MASTER THESIS

Is it feasible to study heart rate values using wearables instead of traditional equipment in research?

L.E.A. Teekens
s1090011

MASTER PSYCHOLOGY
HUMAN FACTORS & ENGINEERING PSYCHOLOGY

EXAMINATION COMMITTEE
Dr. Andrew Spink (Noldus Information Technology BV)
Dr. M.L. Noordzij
H.G. van Lier

25-11-2016

UNIVERSITY OF TWENTE.

Abstract

Wearables are getting more and more common and researchers are wondering if they could be a feasible alternative for traditional research equipment. The reason wearables are not widely used in research is that there are concerns about the differences between the values coming from wearables and traditional research equipment. This study examined heart rate data from four different wearables on their agreement with ECG obtained heart rate values in an ambulatory setting. The Bland-Altman method was used to determine the agreement. None of the wearables was accurate enough to be used in research regarding absolute heart rate. With the results of this study, no conclusions can be done about using wearables for heart rate zones or in other research situations where the absolute value is less important. To be able to use wearables for absolute heart rate values in research, first further development of the technology is needed. After that, there should be validation studies, which employ appropriate statistical measures, as discussed in the present thesis.

Samenvatting

Wearables worden steeds meer gebruikt en in het onderzoek is de vraag of deze een alternatief kunnen bieden aan de traditionele onderzoeksapparatuur. De reden dat dit nog niet gedaan wordt is de vraag in hoeverre wearables accurate data geven. In dit onderzoek werden de hartslagwaardes van vier verschillende wearables vergeleken met hartslagwaardes van een ECG op een ambulante manier. De Bland-Altman methode werd gebruikt om de overeenkomst vast te stellen. Geen van de wearables was accuraat genoeg om gebruikt te worden in onderzoek waar absolute hartslagwaardes belangrijk zijn. Met deze studie kunnen geen conclusies getrokken worden over situaties waar absolute waardes minder belangrijk zijn. Om wearables voor absolute hartslagwaardes te gebruiken, is eerst verdere ontwikkeling van de technologie nodig. Daarna is zijn er validatiestudies nodig met geschikte statistische analyses, zoals omschreven in deze masterthese.

Acknowledgements

This thesis has been written as my graduation project for the master's programme Psychology with as specialization "Human Factors & Engineering Psychology" at the University of Twente. This study has been made possible by Noldus Information Technology, where I had the chance to do an internship. Noldus provided me with the use of several devices, a work place and even a participant pool. I would like to thank all colleagues of Noldus who participated in the study, helped me with setting up the research equipment, restarting my computers, helped me figure out Acqknowledge software, giving me some biology lessons, and even write a small program for me to use. I would like to thank some people in particular, who helped me throughout the whole project.

First of all I would like to thank Matthijs Noordzij for being my supervisor for this thesis and for his inspiration, support and feedback. Thanks for being so interested in this project and motivating me with your energy. I would also like to thank Erika van Lier for being the second supervisor and giving me insightful feedback and asking for clarification when it was needed. Furthermore, I would like to thank Andrew Spink for being my external supervisor from Noldus. Thank you for sharing your knowledge, asking the right questions and telling me when it's time for lunch. At last, I would like to thank my family and especially my husband Peter, for supporting me in all the choices I made. Thank you for being my biggest support and having dinner ready, no matter the time I got home.

Table of Contents

Introduction

Approach

This thesis examines the feasibility of using wearable biosensors in scientific research instead of classic physiological equipment. A requirements analysis was conducted to map the needs and wishes for wearables use in research (Noldus Information Technology, 2016). This showed that the most important physiological measures were heart rate, heart rate variability and skin conductance. Since skin conductance measurement is rare in wearables, the present study focuses on heart rate measurements. Another finding is that wearables in research would be used in real life situations, such as classrooms. Most use cases include a situation in which the participants are in a stationary position. The most mentioned concern for using wearables in research is the validity and reliability. Validity in this case refers to the extent in which wearables take the same measure as traditional research equipment and reliability refers to consistency of the measure when the measurement is repeated. The present study focuses on the extent to which wearables give the same values as traditional research equipment for physiology, the gold standard. Four wearables, Microsoft Band 2 (Microsoft, 2015), Garmin Forerunner 235 (Garmin, 2015), Motorola Moto 360 2nd generation (Motorola, 2015) and a prototype from Philips, Elan (Philips, 2015), were tested against traditional research equipment. Even though the actual results will be outdated in a few years' time, this thesis will be part of the validation literature to illustrate the technical improvement of wearables. This thesis illustrates the importance of correct methodology for validation studies, by examining previous validation studies and review articles about correct methodology.

