Calibration of Contactless Pulse Oximetry

Wim Verkruijsse, PhD,* Marek Bartula, MSc,† Erik Bresch, PhD,‡ Mukul Rocque, MSc,† Mohammed Meftah, BSc,† and Ihor Kirenko, PhD†

BACKGROUND: Contactless, camera-based photoplethysmography (PPG) interrogates shallower skin layers than conventional contact probes, either transmissive or reflective. This raises questions on the calibratability of camera-based pulse oximetry.

METHODS: We made video recordings of the foreheads of 41 healthy adults at 660 and 840 nm, and remote PPG signals were extracted. Subjects were in normoxic, hypoxic, and low temperature conditions. Ratio-of-ratios were compared to reference SpO_2 from 4 contact probes.

RESULTS: A calibration curve based on artifact-free data was determined for a population of 26 individuals. For an SpO_2 range of approximately 83% to 100% and discarding short-term errors, a root mean square error of 1.15% was found with an upper 99% one-sided confidence limit of 1.65%. Under normoxic conditions, a decrease in ambient temperature from 23 to 7°C resulted in a calibration error of 0.1% ($\pm 1.3\%$, 99% confidence interval) based on measurements for 3 subjects. PPG signal strengths varied strongly among individuals from about 0.9×10^{-3} to 4.6×10^{-3} for the infrared wavelength.

CONCLUSIONS: For healthy adults, the results present strong evidence that camera-based contactless pulse oximetry is fundamentally feasible because long-term (eg, 10 minutes) error stemming from variation among individuals expressed as A^*_{rms} is significantly lower ($<1.65\%$) than that required by the International Organization for Standardization standard ($<4\%$) with the notion that short-term errors should be added. A first illustration of such errors has been provided with $A^*_{rms} = 2.54\%$ for 40 individuals, including 6 with dark skin. Low signal strength and subject motion present critical challenges that will have to be addressed to make camera-based pulse oximetry practically feasible. (Anesth Analg 2017;124:136–45)

Since their invention several decades ago,^{1,2} pulse oximeters have successfully used the principle of photoplethysmography (PPG). In PPG, the intensity of light that travels through tissue is modulated by the absorption of pulsatile blood volume.³ In pulse oximetry, the relative PPG amplitudes at 2 or more wavelengths provide a noninvasive estimate, termed SpO_2 , of the arterial oxygen saturation So_2 . Many technical improvements over the past decades have led to SpO_2 as one of the common vital signs to be monitored clinically, but it has always remained a contact measurement.

However, development of inexpensive and sensitive digital imaging sensors has enabled several investigators^{4–8} to change the conventional geometry and measure PPG in a fully contactless manner, with both light source and detector away from the skin. The potential advantages of contactless PPG include avoiding skin damage in fragile patients and a freedom to select a more physiologically central location with a possibly faster response.^{9,10}

In the past decade, a rapidly increasing number of articles¹¹ has been published on contactless PPG addressing pulse or respiration rate, while only a few^{4,5,12–16} explore SpO_2 . Humphreys et al⁵ sought to emulate the conventional source-detector geometry by illuminating the skin at one location and measuring PPG (with a camera) right next to it. Wieringa et al⁴ used wide-field illumination but focused on imaging vasculature. However, none of these has convincingly demonstrated that camera-based SpO_2 is fundamentally calibratable. In this article, we address the question “Can a single calibration curve provide SpO_2 estimates for a population of individuals with acceptable accuracy?”

This is not a trivial question because there is a fundamental difference between the conventional contact source-detector (Figure 1A), and the contactless, wide-field illumination-detection geometries (Figure 1B). Although the former geometry collects light that has travelled through relatively deep vasculature (both in transmissive and reflective, or “surface” mode),¹⁷ the latter predominantly collects light that has travelled through much shallower tissue depths^{15,18} over much smaller distances.

Currently, it is not known whether the PPG signal measured in the camera geometry stems mostly from the deeper arterioles or also (partly) from the shallow capillaries. In fact, there is some controversy on whether the PPG signal stems directly from blood volume changes at all.^{19–21} We can only hypothesize on PPG origins and the consequences for SpO_2 calibratability. Due to their shallow location, superficial capillaries may contribute to PPG amplitude or even be the predominant source (Figure 1B), even if they are only slightly pulsatile compared with arterioles. If their oxygen saturation is not representative of arterial blood, calibratability becomes questionable. Another potential risk is shunt light, light that has travelled through the bloodless epidermis only, or specularly reflected light. The shallow skin layers involved in camera-based PPG are essentially

From the *Standardization Research, †Patient Care and Measurements, and ‡Data Science, Philips Research, Eindhoven, The Netherlands.

Acceptance for publication March 24, 2016.

Funding: None.

The authors declare no conflicts of interest.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (www.anesthesia-analgesia.org).

Reprints will not be available from the authors.

Address correspondence to Wim Verkruijsse, PhD, Standardization Research, Philips Research, HTC 36 p 0.88, Eindhoven, The Netherlands. Address e-mail to wim.verkruijsse@philips.com.

Copyright © 2016 International Anesthesia Research Society. This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially. DOI: 10.1213/ANE.0000000000001381

Figure 1. In conventional contact probes (A), albeit transmissive or reflective, the skin interrogation depth is significantly deeper than in a wide-field illumination camera geometry (B) which has consequences for the signal strength and potentially also for calibration of SpO_2 . Proportions not to scale.

