

SCIENTIFIC DATA

OPEN Data Descriptor: A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in China

Yongbo Liang^{1,2,3}, Zhencheng Chen², Guiyong Liu⁴ & Mohamed Elgendi^{3,5,6}

Received: 6 October 2017
Accepted: 19 December 2017
Published: 27 February 2018

Open clinical trial data provide a valuable opportunity for researchers worldwide to assess new hypotheses, validate published results, and collaborate for scientific advances in medical research. Here, we present a health dataset for the non-invasive detection of cardiovascular disease (CVD), containing 657 data segments from 219 subjects. The dataset covers an age range of 20–89 years and records of diseases including hypertension and diabetes. Data acquisition was carried out under the control of standard experimental conditions and specifications. This dataset can be used to carry out the study of photoplethysmograph (PPG) signal quality evaluation and to explore the intrinsic relationship between the PPG waveform and cardiovascular disease to discover and evaluate latent characteristic information contained in PPG signals. These data can also be used to study early and noninvasive screening of common CVD such as hypertension and other related CVD diseases such as diabetes.

Design Type(s)	observation design • correlation study objective
Measurement Type(s)	blood pressure analysis
Technology Type(s)	photoplethysmography • oxygen monitor
Factor Type(s)	diagnosis
Sample Characteristic(s)	Homo sapiens

¹School of Electronic Engineering and Automation, Guilin University of Electronic Technology, Guilin 541004, PR China. ²School of Life and Environmental Sciences, Guilin University of Electronic Technology, Guilin 541004, PR China. ³School of Electrical and Computer Engineering, University of British Columbia, Columbia, Vancouver V6T 1Z4, Canada. ⁴Guilin People's Hospital, Guilin 541000, PR China. ⁵Department of Obstetrics & Gynecology, University of British Columbia, Columbia, Vancouver V6H 3N1, Canada. ⁶BC Children's & Women's Hospital, Vancouver, Vancouver V6H 3N1, Canada. Correspondence and requests for materials should be addressed to Z.C. (email: chenzhcheng@163.com) or to M.E. (email: moe.elgendi@gmail.com).

Background and Summary

The incidence of cardiovascular disease (CVD) has risen around the world in recent years overtaking the mortality rate of cancer, making CVD the number one killer of humans. Many studies have been conducted using noninvasive early diagnosis and screening techniques for CVDs such as hypertension and coronary artery sclerosis in order to discover more convenient and effective methods for the early identification of CVDs. Of these methods, the photoplethysmography (PPG) has become widely recognized as a low-cost non-invasive detection technology for CVDs. The cardiovascular parameters detected using PPG technology include heart rate, blood oxygen saturation, blood pressure, assessment of arterial stiffness¹, and pulse wave velocity, among others². The PPG signal includes information on the hemodynamic process^{3,4}, hemorheology⁵, and tissue status of the peripheral microcirculation system in the human body⁶. That is, the PPG signal is an aggregated expression of many physiological processes in the cardiovascular circulation system⁷. A physiological information database with high precision and a high sampling rate is urgently needed in PPG technology research in order to extract more cardiovascular parameters for the early screening and diagnosis of CVDs. We provide here a database containing physiological information and PPG waveform data collected over a year that can be used to research arterial blood vessel aging, arterial blood pressure detection⁸, and screening of hypertensive and diabetic patients based on PPG signals.

Electrocardiogram (ECG) signal and PPG signal can effectively estimate blood pressure which has been recently studied improved by researchers⁹. However, there are some limitations with collecting ECG and PPG simultaneously using a mobile phone¹⁰. For these reasons, and for simplicity, a few researchers have attempted to estimate blood pressure based using only PPG signals¹¹. The concept of estimating blood pressure (BP) using only PPG signals seems to promising and is optimally implemented when the PPG signal is of high quality. Therefore providing a database that can help with estimating BP using only PPG will help further research in this area.

