12. Chen, N., Wang, K.-C., and Chu, H.-H. Listen-to-Nose : A low-cost system to record nasal symp- toms in daily life. *14th ACM International Conference on Ubiquitous Computing (UbiComp)*, (2012).
13. Consolvo, S., Klasnja, P., Mcdonald, D.W., et al. Flowers or a Robot Army ? Encouraging Awareness & Activity with Personal, Mobile Displays. *10th International Conference on Ubiquitous Computing (UbiComp)*, (2008).
14. Dell, N. and Borriello, G. Mobile Tools for Point-of-Care Diagnostics in the Developing World Categories and Subject Descriptors. *ACM Annual Symposium on Computing for Development (DEV)*, (2013).
15. Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least Angle Regression. *The Annals of Statistics 32*, 2 (2004), 407–499.
16. Franko, O.I., Bray, C., and Newton, P.O. Validation of a scoliometer smartphone app to assess scoliosis. *Journal of pediatric orthopedics 32*, 8 (2012), e72–5.
17. Gupta, M.R., Garcia, E.K., and Chin, E. Adaptive local linear regression with application to printer color management. *IEEE Transactions on Image Processing*, (2008), 936–945.
18. Larson, E., Goel, M., Boriello, G., Heltshe, S., Rosenfeld, M., and Patel, S. SpiroSmart: Using a Microphone to Measure Lung Function on a Mobile Phone. *ACM UbiComp*, (2012).
19. Larson, E.C., Lee, T., Liu, S., Rosenfeld, M., and Patel, S.N. Accurate and privacy preserving cough sensing using a low-cost microphone. *13th International Conference on Ubiquitous Computing (UbiComp)*, ACM Press (2011), 375.
20. Leartveravat, S. Transcutaneous bilirubin measurement in full term neonate by digital camera. *Medical Journal of Srisaket Surin Buriram Hospitals 24*, 1 (2009).
21. Lee, J., Reyes, B. a, McManus, D.D., Mathias, O., and Chon, K.H. Atrial fibrillation detection using a smart phone. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2012*, (2012), 1177–80.
22. Lee, J., Reyes, B. a, McManus, D.D., Mathias, O., and Chon, K.H. Atrial fibrillation detection using an iPhone 4S. *IEEE Transactions on Bio-Medical Engineering 60*, 1 (2013), 203–6.
23. Longhurst, C., Turner, S., and Burgos, A.E. Development of a Web-based decision support tool to increase use of neonatal hyperbilirubinemia guidelines. *Joint Commission journal on quality and patient safety / Joint Commission Resources 35*, 5 (2009), 256–62.
24. Longhurst, C., Turner, S., and Burgos, A.E. BiliTool™. 2014. http://bilitool.org/.
25. Maisels, M.J., Bhutani, V.K., Bogen, D., Newman, T.B., Stark, A.R., and Watchko, J.F. Hyperbilirubinemia in the newborn infant > or =35 weeks' gestation: an

update with clarifications. *Pediatrics 124*, 4 (2009), 1193–8.

26. National Collaborating Centre for Women's and Children's Health and National Institute for Health and Clinical Excellence. *Neonatal Jaundice Clinical Guideline*. London, 2010.

27. Pamplona, V.F., Mohan, A., Oliveira, M.M., and Raskar, R. NETRA : Interactive Display for Estimating Refractive Errors and Focal Range. *ACM Transactions on Graphics (SIGGRAPH)*, (2010).

28. Patel, P. ClikJaundice: Using mobile technology to detect yellow in newborns. 2013. http://thealternative.in/social-business/clickjaundice-using-the-phone-to-prevent-jaundice-in-newborns/.

29. Pedregosa, F., Weiss, R., and Brucher, M. Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research (JMLR) 12*, (2011), 2825–2830.

30. Poh, M.-Z., Kim, K., Goessling, A.D., Swenson, N.C., and Picard, R.W. Heartphones: Sensor Earphones and Mobile Application for Non-obtrusive Health Monitoring. *2009 International Symposium on Wearable Computers*, (2009), 153–154.

31. Poland, R.L., Hartenberger, C., McHenry, H., and Hsi, A. Comparison of skin sites for estimating serum total bilirubin in in-patients and out-patients: chest is superior to brow. *Journal of perinatology : official journal of the California Perinatal Association 24*, 9 (2004), 541–3.

32. Ramer, U. An Iterative Procedure for the Polygonal Approximation of Plane Curves. *Computer Graphics and Image Processing 1*, 3 (1972), 244–256.

33. Shen, L., Hagen, J. a, and Papautsky, I. Point-of-care colorimetric detection with a smartphone. *Lab on a chip 12*, 21 (2012), 4240–3.

34. Slusher, T.M., Zipursky, A., and Bhutani, V.K. A global need for affordable neonatal jaundice technologies. *Seminars in perinatology 35*, 3 (2011), 185–91.

35. Smola, A.J. and Schölkopf, B. A tutorial on support vector regression. *Statistics and computing 14*, 3 (2004), 199–222.

36. Viggiano, J.A.S. Comparison of the accuracy of different white balancing options as quantified by their color constancy. *Sensors and Camera Systems for Scientific, Industrial, and Digital Photography ApplicationsV: Proceedings of the SPIE 5301*, (2004).

37. Wadhawan, T., Situ, N., Rui, H., Lancaster, K., Yuan, X., and Zouridakis, G. Implementation of the 7-point checklist for melanoma detection on smart handheld devices. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. (2011), 3180–3.

38. Wang, L., Pedersen, P.C., Strong, D., Tulu, B., and Agu, E. Wound image analysis system for diabetics. *International Society for Optics and Photonics (SPIE)*, (2013), 866924.

39. Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*, 2 (2005), 301–320.

40. Management of Hyperbilirubinemia in the Newborn Infant 35 or More Weeks of Gestation. *Pediatrics 114*, 1 (2004), 297–316.