Figure 2. Normalized raw PPG signals at a normoxic (A) and hypoxic (B) segments. Breath-holds are shown along with SpO_2 traces from finger probe and camera in (C). Time segments in (A) and (B) are indicated in (C) by “n” and “h.” The camera SpO_2 trace is delayed by 21 seconds for a good visual match. Tentative calibration constants were used.

“uncharted territory” in terms of pulse oximetry. To address this fundamental question of calibratability, we have taken an experimental approach. However, we did not address technical challenges and used only artifact-free data.

We have found that some individuals can induce significant changes in SpO_2 by holding their breath. Trending of contactless SpO_2 can thus be demonstrated without the aid of hypoxic environments (Figure 2). However, to calibrate and study accuracy on a broader population, stable desaturation levels are needed to exclude discrepancies related to the temporal filtering in contact references and synchronization between reference and

camera traces. Guazzi et al¹⁶ were the first to report contactless SpO_2 in a more controlled manner in a hypoxic chamber but their analysis aimed at trending and resulted in different calibration constants for each of the 5 individuals studied.

In the present study, we use a dedicated hypoxic chamber with the goal of investigating whether a single calibration curve can describe a population of individuals with acceptable accuracy. We realize that calibratability may depend on many variables such as physiological status, age, and skin condition. We only included healthy adults as subjects under normoxic and hypoxic conditions. Additionally, we explored the impact of simulated centralization on the calibratability of contactless pulse oximetry. To this goal, we exposed subjects to low temperatures in a climate room study under normoxic conditions.

METHODS

IRB/Consent

All experiments were approved by the Philips IRB (Internal Committee on Biomedical Experiments) before experimentation. Written informed consent was obtained from all subjects. Three different experimental protocols (studies N, H, and C) were submitted to and approved by the Philips IRB. The methods used for the 3 experiments are described in separate sections below. IRB contact: PHM Keizer, Senior Ethical & Biomedical Officer, High Tech Campus 34, 5656 AE Eindhoven, The Netherlands.

Clinical Trial Registration

The study was not registered before patient enrollment because it was not a clinical trial; it was a basic research study on healthy volunteers. It did not include treatment or control groups.

The Basic Set-up (Study N)

During data collection, the subjects were seated, and head motion was minimized using a head support mounted on the chair. Two tripod-mounted monochrome Stingray F046-B FireWire cameras (Allied Vision Technologies, Stadtroda, Germany) with Computar 50 mm 1:1.3 lenses (CBC, Commack, NY) were each equipped with spectral bandpass filters with red and infrared (IR) center wavelengths of 675 and 842 nm, respectively (Semrock FF02-675/67-24-D and FF01-842/56-24-D; IDEX Corp., Lake Forrest, IL). The cameras recorded movies of the subject's face from a distance of about 1.6 meters at 15 frames per second.

Illumination was provided by 2 armatures (Falcon Eyes, Hong Kong, China), each equipped with 9 incandescent lamps (Philips 40 W) at a distance of about 1 m from the subjects. Current limited DC power supplies set to 150 V, 1.25 A (SM3004-D, Delta Electronica, Zierikzee, The Netherlands) powered the lamps.

For SpO_2 reference, rather than blood gas analysis, we used 4 conventional SpO_2 probes coupled to Philips MP2 patient monitors: a Philips finger sensor, a Philips ear sensor (M1191B and M1194A, Philips Medizin-Systeme, Böblingen, Germany), a Masimo finger sensor (LNCS DC-I, Masimo Corporation, Irvine), and a Nellcor finger sensor (DS-100A Medtronic, Dublin, Ireland). A sample-wise (1 Hz) median of all 4 probes was defined as the reference signal $\text{SpO}_{2,\text{ref}}(t)$. Twenty-five individuals participated in study N (Supplemental Digital Content, Supplemental Table 1, <http://links.lww.com/AA/B428>).

The Hypoxic Tent (Study H)

For the second study, the setup described above was used in a normobaric hypoxic tent (At Home Cubicle, Hypoxico Inc, New York, NY). Oxygen concentration and temperature inside the tent were continuously monitored (Model AD300, Teledyne Analytical Instruments, City of Industry, CA and Model 54II B, Fluke Corp, Everett, WA, respectively). After about 4 hours of pumping with the hypoxic generator (Everest Summit 2 Hypoxic Generator, Hypoxico Inc), the oxygen concentration stabilized at approximately 15% at which point the subject entered the tent, carefully and quickly to minimize air exchange. An electric fan (Cool Air System 40, Philips) homogenized the air inside the tent while the outlet of the oxygen pump was relatively close (eg, 20 cm) to the mouth of the subject and the probe of the oxygen meter at about 1 m away, probing homogenized air. During data collection, oxygen concentration was measured to be $15.0\% \pm 0.5\%$ (average and standard deviation over all recordings) and air temperature typically increased to approximately 25°C . While results varied strongly among subjects, an SpO_2 reduction of about 7% on average was induced by the hypoxic conditions: median SpO_2 values were 98.2% and 90.6% for studies N and H, respectively (Supplemental Digital Content, Supplemental Table 1, <http://links.lww.com/AA/B428>). Twenty-one individuals participated in study H (Supplemental Digital Content, Supplemental Table 2, <http://links.lww.com/AA/B428>), three were also in study N.