This PPG and BP (PPG-BP) database integrates the deidentified, comprehensive clinical data of patients admitted to the Guilin People's Hospital in Guilin, China. The openness of the data allows clinical studies to explore and improve the understanding of relationships between cardiovascular health and PPG signals, with the final goal of creating a simple, effective non-invasive detection technology that is easy to use and wearable. This dataset has been collected from 219 subjects, aged 21–86 years, with a median age of 58 years. Males accounted for 48%. The dataset covers several diseases including hypertension, diabetes, cerebral infarction, and insufficient brain blood supply.

In summary, this unique non-invasive detection dataset for cardiovascular disease can be used in a wide range of in-depth research. In the following section, we describe the database detail records, database usage descriptions, and we also explain how to fully use the dataset. We also provide an example of the basic properties of the database that allows researchers to conduct research.

Methods

Experimental design and data acquisition

The dataset collection program involved acquiring information on the basic physiology of individuals, extracting information on cardiovascular diseases from hospital electronic medical records, collecting PPG waveform signals, and detecting instant arterial blood pressure at the same time. The data acquisition was conducted at the Guilin People's Hospital.

A customized portable hardware platform was designed, and consisted of a PPG sensor probe, microcontroller, and a matching app. Data were transmitted via Bluetooth. The PPG sensor model was SEP9AF-2 (SMPLUS Company, Korea), which contains dual LED with 660nm (Red light) and 905 nm (Infrared) wavelengths, with a sampling rate of 1 kHz and 12-bit ADC, and the hardware filter design is 0.5–12Hz bandpass. The microcontroller model was MSP430FG4618 (Texas Instruments company, USA) embedded on the probe's board to configure the ADC, fetch the data and send the data to the matching app via Bluetooth.

Waveform data is collected using a set of customized probes and a matching app that was developed based on Android Studio. The PPG detection probe used the infrared light and transmission method to collect fingertip PPG waveform data. These real-time data are transmitted to the matching app via Bluetooth. The app can control the detection probe, display the real-time waveform, and conduct a signal quality assessment of the PPG waveform in order to save the high-quality PPG wave segment. The arterial blood pressure is measured using the Omron HEM-7201 (Omron Company, Kyoto, Japan) upper arm blood pressure monitor, which is validated in ref. ¹².

The study was approved by the ethics committee of the Guilin People Hospital and the Guilin University of Electronic Technology in China. All participants gave written and informed consent before the study. They were compensated monetarily at 10 Yuan/h. Participants answered questions about age, gender, height, and weight and all initial data acquisition was conducted in a private, and comfortable clinical room.

As shown in Figure 1, before beginning with signal collected, each individual was asked to sit in an office chair in the most comfortable posture and to relax their arms on an empty desk. Each individual had 10 min to adapt to the environment and adjust their breathing after entering the data collection room. The specific collection settings were as follows: The PPG signal was collected at the fingertip of the left index finger, the arterial blood pressure was collected from the right forearm, all of which was

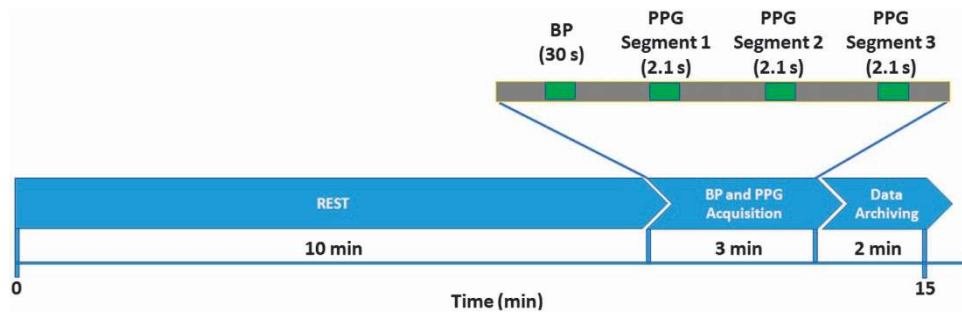


Figure 1. Measurement Protocol. The duration of the whole experiment was approximately 15 minutes. The photoplethysmogram (PPG) and blood pressure (BP) were collected within 3 minutes. Three PPG segments collected per subject, with a duration of 2.1 seconds each.