In the beginning of this introduction the current use of wearables is discussed, next there is a short explanation how heart rate is measured, furthermore some validation studies are reviewed and finally the correct way to determine agreement between devices is examined.

Wearables

There is no clear definition for a wearable in the literature (Chuah et al., 2016); the terms wearable, wrist-worn device and smartwatch are used interchangeably, even though there are wearables worn on other places than the wrist. Smartwatches are watches with extra functionalities, such as the access to email by using a connection to a smartphone, or step count. A well-known example is the Apple Watch (Apple, 2015). There are also wrist-worn devices without a screen to use for activity tracking, such as the Fitbit Flex (Fitbit, 2013). An example

of a wearable not worn on the wrist is the My UV-patch (L'Oreal, 2016), which can be attached anywhere on the skin to measure ultraviolet radiation.

Wearables occur in many different forms, varying from wristbands to inner ear pieces. Their functions and uses are also diverse, ranging from step count to fertility tracking. Schwartz and Baca (2016) defined three types of wearables: commercial wearables intended for the general public, advanced wearables for use in research and experimental wearables, usually in a developmental stage. Commercial wearables are the most common wearables, focusing on measuring physical activity, such as travel distances, the other two types of wearables have broader measurements, for example skin temperature or oxygen saturation. Other differences in the types of wearables are the price, implemented algorithms and possibility to access the raw data (Schwartz & Baca, 2016). Consumer wearables have algorithms to smooth and structure the data, while with advanced and experimental wearables the raw data is given. For the present study three commercial wearables and an experimental wearable were used, all wrist-worn with heart rate measurement. These wearables were chosen because they are assumed to have decent accuracy, as claimed by manufacturers' research and consumer tests. Furthermore, these wearables were provided by Noldus Information Technology and they were appropriate for use in other project.

In the next paragraphs, the current state of wearable use in research will be explained, after which there will be an elaboration on consumer wearable use, to shine light on the different uses of wearables.

Use in research. Wearables are getting used more commonly in research (Wac & Tsiourti, 2014). Sometimes commercial devices are used (e.g. Swift et al., 2015; Tiedemann, Hassett & Sherrington, 2015), but there are also devices marketed towards research, such as the ActivPal (PAL technologies, 2007) and the Empatica E-series (Empatica, 2012). A great advantage of wearables is their unobtrusive nature, so the user is not restricted in their tasks and movements (Wac & Tsiourti, 2014). This means that the data tracking feels more natural for the user and that there is no need for a lab environment, making the data more ecologically valid (Fahrenberg, Myrtek, Pawlik & Perrez, 2007; Wac & Tsiourti, 2014). Other advantages of wearables include ease of use and higher participant compliance (Noldus, 2016), since it is likely that participants are more at ease with attaching a device to their wrist than having electrodes placed on their chest as with the traditional research equipment (Wac & Tsiourti, 2014). It is also possible to easily test multiple participants at once when using wearables (Wac & Tsiourti, 2014).

An area in which wearables could be a promising alternative for traditional research equipment is ambulatory assessment. Ambulatory assessment is the use of computer-assisted methodology for monitoring purposes while the participant undergoes their daily activities (Trull & Ebner-Priemer, 2012). This type of observation originates from medicine, but is getting more commonly used in other fields for research purposes, such as psychology (Fahrenberg et al., 2007; Trull & Ebner-Priemer, 2012). Wac and Tsiourti (2014) suggested exploratory research when using wearables in ambulatory assessment, given the complexity of the data and the unknown circumstances. This is because with ambulatory assessment, there are many possible factors that could influence that data which cannot be quantified. In an experiment setting, most of these factors are known, or equal between participants.