The Climate Chamber (Study C)

In the third study, the subject was seated comfortably in a chair in a climate-controlled room (Environmental Stress

Chamber GMAPX-15CW, Hielkema Testequipment, Uden, NL), with an initial temperature of approximately 22°C . The chamber temperature was then gradually reduced to about 5°C . This process took approximately 30 minutes. The room temperature was continuously monitored using the temperature probe of a patient monitor (MP5, Philips Medizin-Systeme). Video recordings of the subject's head were made before the cooling phase at about 22°C , in the middle of the cooling phase at about 11°C , and at the end at about 5°C . The climate chamber's cooling fan was turned off during the recordings to avoid any air turbulence moving the cameras or illumination. Due to limited space in the climate room, only 1 illumination armature was used instead of 2 in studies N and H. Other than this difference, the temperature was the only variable that changed.

To obtain feedback on cold-induced skin physiological changes, skin reflectance spectra (380–740 nm) were measured on the forehead with a spectrophotometer (CM-2600d, Konica-Minolta, Osaka, Japan). Immediately after each video recording at each temperature regime, 5 measurements were made and used as input for a median spectrum. Five individuals participated in study C (Supplemental Digital Content, Supplemental Table 2, <http://links.lww.com/AA/B428>). Three of these individuals also participated in studies N and/or H.

Signal Processing for All Experiments

The camera and reference probe signals were acquired and processed by LabVIEW (National Instruments, Austin, TX). The raw PPG signals $R(t)$ and $IR(t)$ (Appendix) are the spatially averaged pixel intensity in a manually selected region of interest on the forehead. Next, using a window size of 10 seconds and step size of 1 second, $R(t)$ and $IR(t)$ were normalized (by the respective temporal mean values over the window) and temporally filtered (second-order Butterworth bandpass with the cutoff frequencies 0.5 and 1.5 Hz), giving signals $R_n(t)$ and $IR_n(t)$. For each channel, the difference between median values of peaks and valleys was defined as the pulsatility amplitude, now at a temporal resolution of the window step size. The ratio-of-ratio signal $RR(t)$ was finally computed as the sample-wise quotient of red and IR pulsatility amplitudes, which is subsequently linearly transformed into an SpO_2 trace with calibration constants C_1 and C_2 :

$$\text{SpO}_{2,\text{cam}}(t) = C_1 - C_2 * RR(t). \quad (1)$$

Camera SpO_2 traces were typically about 20 seconds faster than the $\text{SpO}_{2,\text{ref}}$, due to a limited signal processing of the camera signals compared with the reference SpO_2 probes and likely also a physiological delay at the periphery with respect to the forehead. Most traces have insufficient features to individually determine a delay. To avoid fitting the camera data to the reference probe, we tried to determine an average delay for the whole data set. We identified 12 traces with distinct temporal features in the SpO_2 traces. For each of these traces, we determined the optimal delay τ for temporal alignment of the reference and camera SpO_2 traces, using Equation 2:

$$\tau = \underset{t'}{\operatorname{argmin}} | \text{SpO}_{2,\text{ref}}(t - \tau') - \text{SpO}_{2,\text{cam}}(t) | \quad (2)$$

Figure 3. Example SpO_2 traces for study H. Values for (tentative) calibration constants C_1 and C_2 were the same as those used for the breath hold experiment (Figure 2). For a good temporal match with the SpO_2 ref, delays τ were determined as 20, 22, 24, and 24 seconds in (A–D), respectively.

Four of the 12 traces are shown in Figure 3, A–D. For the 12 traces, an average delay of 21.1 seconds was found. An SD of 2.1 seconds partly represents physiological variation but also expresses accuracy of determination of τ .

Next, we computed the temporal median RR of the signal $\text{RR}(t)$. This procedure was justified since we considered individuals in steady-state conditions and we were primarily interested in long-term calibratability of camera-based SpO_2 measurements. Further advantages are that each individual entered the calibration process with the same weight and that recording length did not introduce a bias. It also further minimized synchronization issues between reference and camera-based traces and among the reference probes themselves. Similarly, for each recording, 1 median $\text{SpO}_{2,\text{ref}}$ value was computed for each $\text{SpO}_{2,\text{ref}}(t)$ trace. With this approach, each recording in studies N and H provided 1 data point (median RR, median $\text{SpO}_{2,\text{ref}}$) to populate a calibration data set.

Despite head fixation, some recordings were frequently disrupted due to motion. In other recordings, the signal

amplitude of the red PPG signal was so small that noise from illumination or camera disrupted proper measurement of the pulsatility. Because we sought a fundamental assessment of calibratability, we excluded such recordings from the data set in an objective manner. To this goal, a signal quality $Q(t)$, essentially a signal-to-noise ratio, was derived from the raw red PPG signal to allow the deselection of data of low signal quality (see Results, Figure 4B). Hereby, $R_n(t)$, the weakest signal, was processed in 20-s sliding windows, in steps of 1 second. Within each window, the Fourier transform of $R(t)$ was computed. The magnitude of the Fourier spectrum at the heart rate was used as a measure of signal strength, whereas the average magnitude at frequencies of heart rate ± 0.3 Hz was used as a measure for noise strength. The Q -value for the window was finally computed as $\log_{10}(\text{signal amplitude}/\text{noise amplitude})$. Example signals of strong, weak, and motion-distorted signals are shown in Figure 4, A–C, respectively.