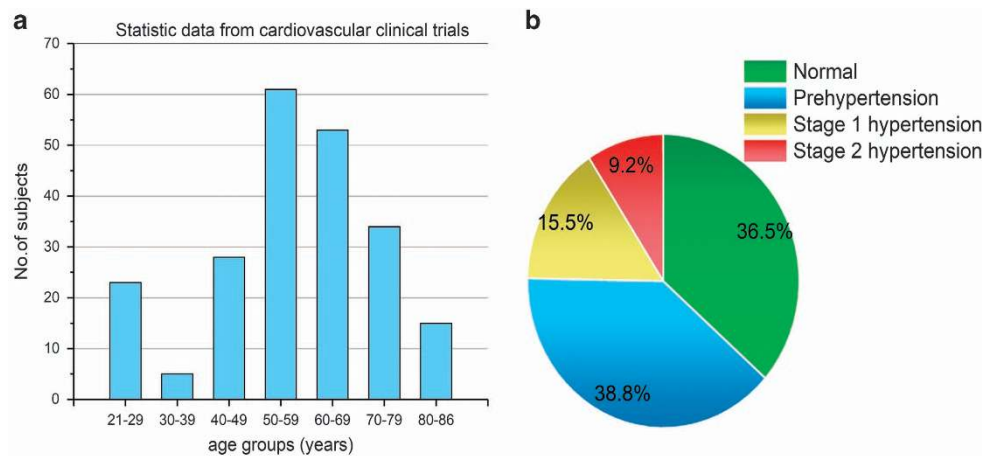


Figure 2. Statistics about the PPG-BP dataset. (a) histogram of age groups; (b) pie chart of blood pressure stages.

completed within three minutes. The arterial blood pressure measurement was performed by the hospital nurse.

During signal acquisition the sampling precision of waveform data was set to a sampling rate of 1 kHz, with 12 bits AD conversion precision. Three segments were recorded and saved per subject, each segment included 2100 sampling points, which corresponds to a length of 2.1 seconds. During the 3 min data collection phase, every PPG segment of a particular subject scored a Skewness SQI value; values greater than Zero¹³ were saved, and if a value was less than Zero the app prompted the user to recollect the PPG signal. This step was developed to reduce including PPG segments with high noise and motion artifacts.

The BP collection device (Omron HEM-7201) requires at a minimum a 30 second waveform to detect the systolic and diastolic period. The BP reading represents the blood pressure value for the 3 minute data recording for each subject, as shown in Figure 1. During the data collection process, we aimed to collect the BP and PPG data immediately after each other respectively. Three PPG segments were saved during the data collection period in addition to the BP recording. Every participant was asked to breathe as they normally would on day-to-day basis for practical applications. Note that we did not investigate the baroreflex response to stress. The dataset includes BP and PPG information from subjects that were diagnosed with normotension, prehypertension, and stage I/stage II hypertension, which can be helpful and valuable for researchers.

Patient characteristics

The dataset was collected from 219 adult subjects and currently contains 657 PPG waveform segments. The dataset covers individuals aged 21–86 years, and males account for 48% of participants. The dataset also covers several different CVDs, including hypertension, cerebral infarction, and insufficient brain blood supply and other related diseases such as diabetes. The statistical results are shown in Figure 2.