Consumer use of wearables. Wearables are most often used by consumers. Their main motivation for using wearables is self-tracking. This can be done for specific reasons, such as measuring heart rate while working out (Thompson, 2016) and adjusting their training using it (Tholander & Nylander, 2015). Self-tracking can also be used to keep track of certain habits, like sleep patterns or food intake. Self-tracking is the basis of the movement of Quantified Self (Lupton, 2014). The Quantified Self is a community in which people track aspects of their everyday life, these aspects could be anything, as long as it is quantifiable. Examples include blood pressure, amount of exercise and food intake (Swan, 2012). The rise of wearables has made the tracking easier, because some of the data of interest can be tracked automatically now, like sleep or activity (Lupton, 2014; van Dijk, Beute, Westerink & Ijsselstein, 2015). Self-tracking is an interesting movement for research, because of the availability of data about behavioral patterns from various persons.

There are different modes of self-tracking. The first mode is private self-tracking, in which case the user decides what and how they track and with whom they share their data (Lupton, 2014). Pushed self-tracking is a mode in which the user is nudged to use self-tracking. This occurs for example in the United States, where employers pay for their employee's health insurance and where the employer wants to increase the employee's activity (Lupton, 2014). Self-tracking with an app or wearables creates a database with all data from the users. This leads to the mode of self-tracking in which data is shared within a community; communal self-tracking. The availability of big data has also sparked interest of third parties, leading to the last two modes of self-tracking. There is imposed self-tracking, where users have limited choice in whether they self-track or not and the data is used for other's benefit (Lupton, 2014). For example in a company, the location of employees is recorded, even when they are not working (Fort, Raymond & Shackelford, 2016). The last mode is exploited self-tracking, where data

from the user is sold or used for other's benefit. This is for example the case with Strava (Strava, 2009), an app in which users can track their training and performance in different sports. Strava uses this data for example to make heat maps with the most used cycling roads. Self-tracking has different forms and users might not always track for their own benefit. Regardless of the mode of self-tracking, it leads to a database full of personal information, but the users may not be aware that the tracking is not primarily for their own benefit. (Lupton, 2014).

The availability of self-tracking data could contribute to research, as is for example the case with the period tracking app Clue (Clue, 2013). The Clue company uses the information of its users to do research. Users indicate if they experienced certain symptoms, events and feelings per day. With some of their research projects, like when testing the claim that periods can synchronize when spending much time with someone else, they ask for volunteers to track more factors relevant for the project. In other cases, it might not be so clear that data is being used by others than the person doing the tracking. The use of other's data is possible because of ambiguities about the ownership of this data, so users could be unaware that their data is used by third parties (Lupton, 2014).The ethical implications of the use of wearables are not clear for users, and might not be for researchers.

Heart rate

Wearables have the possibility to gather various information about the user, as stated above, but the present study focuses on heart rate. The different ways to measure heart rate will be explained in the next paragraphs.

Measurement. Heart rate is traditionally measured by an electrocardiogram (ECG). ECG is seen as the gold standard in monitoring heart rate (Lemay et al., 2014). An ECG is obtained by using electrodes measuring the electrical activity of the heart (Berntson, Quigley & Lozano, 2007). There are different methods of placing the electrodes in order to get an ECG, but for many studies three electrodes are placed on the chest and shoulder. The electrical activity obtained by the electrodes is visualized, the ECG signal, shown in Figure 1.

The cardiac cycle is visible in the ECG, it starts at the P wave, where the heart does not pump and fills with blood due to an electrical signal generated by the heart. After this the heart contracts, shown by the QRS complex. Then a recovery phase occurs, the T wave, in which the heart is prepared for the next pump movement (Berntson et al., 2007). The number of R-peaks happening in the timeframe of a minute is used to determine the heart rate, which is most commonly expressed as beats per minute (bpm) (Lemay et al., 2014).

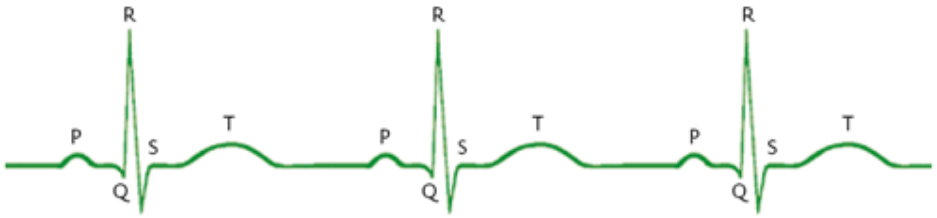


Figure 1. ECG waves with typical points (Nederlandse Vereniging voor Cardiologie).