Analogous to the temporal medians for $\text{SpO}_{2,\text{ref}}$ and RR, we also defined a single Q value for each recording, as the temporal median of $Q(t)$. As shown in the Results section, a threshold value Q_{thr} was determined as an objective measure to only use reliable data in the calibration procedure.

Statistical Approach and Analysis

Before our measurements, we faced many unknowns such as variations in PPG signal strength among individuals, subject motion, and the extent to which hypoxia could be induced in the normobaric hypoxic tent. Rather than conducting a pilot study to investigate each of these unknowns, we recruited relatively large numbers of subjects for studies N and H (25 each), in the anticipation that this would provide a sufficiently large and clean data set to determine a stable linear regression between RR and $\text{SpO}_{2,\text{ref}}$. Thus, we enrolled many subjects in the study and determined a posteriori if we succeeded in establishing a stable linear regression. This process is described in the Results section.

All measurements for studies N and H were performed consecutively without intermediate analysis to adjust the study design. Inclusion and exclusion criteria for studies N, H, and C are in Supplemental Digital Content, Supplemental Table 3 (<http://links.lww.com/AA/B428>). For study N, all 25 recruited individuals were included while for study H, 21 individuals were included after checking for inclusion and exclusion criteria. For study C, we were somewhat limited in inclusion because the climate room had limited availability. The purpose of study C was exploratory.

For studies N and H, an objective measure was used to discard recordings with unreliable RR. As shown in the Results section, indeed the 17 cleanest data points resulted in a regression line that was essentially the same as the regression for the 31 cleanest data points by which we mean that any error due to variation in regression was irrelevant (much smaller) compared with inaccuracies caused by variations among subjects. The determination of C_1 and C_2 was eventually based on the 31 cleanest data points. Next, with these calibration constants we proceeded to answer the overall study goal (Can a single calibration curve provide SpO_2 estimates for a population of individuals with acceptable accuracy?). For this, we adopted A_{rms}

Figure 4. Normalized PPG signal segments with Q values of 2.2, 1.0, and 0.6 in (A), (B), and (C), respectively. The PPG signals may be described as “strong,” “weak,” and “motion corrupted,” respectively.

a metric used for pulse oximeters to describe accuracy (ISO 80601-2-61, 2011, section 201.12.1.101.2.2). An * superscript is used to indicate at least 1 major difference with the International Organization for Standardization (ISO) standard, which is that in our formulation (Equation 3) we used median values of the traces, thus discarding short-term errors.

$$A_{rms}^* = \sqrt{\frac{\sum (SpO_{2,ref} - SpO_{2,cam})^2}{n}} \quad (3)$$

Because, in our calibration approach, this expression was mathematically similar to that of a standard deviation, we used the χ^2 approach to calculate an error estimate for A_{rms}^* .

A mean difference or “bias” was according to Equation 4, also following the ISO standard.

$$MD = \frac{\sum (SpO_{2,ref} - SpO_{2,cam})}{n} \quad (4)$$

In study C, the protocol was repeated once for each of the 5 individuals. To assess statistical significance among parameters (eg, DC reflectance, PPG signal strength or $SpO_{2,cam} - SpO_{2,ref}$), caused by exposure to different temperatures, we averaged the values from the 2 sessions and then treated these averages as 5 independent samples (from 5 individuals).

Throughout this study, we reported 99% confidence intervals (CIs) because no prior data were used to choose our sample sizes. When Student t tests were used, we reported the P values.

RESULTS

Filtering the Data Set and Determination of Q_{thr} and Calibration Constants C_1 , C_2

A total of 46 data points from 39 individuals for studies N and H are shown in Figure 5A and listed in Supplemental Table 1 (Supplemental Digital Content, <http://links.lww.com/AA/B428>). As expected, similar to calibration curves for conventional pulse oximetry, a negative correlation between median RR and median $SpO_{2,ref}$ is seen. Some outlier data points are from signals similar to those in Figure 4B or C, that are obviously corrupted. As expected, similar to calibration curves for conventional pulse oximetry, a negative correlation between median RR and median $SpO_{2,ref}$ is seen. Some outlier data points are from signals

Figure 5. All candidate data points for calibration are shown in (A) with Q values indicated and regression lines for different values for Q_{thr} . Using data with $Q > 0.5$ shows a regression line (dashed) that is biased by outlier points. Too strict filtering with $Q_{thr} = 2.3$ gives a regression line (dashed) determined by too few data points and low value for R^2 . Regression has high values for R^2 and, more importantly, is stable for $1.4 < Q_{thr} < 1.9$, (A and B). Fourteen regression lines (solid) are shown for this range of Q_{thr} .

similar to those in Figure 4B or C that are obviously corrupted. As mentioned, we wished to discard such data in an objective manner. We defined an objective quality metric Q and indeed, the signals in Figure 4B and C, had low Q values. However, the question remains which threshold value Q_{thr} to use for this metric to discard data for which the signal corruption is less obvious. In the following paragraph, we describe the process of determining Q_{thr} in an objective manner.