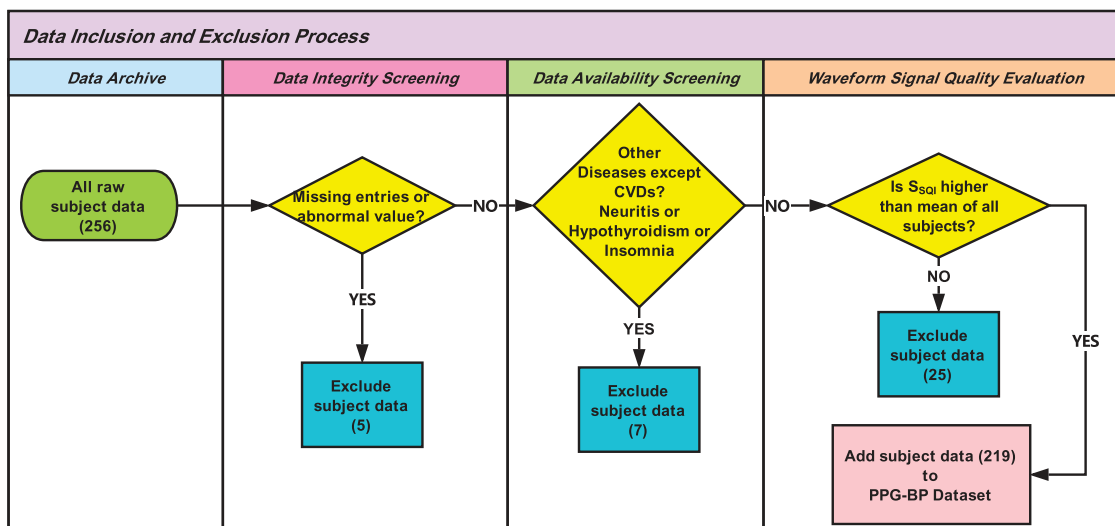


Figure 3. A process flowchart of data inclusion and exclusion. Note, S_{SQI} stands for Skewness signal quality index.

De-identification

In the process of creating the dataset, the first task was to de-identify each participant and remove personal information such as name, telephone number, address, date, and so on.

Data Records

The dataset has been fully uploaded to the network, and users can download them through the *Figshare* repository with the title *PPG-BP Database* and reference (Data Citation 1).

The dataset comprises 1 table file and 219 waveform data folders, which include three 2.1-second-length infrared PPG signal text files and physiological information recording files. Among these, the PPG signal data is the 2.1-second-length 12 bits AD raw value. The ID_1, ID_2, and ID_3 text files represent three separate segments of waveform data Table 1 (available online only).

The "PPG-BP database.xlsx" table file contains aggregated subjects of physiological information and disease information. Information records include ID, sex, age, height, weight, systolic pressure, diastolic pressure, heart rate, and disease records.

Before the participant record is archived, it was required to conduct data integrity screening, data availability screening, and a waveform signal quality evaluation (to remove inconsistent, abnormal, and high noise data) in order to form a high-quality dataset. The detailed process of inclusion and exclusion, as shown in Figure 3, is described as follows:

- 1 **Data integrity screening:** This process includes the screening of missing and abnormal values for: basic physiological information, disease information, blood pressure, heart rate, and 3 segments waveform data. If one or more items are missing or if there was an abnormal value, the participant record was removed.
- 2 **Data availability screening:** This dataset is designed to focus on the clinical information for CVDs and other closely related diseases such as diabetes. Data from the CVD patients who were diagnosed with non-CVD diseases (except diabetes) were excluded during the screening process to ensure that the dataset only contains data from participants who were diagnosed with the disease of interest.
- 3 **Waveform signal quality evaluation:** All 3 segments for each participant went through a signal quality evaluation, and a robust signal quality index (SQI) method was applied in order to achieve this step. If the SQIs of the 3 segments in one subject were lower than the mean SQI calculated from the segments of all subjects, the subject data was removed.

Technical Validation

In order to thoroughly analyze the abundant information on cardiovascular physiology contained in the PPG signals, it was necessary to obtain the most authentic, high-precision, and high-quality PPG signal. In addition to controlling the process of data collection as accurately as possible, the PPG signal quality evaluation process was important in acquiring excellent quality signals^{14,15}. This process can effectively prevent unfit waveform data from being saved and ensure that the saved data have complete heartbeat cycles, less noise, and lower drift and motion artifacts.