Photoplethysmogram (PPG) is another often used method for measuring heart rate (Lemay, 2014). Wearables that allow for heart rate measurements make use of PPG. This method makes use of light and measures changes in light absorption, to indicate blood volume changes. When a heart beats occurs, the blood flows in bigger volumes trough the body, which is visible due to the change in light absorption this causes. This entails that with a PPG sensor the Blood Volume Pulse (BVP) is measured, the phasic change in blood volume with each heartbeat. Based on the change in BVP, the heart rate is determined. The visual representation of this can be found in Figure 2, which also shows how this differs from an ECG signal. In the PPG signal, every peak represents a heartbeat.

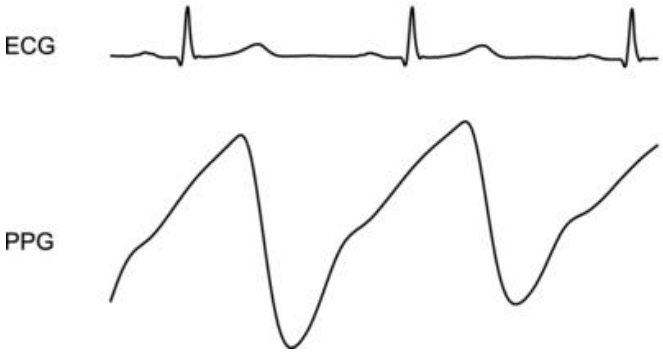


Figure 2. Visualization of the ECG signal and PPG signal (Allen, 2007).

The important difference in ECG and PPG are the way of obtaining information, with ECG using a bio-potential technique, and PPG using an optical method (Lemay et al., 2014). This means that with an ECG a bio-potential, an electric signal, is measured, while with PPG the measurement is done by detecting visual differences.

There are several studies comparing both methods, leading to the conclusion that PPG can be used to accurately measure heart rate, as pulse rate is in agreement with heart rate (e.g. Lemay et al., 2014; Schäfer & Vagedes, 2013). These studies did experiments with PPG sensors designed for research or health care, which were most of the time attached to the participant's

finger. In the present study, the PPG sensors of consumer wearables worn on the wrist are tested in an ambulatory setting.

Use. Heart rate is a measure of activity of the autonomic nervous system. The autonomic nervous systems receives and sends information from the internal organs (Kalat, 2007). This means that heart rate is dependent on the responses of the autonomic nervous system. This entails that some factors, such as stress, emotion and activity, indirectly influence heart rate (Wilhelm, Pfaltz & Grossman, 2006; Kreibig, 2010). Heart rate in humans in a resting position is normally between 60 and 100 bpm (American Heart Association, 2015). Very fit persons could have a heart rate as low as 40 bpm and while exercising, the heart rate could go up to 200 bpm for young healthy adults (American Heart Association, 2015). Since the normal heart rate ranges are broad, it is important to keep in mind that average heart rate can vary per person, it is for example very dependent on age (Acharya, Kannathal, Sing, Ping & Chua, 2004).

Validation studies

As mentioned before, PPG is a different method to obtain heart rate than ECG, which causes concern to use the two method interchangeably. This is why validation studies are needed in which the agreement is assessed between data from a PPG and data from an ECG. Wearables make use of PPG, while most chest straps use electrodes to obtain heart rate. Consumers also have concerns about the measurements of their wearables. Without access to scientific validation papers about wearables, many web logs try to validate their new wearables for themselves by comparing with a chest strap (e.g. DC Rainmaker (www.dcrainmaker.com); Wareable (www.wareable.com)). These kind of tests most often include only one participant, but most bloggers do use the same setting, such as training scheme and route, when testing new wearables. The tests almost never include a statistical calculation, only an estimate of agreement after visually comparing the data. This gives some information about the validity of the device, but no generalizable conclusions. This means that even though there are many tests regarding wearables, there are no clear conclusions about the reliability and validity of consumer wearables. It is important to note that the demands for consumers might be different than demands for scientific purposes. For consumer the approximate heart rate might be enough, while for scientific use the absolute heart rate is often necessary.