In Figure 5B, we used Q_{thr} as a parameter to discard data with Q is less than Q_{thr} and plot the coefficient of determination (R^2) for regression on the remaining data. Three domains can be identified: insufficient filtering, appropriate filtering, and too aggressive filtering. At $Q_{thr} < 1.4$, recordings with corrupted data had a large impact on the regression; only moderate R^2 values were found. For $1.4 < Q_{thr} < 1.9$, R^2 was high but more importantly, regression lines were very similar to each other (Figure 5A). For $Q_{thr} > 1.9$, both the R^2 values and regression slopes were unstable. Here, the filtering was too aggressive and inter-individual variations combined with poor statistics (too few data points) dominated the regression slope. We chose $Q_{thr} = 1.4$ as the filtering value to obtain a clean set of data for the calibration although any value between 1.4 and 1.9 gave essentially equal calibration results. In other words, the 17 cleanest recordings gave the same regression as the 31 cleanest recordings and those in between (14 regression lines shown in Figure 5A), indicating that the data set was sufficiently large to determine a population calibration.

The regression line for $Q_{thr} = 1.4$ defined the calibration curve with constants $C_1 = 118.0$ and $C_2 = 45.9$ (Figure 6). With calibration defined, the next step was to analyze the A^*_{rms} as this determined an estimate of the system accuracy. After linear regression on 31 data points, the point estimate of the A^*_{rms} was 1.15%. We applied a 1-sided, 99% CI based on the χ^2 distribution with 29 degrees of freedom for the root mean

square error after a linear regression of our data. This gave an upper limit of 1.65% for A^*_{rms} .

Perturbation of Physiological Conditions and Validation of Calibration Under Normoxic Conditions

One source of feedback on the effect of cooling was obtained with reflectance spectra on the forehead. Typically, the reflectance spectra at low temperatures were slightly higher than that at room temperature for wavelengths between 450 and 600 nm that was likely due to a lower blood volume caused by vasoconstriction. Figure 7A shows reflectance spectra for room and cold temperatures for one individual, as well as the average difference for 5 individuals. The difference just reached statistical significance (at 99% CI) for a few yellow and green wavelengths. An increased reflectance in this wavelength region only was consistent with a reduced blood volume at skin depths of about 0.1 to 0.3 mm (Figure 7B).²² These skin layers were dominant in camera-based PPG measurements (Figure 1B). While barely significant, we considered this a first sign of successful perturbation relevant for investigating calibration of camera-based pulse oximetry.

A stronger manifestation of physiological perturbation was the dependency of PPG signal strength on temperature, for both contact and contactless PPG (Figure 7C). Pulsatile strengths varied considerably at any temperature, which was due to a natural variation among individuals (Supplemental Digital Content, Supplemental Table 3, <http://links.lww.com/AA/B428>) and the fact that temperature control was not perfect. A strong wind from the cooling ventilators caused us to only record when they were off. Thus, actual temperatures increased quite rapidly during recordings and indicated temperatures per data point were only averages. Nevertheless, the impact of ambient temperature on pulsatility in the finger was significant. A paired, 1-sided t test analysis of the changes of pulsatile strengths with temperature shows that pulsatility at room temperature was larger than that at medium temperatures ($P = 0.0096$) and similarly for medium and cold temperatures ($P = 0.006$) for the finger probe. For the forehead, this was similar for room and medium temperature ($P = 0.002$). For medium and cold temperatures, the difference did not reach a significant level ($P = 0.034$); here, only 3 data were available with Q is greater than Q_{thr} . Overall, the amplitude reduction at the finger was stronger than on the forehead by approximate factors of 30 and 4, respectively, for the temperature range of 23 to 7°C. More interestingly, however, we observed that the regression lines for red and IR were more or less parallel, which illustrates that the RR, critical for Spo_2 calibratability, did not appear to be negatively affected. Please note that the wavelengths used in contact probes were different than those used for the camera, which explains the different red/IR ratios. To investigate calibration robustness with temperature, we did not further analyze the red and IR regression lines because this would have implied that we assumed the Spo_2 did not change with temperature. In fact, on average, $Spo_{2,ref}$ increased slightly by 0.21% ($\pm 0.08\%$, 99% CI) per degree Celsius cooling. Instead, we computed $Spo_{2,cam}$

Figure 6. The calibration curve (regression line) based on filtered data ($Q > 1.4$) along with its corresponding 99% confidence interval. Discarded data ($Q < 1.4$) not used in calibration are also shown.

Figure 7. Reflectance spectra for one individual (A, left axis) and the average difference for five individuals (B, right axis) shown along with the 99% confidence interval. Typically, reflectance for green and yellow light increases slightly at low temperature although the difference reaches significance at a few wavelengths only. Such a difference is consistent with reduced blood volume in superficial skin (B). B, The relative penetration depths of the various wavelengths indicated. The PPG signal strength strongly reduces with temperature (C) for both finger contact probe and camera.

Figure 8. Calibration data from studies N and H, shown along with data from study C (normoxic conditions), all computed with one set of calibration constants. The labels “Low,” “Medium,” and “Room” refer to temperatures of approximately 7, 14, and 22°C, respectively.

values from the RR and the earlier established calibration constants. The results are shown in Figure 8 combined with the main results of studies N and H to allow an easy visual comparison. The data used for calibration are shown along with the computed upper limit for A^*_{rms} (99% CI), indicated as dashed lines. The ISO requirement of $A_{rms} < 4\%$ is also indicated for reference.