Although the potential value of the PPG signal is well known^{16,17}, it is difficult to acquire authentic and rich PPG signals and to extract subtle characteristics that signify problems such as difference in body

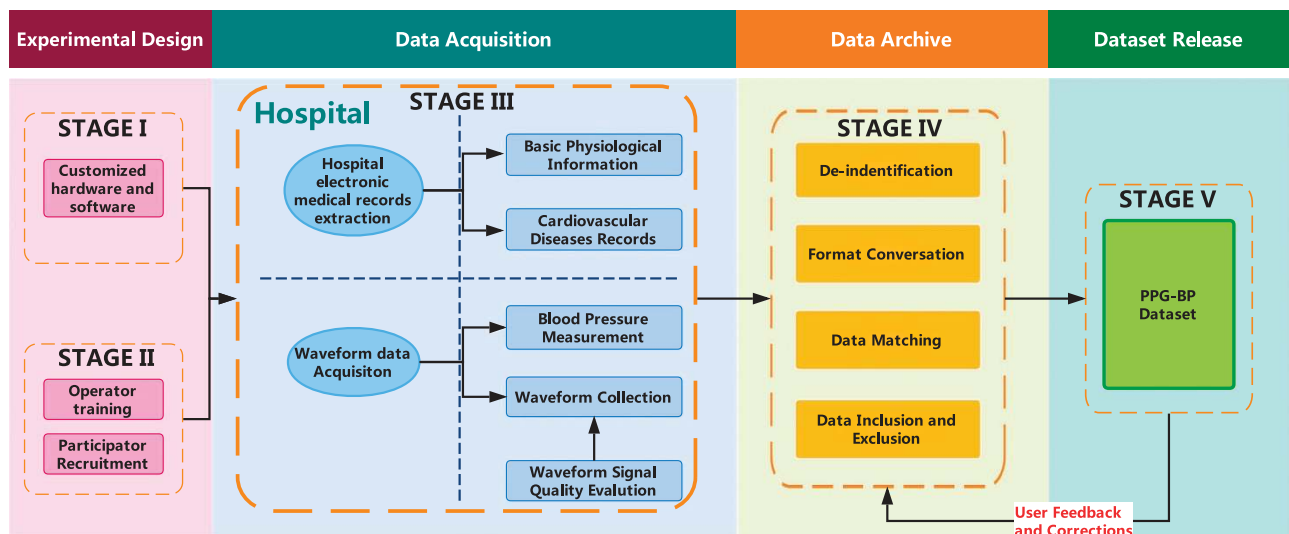


Figure 4. Overview of the PPG-BP Dataset.

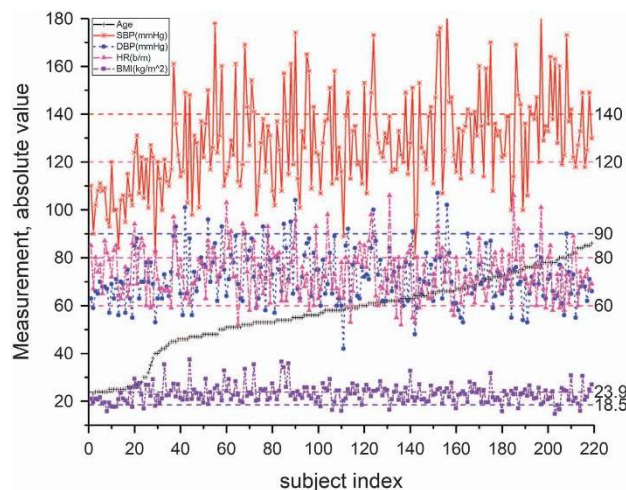


Figure 5. Physiological records for all subjects ordered in ascending order according to their age.

tissue^{18,19}, disturbance of motion^{20–23}, control of the acquisition process^{24,25}, etc. This is worth more research and exploration²⁶. In order to avoid or eliminate the above mentioned issues, the data collection process was conducted in a controlled, quite environment with minimal interferences. Figure 4 shows an overview of the data collection process. Figure 5 shows a statistics table of the physiological records for all subjects in the database.