Accelerometers have been scientifically tested, and dependent on the brand, are deemed a valid measure of activity (e.g. Welk, Schaben & Morrow, 2004; Ferguson, Rowlands, Olds & Maher, 2015). There is no such clear claim about the use of heart rate measurements from wearables (Patel, Asch & Volpp, 2015). However, there are some validation studies for heart

rate measurements from wearables. For example, Stahl, An, Dinkel, Noble and Lee (2016) tested six different consumer wearables and stated those as acceptable for the recreational athlete and for use in research. In contrast, Wang et al. (2016) also tested six consumer wearables, four of them the same as Stahl et al. (2016) and concluded variable accuracy and that none of the wearables achieved the same accuracy as a chest strap based monitor. They recommended to use electrode-containing chest monitors for situations in which accurate heart rate is needed. Spierer, Rosen, Litman and Fujii (2015) compared different wearables, including the Alpha Mio with a gold standard, Polar. They concluded that on average the wearable devices were accurate, but, especially the Mio Alpha, were less accurate in intensive activities and for participants with a darker skin color (Spierer et al., 2015). A few other factors that could influence measurement with a wearable are personal characteristics (e.g. skin color), ambient temperature, ongoing behavior and potential misplacement (Wac & Tsiourti, 2014).

Reproducibility

It is important that validation study results are reproducible, meaning in this case that devices give similar results in different settings or with different participants. De Vet, Terwee, Knol and Bouter (2006) stated that reproducibility includes agreement and reliability parameters. The agreement concerns the measurement error and assesses how close the scores for repeated measures are, while the reliability has to do with the variability between subjects, despite the measurement errors (de Vet et al., 2006).

To draw valid conclusions about the measurements of the wearables, relevant statistical calculations should be used. Correlations are in general used as a measure of association, and in some cases for assessing agreement between devices (e.g. Poh, Swenson & Picard, 2010). Bland and Altman (1986) were to first to state that correlations are not an appropriate measure for assessing agreement, since a high correlations indicates that the output of the devices are related, but not that they necessarily agree. Bland and Altman (1986) stated that data in poor agreement can give high correlations: Any straight line in a figure of measurement points would give a high correlation, but high equality requires a straight line with the same x and y values. An example with simulated data is shown in Figure 3.

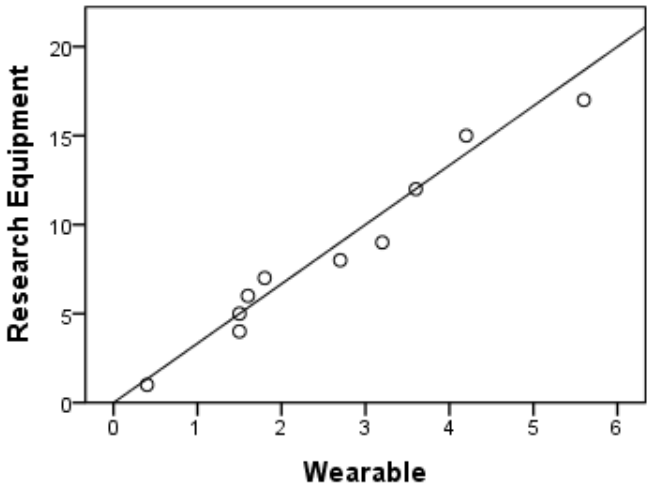


Figure 3. Example data to show the difference between related variables and variables in agreement. The Pearson correlation coefficient for this data set is set to be .98.

Bland and Altman (1986) also pointed out that it would be unlikely that two methods designed to measure the same quantity, would not give an adequate correlation. Other reasons for not using correlations are the dependency on the range of quantity, if the quantity is wide, the correlation will be stronger and a change in scale would change the agreement, but not the correlation (Bland & Altman, 1986). Ludbrook (2002) stated that correlations should be seen as an index of goodness-of-fit of a linear regression model.