It is encouraging that with a different experimental setting, the $Spo_{2,cam}$ values from study C were estimated quite well with the calibration derived from studies N and H. Even though study C explored the impact of ambient temperature on calibrability rather than validation of the calibration, the data for 2 individuals (40 and 41) that were not in the calibration studies at room temperature can be considered a first validation of the calibration at normoxic conditions. While the mean difference values for room, medium, and cold temperature were small (-0.7% , -0.5% , and -0.7%), the 99% CI intervals were relatively large ($\pm 0.7\%$, $\pm 2.1\%$, and $\pm 2.4\%$, respectively), expressing the poor statistics in study C. An indication of whether calibration is robust for cold perturbation is to analyze the pairwise changes in Spo_2 discrepancies ($Spo_{2,cam} - Spo_{2,ref}$) per individual for a change in temperature because such an analysis discards any potential systematic errors resulting from the different experimental setting. With Q is greater than Q_{thr} , we had 5 $Spo_{2,cam} - Spo_{2,ref}$ values at room and medium temperature but only 3 at cold temperature, which gave 5 pairs for room-medium and 3 pairs for medium-cold. An additional comparison was made by comparing the values for room and cold temperature. The average changes in discrepancies for these 3 temperature steps were small: 0.3% , 0.5% , and -0.1% for room-medium, medium-cold, and room-cold, respectively. However, the 99% CI were relatively large: $\pm 2.0\%$, $\pm 0.7\%$, and $\pm 1.3\%$, respectively.

Figure 9. Differences between $\text{SpO}_{2,\text{cam}}$ and $\text{SpO}_{2,\text{ref}}$ for 10 second segments with $Q > 1.4$ from all recordings of studies N, H, and C. The calibration constants are validated for a population of 26 individuals that were in the calibration (“+” symbols), and 16 new individuals (“—” symbols) that were not in the calibration set. Six individuals with fotoype IV or higher are now represented in total and are indicated by “*” in the legend and emphasized by larger symbols in the chart. The A_{rms}^* is 2.54% where the asterisks serve to indicate that the individuals are not equally weighed for the calculation.

DISCUSSION

The presented data provide strong evidence that a single calibration curve for a population of healthy adult individuals can be used to estimate SpO_2 contactlessly with an acceptable accuracy of $A_{\text{rms}}^* < 1.65\%$ (upper 99%, one-sided confidence limit). While noting that we discarded short-term errors, we consider this accuracy acceptable in view of typical A_{rms} values of 2% or 3% for commercial pulse oximeters and the maximum of 4% as specified by the ISO standard (80601-2-61, 2011). In our view, this finding was not trivial because the contactless nature of camera-based pulse oximetry implies interrogation of much shallower skin layers than conventional pulse oximetry.

The calibration was performed on the basis of data for 26 individuals from a data pool of 39. Using an objective measure, the data for 15 individuals were discarded because subject motion and/or low pulsatile strengths caused the signals to be unreliable for accurate assessment of pulsatile strength. Overall pulsatile strength varied significantly among individuals: the lowest and highest median IR pulsatile strengths in the data with $Q > 1.4$ were 0.9×10^{-3} and 4.6×10^{-3} , respectively. The discarded data tended to have lower average pulsatility than the nondiscarded data (IR pulsatility: 0.9×10^{-3} and 1.2×10^{-3} , respectively). Even though we found that 14 regression lines were stable for $1.4 < Q < 1.9$, we wanted to verify that the corresponding measured pulsatile strengths were clean, that is, the signal was due to true PPG and not due to motion or noise (camera or lamp). In our approach of assessing pulsatile strength, noise artificially increased pulsatile strength in both channels

(red and IR) to the same extent, pushing the RR value closer toward one. If this were the case, the discrepancy ($\text{SpO}_{2,\text{cam}} - \text{SpO}_{2,\text{ref}}$) would then tend to become negative for traces with relatively low pulsatile strength. Analysis of our data (Supplemental Digital Content, Supplemental Table 1, <http://links.lww.com/AA/B428>) showed that ($\text{SpO}_{2,\text{cam}} - \text{SpO}_{2,\text{ref}}$) in fact slightly trended in the opposite direction (not significant) illustrating that the measured pulsilities were caused by true PPG (skin color changes) and did not have a significant noise component. Apparently, the approach of discarding unreliable data based on Q provided satisfactory filtering of the data. Moreover, in study C, where PPG amplitudes were reduced significantly, we measured traces with $Q > 1.4$ at a pulsatile strength as low as 0.7×10^{-3} .

PPG pulse amplitude on the forehead decreased to a much smaller extent than at the finger. This was qualitatively consistent with observations of Bebout and Mannheimer⁹ who also compared pulsatile strengths on the forehead with the finger. It should be noted, however, that the contact forehead sensor interrogated deeper skin layers than the camera (Figure 1 and comments by Mannheimer¹⁷ on referring to such probes as “surface” versus “reflective” probes). Other differences in the present study were that the temperature reduction was slightly larger and acclimatization was shorter.