The process of data collection experiment consists of five stages. The stage I and stage II conduct some preparations of the customized hardware and software, training of operators and recruitment of participants. Stage III is the phase of data collection in hospitals, including the acquisition of basic physiological information of participants, hospital electronic medical records, PPG signals and blood pressure data. Stage IV is the data archiving part, including de-identification, format conversion, data matching, data inclusion and exclusion for all the collected data. Stage V is the public release of the PPG-BP dataset; researchers can download the dataset and validate their algorithms.

At present, Perfusion Indices (P_{SQI}) are regarded as the gold standard of PPG Signal Quality Indices (SQI). Various other signal quality evaluation methods have been proposed and studied in order to identify more simple and accurate evaluation methods for signal quality assessment. Elgendi¹³ compared eight different signal quality indices: P_{SQI} , Skewness (S_{SQI}), Kurtosis (K_{SQI}), Entropy (E_{SQI}), Signal-to-noise ratio (N_{SQI}), Zero-crossing (Z_{SQI}), Matching of multiple systolic wave detection algorithms (M_{SQI}), and Relative power (R_{SQI}). For varying lengths of PPG waveform recordings (i.e., from 2 s to 30 s), the S_{SQI} method demonstrated better performance when compared to other methods (P_{SQI} , K_{SQI} , E_{SQI} , N_{SQI} , M_{SQI} , Z_{SQI} , and R_{SQI})¹³. Moreover, PPG waveform classification is possible with 2 s length recording (excellent vs. unfit) using the S_{SQI} index¹³. These results motivated collecting PPG signals with 2 s length.

Krishnan et al.²² introduced and tested the Skewness statistic and found that S_{SQI} had a certain connection with the quality of PPG waveform. Other researchers validated this observation, and Elgendi¹³ found it to be the optimal method for assessing SQI in PPG signals. Skewness is used to measure the probability distribution of symmetric signals, which can distinguish the periodic, symmetrical, stationary signals and sudden jumps, periodic signals, and irregular signals. The specific definition is as follows:

$$S_{SQI} = 1/N \sum_{i=1}^N [x_i - \hat{\mu}_x / \sigma]^3,$$

where N is the sample number of the PPG signal, and $\hat{\mu}_x$ and σ are the empirical estimates of the mean and standard deviation of x_i , respectively.

In the process of data collection, the data is evaluated using the PPG signal quality before it was saved, and the evaluation method adopted the S_{SQI} index. Each segment of PPG signal was evaluated against the classification threshold of excellent, acceptable, or unfit PPG waveform in order to determine whether it should be saved.

During the evaluation of signal quality for each participant, the S_{SQI} for the three segments were compared. Among the three the segments, the segment with the highest S_{SQI} was deemed as "high quality", the segment with the lowest S_{SQI} was deemed as "low quality" and the remaining segment was deemed as "medium quality". Note, we are providing the PPG segments and their corresponding S_{SQI} values to make it easier for investigators to select the segment with highest quality. Additionally, with the availability of the three S_{SQI} values, researchers will be able to analyze each segment, if needed, for validation, etc.

Usage Notes

The dataset is distributed in the normal standard file format (text, xlsx) and can be read and processed by a variety of packages, including Matlab, Python, and R. In addition, when using this dataset, users should first perform an effective waveform quality evaluation to select the high-quality waveform segment from the three waveform segments from each participant.

There are multiple potential uses for this dataset, the most obvious of which is to validate various CVD diseases and diabetes through pattern recognition and machine learning methods. About one-third of the dataset includes hypertensive or diabetic patients. A series of PPG characteristics can be defined and mined, the intrinsic association between characteristic and physiological process can be studied, and the prediction of blood pressure or classification of hypertension can be explored. It is important to analyze the physiological information contained in the PPG waveform comprehensively.

Example usage

This dataset can be used to analyze the definition and extraction of morphological information of the PPG waveform. The PPG wave mainly consists of the systole period and the diastole period. At the same time, depending on different human cardiovascular health statuses, the tidal and diastolic wave may also be shown. These characteristics²⁷ can be defined and extracted by means of digital signal processing such as time and frequency domain processing or wavelet transform. The time span, amplitude, area, ratio, frequency, and energy parameter are the main features that can be extracted from the PPG signal.²⁸ The accurate identification and extraction of these characteristics has potential value in analyzing vascular aging, blood pressure, and arteriosclerosis, among others.

References

1. von Wowern, E., Ostling, G., Nilsson, P. M. & Olofsson, P. Digital Photoplethysmography for Assessment of Arterial Stiffness: Repeatability and Comparison with Applanation Tonometry. *PLoS ONE* **10**, e0135659 (2015).
2. Allen, J. Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement* **28**, R1–39 (2007).
3. Leguy, C. A., Bosboom, E. M., Gelderblom, H., Hoeks, A. P. & van de Vosse, F. N Estimation of distributed arterial mechanical properties using a wave propagation model in a reverse way. *Medical engineering & physics* **32**, 957–967 (2010).
4. Alian, A. A. & Shelley, K. H. Photoplethysmography. *Best practice & research. Clinical anaesthesiology* **28**, 395–406 (2014).
5. Njoun, H. & Kyriacou, P. A. Photoplethysmography for the Assessment of Haemorheology. *Scientific reports* **7**, 1406 (2017).
6. Lee, C., Sik Shin, H. & Lee, M. Relations between ac-dc components and optical path length in photoplethysmography. *Journal of biomedical optics* **16**, 077012 (2011).
7. Mukkamala, R. et al. Toward Ubiquitous Blood Pressure Monitoring via Pulse Transit Time: Theory and Practice. *IEEE transactions on bio-medical engineering* **62**, 1879–1901 (2015).
8. Xing, X. & Mingshan Sun. Optical blood pressure estimation with photoplethysmography and FFT-based neural networks. *Biomedical optics express* **7**, 3007–3020 (2016).
9. Zheng, D. & Murray, A. Non-invasive quantification of peripheral arterial volume distensibility and its non-linear relationship with arterial pressure. *Journal of Biomechanics* **42**, 1032–1037 (2009).
10. Jung, S.-J., Myllylä, R. & Chung, W.-Y. Wireless machine-to-machine healthcare solution using android mobile devices in global networks. *IEEE Sensors Journal* **13**, 1419–1424 (2013).
11. Addison, P. S. Slope transit time (STT): A pulse transit time proxy requiring only a single signal fiducial point. *IEEE Transactions on Biomedical Engineering* **63**, 2441–2444 (2016).
12. Cho, K., Tian, M., Lan, Y., Zhao, X. & Yan, L. L. Validation of the Omron HEM-7201 upper arm blood pressure monitor, for self-measurement in a high-altitude environment, according to the European Society of Hypertension International Protocol revision 2010. *Journal of Human Hypertension* **27**, 487 (2013).