Unfortunately, the discarded group included 5 of 7 people with dark skin, leaving only 2 remaining subjects in the calibration set (subjects 7 and 29, Supplemental Digital Content, Supplemental Table 3, <http://links.lww.com/AA/B428>). Although not enough data were available for statistical proof, the 2 data points had relatively small

errors (0.5% and 1.5%), suggesting that there was not necessarily relevant bias due to melanin concentration. While melanin reduces the signal to noise ratio, we did not see evidence, experimental nor theoretical, of a potential calibration problem in contactless pulse oximetry caused by melanin.

In this study, we focused on fundamental calibratability based on median values of traces, discarding short-term errors. However, to obtain an impression of general accuracy, we considered all recordings of all individuals and considered segments of 10 seconds. In Figure 9, each point represents the average value for a 10-second segment. We filtered with $Q > 1.4$, per segment, resulting in 2500 segments (59% of all segments). Please note that the data in Figure 9 now include 6 individuals with skin phototype IV or higher (indicated by asterisks in the legend). For data that were in the calibration set, “+” symbols are used, while for the additional data we use “-.” Data for dark-skinned individuals are emphasized by using larger symbols. Some data points feature a relatively large error. It should be noted that the Q_{thr} of 1.4 was derived for traces and is likely not strict enough for the 10-second shorter segments. Also, the data that were not used in the calibration stem from traces with lower Q values; thus, it is not surprising that these data feature larger errors than the segments stemming from the traces with higher Q values. We include this figure to give the reader an impression of data that were discarded in the calibration. Also, it serves to remind that the calibration error A^*_{rms} is an underestimate because it discards short-term errors. While noting that some individuals contribute much more than others, the overall A^{**}_{rms} for Figure 9 is 2.54%. Short-term errors can be mitigated with various methods (eg, filtering and error concealment) and of course algorithms that address motion, unlike the approach used in this study. Many factors therefore will change (reduce) the eventual A_{rms} . The value of 2.54% is therefore merely illustrative which is why we did not compute a confidence interval. Nevertheless, we believe that the finding of an A^{**}_{rms} of 2.54%, together with the $A^*_{rms} < 1.65\%$ (upper 99% 1-sided confidence limit) for long-term calibratability, provides strong evidence for the fundamental feasibility of contactless pulse oximetry.

CONCLUSIONS

Calibration of contactless, camera-based pulse oximetry was performed by robust linear regression on 31 data points measured on a population of 26 healthy individuals under normoxic and hypoxic conditions (SpO_2 83% – 100%). Discarding short-term errors, an accuracy of $A^*_{rms} < 1.65\%$ (99% one-sided, upper confidence limit) was found which compared well with A_{rms} values for conventional pulse oximeters (typical values are 2% – 3%) and the maximum (4%) allowed by ISO 80601-2-61, 2011.

By exposing subjects to temperature changes from room temperature down to cold (about 5°C), discrepancies between $SpO_{2,ref}$ and $SpO_{2,cam}$ were 0.3%, 0.5%, and –0.1% for room-medium, medium-cold, and room-cold, respectively. With 99% CI intervals of $\pm 2.0\%$, $\pm 0.7\%$, and $\pm 1.3\%$, respectively, these accuracies do not necessarily violate the

ISO standard. Although further research on more than just 5 individuals is needed to narrow the intervals, these results are encouraging.

Challenges such as subject motion and low pulsatile strength have to be addressed to make this new measurement practical and successful. ■■

APPENDIX

$R(t), IR(t)$	Raw PPG signals: time series of the spatial averaged of intensity of camera pixels in the region of interest for red and infrared, respectively	At framerate 15 Hz
$Rn(t), IRn(t)$	Normalized and filtered PPG signals	At framerate 15 Hz
$RR(t)$	Quotient of red and IR pulsatility amplitudes	At 1Hz
RR	Temporal median value of $RR(t)$ over length of recording	Scalar
$SpO_{2,ref}(t)$	Sample-wise (1Hz) median of all four contact SpO_2 probes	At 1Hz
$SpO_{2,ref}$	Temporal median value of $SpO_{2,ref}(t)$ over length of recording	Scalar
$SpO_{2,cam}$	Camera based SpO_2 ($C1 + C2 * RR$)	Scalar
$Q(t)$	Signal-to-noise ratio of the raw red PPG signal	At 1Hz
Q	Temporal median value of $Q(t)$ over length of recording	Scalar
Q_{thr}	Threshold value to exclude noisy data	Scalar
C_1, C_2	Calibration constants	Scalar
τ	Temporal delay of $SpO_{2,ref}(t)$ with respect to the camera determined $SpO_2(t)$	Scalar
A^*_{rms}	Root mean square difference between $SpO_{2,cam}$ and $SpO_{2,ref}$ Asterisk indicates that temporal medians of traces are used	Scalar
A_{rms}	Root mean square difference between $SpO_{2,cam}$ and $SpO_{2,ref}$ for short time segments	Scalar
A^{**}_{rms}	Root mean square difference between $SpO_{2,cam}$ and $SpO_{2,ref}$ for short time segments. Asterisks indicate that individuals are unequally balanced.	Scalar
MD	Mean difference between $SpO_{2,cam}$ and $SpO_{2,ref}$	Scalar

DISCLOSURES

Name: Wim Verkruijsse, PhD.

Contribution: This author was involved in study design, conduct of the study, and manuscript preparation.

Name: Marek Bartula, MSc.

Contribution: This author was involved in study design and manuscript preparation.