13. Elgendi, M. Optimal Signal Quality Index for Photoplethysmogram Signals. *Bioengineering* **3**, 21 (2016).
14. Orphanidou, C. *et al.* Signal-quality indices for the electrocardiogram and photoplethysmogram: derivation and applications to wireless monitoring. *IEEE journal of biomedical and health informatics* **19**, 832–838 (2015).
15. Silva, I., Lee, J. & Mark, R. G. Signal quality estimation with multichannel adaptive filtering in intensive care settings. *IEEE transactions on bio-medical engineering* **59**, 2476–2485 (2012).
16. Tang, S. C. *et al.* Identification of Atrial Fibrillation by Quantitative Analyses of Fingertip Photoplethysmogram. *Scientific reports* **7**, 45644 (2017).
17. Shin, H. & Min, S. D. Feasibility study for the non-invasive blood pressure estimation based on ppg morphology: normotensive subject study. *Biomedical engineering online* **16**, 10 (2017).
18. Bruno, R. M. *et al.* Different Impact of Essential Hypertension on Structural and Functional Age-Related Vascular Changes. *Hypertension* **69**, 71–78 (2017).
19. Liu, J. *et al.* Multi-wavelength photoplethysmography method for skin arterial pulse extraction. *Biomed Opt Express* **7**, 4313–4326 (2016).
20. Park, C., Shin, H. & Lee, B. Blockwise PPG Enhancement Based on Time-Variant Zero-Phase Harmonic Notch Filtering. *Sensors* **17** (2017).
21. Peng, F., Liu, H. & Wang, W. A comb filter based signal processing method to effectively reduce motion artifacts from photoplethysmographic signals. *Physiological measurement* **36**, 2159–2170 (2015).
22. Krishnan, R., Natarajan, BB & Warren, S Two-stage approach for detection and reduction of motion artifacts in photoplethysmographic data. *IEEE Trans. Biomed. Eng.* **57**, 1867–1876 (2010).
23. Jindal, V., Birjandtalab, J., Pouyan, M. B. & Nourani, M. An adaptive deep learning approach for PPG-based identification. Conference proceedings. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. *IEEE Engineering in Medicine and Biology Society. Annual Conference 2016* 6401–6404 (2016).
24. Grabovskis, A., Marcinkevics, Z., Rubins, U. & Kviesis-Kipge, E. Effect of probe contact pressure on the photoplethysmographic assessment of conduit artery stiffness. *Journal of biomedical optics* **18**, 27004 (2013).
25. Hsiu, H., Hsu, C. L. & Wu, T. L. Effects of different contacting pressure on the transfer function between finger photoplethysmographic and radial blood pressure waveforms. *Proceedings of the Institution of Mechanical Engineers. Part H, Journal of engineering in medicine* **225**, 575–583 (2011).
26. Palmeri, L. *et al.* Photoplethysmographic waveform characteristics of newborns with coarctation of the aorta. *Journal of perinatology : official journal of the California Perinatal Association* **37**, 77–80 (2017).
27. Resit Kavsaoglu, A., Polat, K. & Recep Bozkurt, M. A novel feature ranking algorithm for biometric recognition with PPG signals. *Computers in biology and medicine* **49**, 1–14 (2014).
28. Elgendi, M. On the Analysis of Fingertip Photoplethysmogram Signals. *Current Cardiology Reviews* **8**, 14–25 (2012).

Data Citations

1. Liang, Y., Liu, G, Chen, Z & Elgendi, M. *Figshare* <https://doi.org/10.6084/m9.figshare.5459299> (2017).

Acknowledgements

This work has been partly supported by the development project of major scientific research instruments of the National Natural Science Foundation of China (grant No. 61627807), Guangxi Postgraduate Education Innovation Program and the Innovation Project of GUET Graduate Education (grant No. 2016YJ CXB01).

Author Contributions

Y.L. designed the data acquisition device and system, managed the database development process and drafted the manuscript. Dr. G.L. organized and managed the data collection process in hospital. Dr. Z.C. advised and supervised the project. Dr. M.E. led the data processing, audited data quality, and main contributor to the writing of the paper.

Additional Information

Table 1 is only available in the online version of this paper.

Competing interests: The authors declare no competing financial interests.

How to cite this article: Liang, Y. *et al.* A new, shortrecorded photoplethysmogram dataset for blood pressure monitoring in China. *Sci. Data* **5**:180020 doi: 10.1038/sdata.2018.20 (2018).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files made available in this article.

© The Author(s) 2018