Name: Erik Bresch, PhD.

Contribution: This author was involved in conduct of the study, data analysis, and manuscript preparation.

Name: Mukul Rocque, MSc.

Contribution: This author was involved in conduct of the study, data analysis, and manuscript preparation.

Name: Mohammed Meftah, BSc.

Contribution: This author was involved in study design, conduct of the study, and manuscript preparation.

Name: Ihor Kirenko, PhD.

Contribution: This author was involved in conduct of the study and critical reading.

This manuscript was handled by: Maxime Cannesson, MD, PhD.

ACKNOWLEDGMENTS

We greatly appreciate the invaluable suggestions from Siegfried Kaestle, Andreas Schlack, Alexander Dubielczyk, and Rolf Neumann from Philips Medizin-Systeme, Böblingen, Germany. Furthermore, we also thank our software expert Patriek Bruins from Philips Innovation Services, Eindhoven, The Netherlands.

REFERENCES

1. Aoyagi T, Kishi M, Yamaguchi K, Watanabe S. Improvement of the earpiece oximeter. In: Abstracts of the 13th annual meeting of the Japanese Society for Medical Electronics and Biological Engineering. Osaka; 1974:90–91
2. Severinghaus JW. Takuo Aoyagi: discovery of pulse oximetry. *Anesth Analg* 2007;105:S1–4
3. Hertzman A, Spielman C. Observations on the finger volume pulse recorded photoelectrically. *American Journal of Physiology* 1937;119:334–335
4. Wieringa FP, Mastik F, van der Steen AF. Contactless multiple wavelength photoplethysmographic imaging: a first step toward “SpO₂ camera” technology. *Ann Biomed Eng* 2005;33:1034–41
5. Humphreys K, Ward T, Markham C. Noncontact simultaneous dual wavelength photoplethysmography: a further step toward noncontact pulse oximetry. *Rev Sci Instrum* 2007;78:044304–6
6. Huelsbusch M, Blazek V. Contactless mapping of rhythmical phenomena in tissue perfusion using PPGI. In: Clough AV, Chen CT. *SPIE Proceedings of the Medical Imaging*. San Diego, CA: Physiology and Function from Multidimensional Images, 2002:110–117
7. Takano C, Ohta Y. Heart rate measurement based on a time-lapse image. *Med Eng Phys* 2007;29:853–7
8. Verkruysse W, Svaasand LO, Nelson JS. Remote plethysmographic imaging using ambient light. *Opt Express* 2008;16:21434–45
9. Bebout DE, Mannheimer PD. Effects of cold-induced peripheral vasoconstriction on pulse amplitude at various pulse oximeter sensor sites. *Anesthesiology* 2002;96:A558
10. Cooke J, Scharf J. Improving pulse oximeter performance. *Anesthesiology* 2002;96:A593
11. Sun Y, Thakor N. Photoplethysmography Revisited: From Contact to Noncontact, From Point to Imaging. *IEEE Trans Biomed Eng* 2016;63:463–77
12. Kong L, Zhao Y, Dong L, Jian Y, Jin X, Li B, Feng Y, Liu M, Liu X, Wu H. Non-contact detection of oxygen saturation based on visible light imaging device using ambient light. *Opt Express* 2013;21:17464–71
13. Bal U. Non-contact estimation of heart rate and oxygen saturation using ambient light. *Biomed Opt Express* 2015;6:86–97
14. Shao D, Liu C, Tsow F, Yang Y, Du Z, Iriya R, Yu H, Tao N. Noncontact Monitoring of Blood Oxygen Saturation Using Camera and Dual-Wavelength Imaging System. *IEEE Trans Biomed Eng* 2015; Epub ahead of print
15. Corral F, Paez G, Strojnik M. A photoplethysmographic imaging system with supplementary capabilities. *Optica Applicata* 2014;44:191–204
16. Guazzi AR, Villarroel M, Jorge J, Daly J, Frise MC, Robbins PA, Tarassenko L. Non-contact measurement of oxygen saturation with an RGB camera. *Biomed Opt Express* 2015;6:3320–38
17. Mannheimer PD. The light-tissue interaction of pulse oximetry. *Anesth Analg* 2007;105:S10–7
18. Farrell TJ, Patterson MS, Wilson B. A diffusion theory model of spatially resolved, steady-state diffuse reflectance for the non-invasive determination of tissue optical properties in vivo. *Med Phys* 1992;19:879–88
19. Fine I. The optical origin of the PPG signal. In: Genina EA, Derbov VL, Meglinski I, Tuchin VV. *SPIE Proceedings of the Saratov Fall Meeting. Saratov: Optical Technologies in Biophysics and Medicine XV; and Laser Physics and Photonics XV*, 2013:903103–1
20. Shvartsman LD, Fine I. Optical transmission of blood: effect of erythrocyte aggregation. *IEEE Trans Biomed Eng* 2003;50:1026–33
21. Kamshilin AA, Nippolainen E, Sidorov IS, Vasilev PV, Erofeev NP, Podolian NP, Romashko RV. A new look at the essence of the imaging photoplethysmography. *Sci Rep* 2015;5:10494
22. Verkruysse W, Lucassen GW, van Gemert MJ. Simulation of color of port wine stain skin and its dependence on skin variables. *Lasers Surg Med* 1999;25:131–9