The claim that correlations are not an appropriate method for measuring agreement is widely agreed upon (e.g. Lin, 1989; Ludbrook, 2002; Zaki, Bulgiba, Ismail & Ismail, 2012; Schäfer & Vagedes, 2013). Bland and Altman (1986) proposed a method for assessing agreement between devices, similar to the Tukey mean-difference plot (Zaki et al., 2012). This method identifies how much the new method is likely to differ from the old instead of quantifying the actual agreement. Other appropriate calculations are the concordance correlation coefficient (Lin, 1989; Ludbrook, 2002), or the intra class correlation coefficient (de Vet et al., 2006; Zaki et al., 2012), which are similar methods (Atkinson & Nevill, 1998). A least products regression analysis can also be used to assess the agreement between two devices (Ludbrook, 2002).

In the present study there is chosen to use the Bland-Altman method, since the focus is on absolute heart rate values. The other methods take slope into account as well. Ludbrook (2002) wrote in a review about different methods for assessing agreement that the Bland-Altman method is an appropriate statistical method for continuous variables, for example lung function or blood pressure. It was added that a small sample size could contribute to unacceptably wide limits of agreement. Another reason for using the Bland-Altman method is

that this method is used more often than the other, statistically more complicated, calculations (Zaki et al., 2012). This has the effect that the results are easier to interpret when compared to other literature. For example, the concordance correlation coefficient shows poor agreement when a value is < .90 (McBride, 2005). This is confusing, since this is very different compared to the Pearson correlation coefficient. Wang et al. (2016) used the concordance correlation coefficient and found a value of .91 for the Apple Watch, which led to articles on several websites claiming that the Apple Watch had 90% accuracy compared to an ECG (e.g. Sawh, 2016) instead of moderate agreement as stated by the authors. A Bland-Altman analysis was also provided and gave limits of agreement ranging from -27 to +29 bpm, which shows that even with moderate agreement, the true absolute value could be in a range of almost 60 bpm.

The Bland-Altman method is set up to compare two devices by means of setting up limits of agreement. It requires the data of all participants together, so individual differences will be minimalized. For each data point, the difference and the mean of the devices is calculated. A graph is produced, where the differences between the methods are plotted against the mean of the two methods, as can be seen in Figure 4.

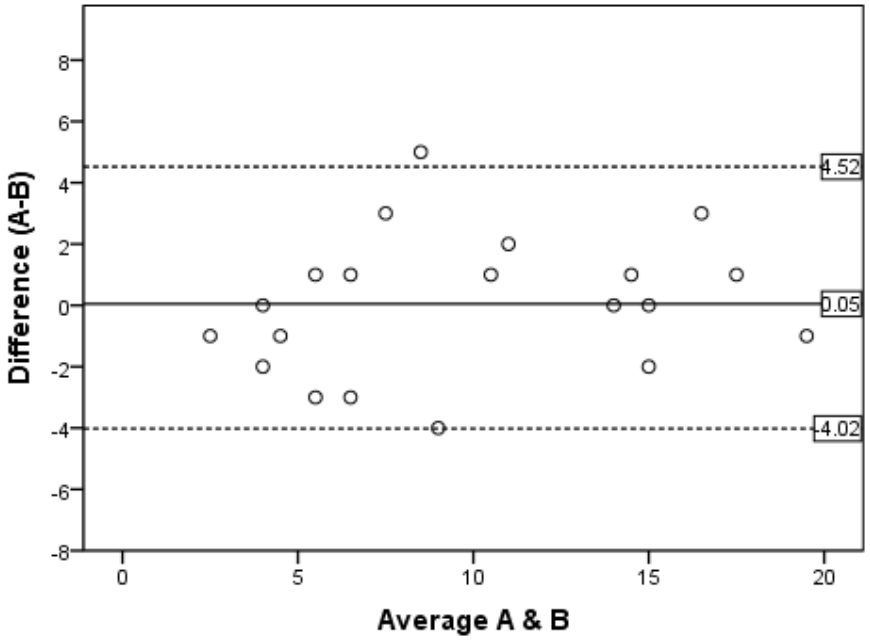


Figure 4. Example of a Bland–Altman plot with simulated data.

By plotting the differences against the mean, the distribution of the differences is made visual. Two horizontal lines are placed at two standard deviation above and below the mean of the differences, the limits of agreement. Since the lines are two standard deviations from the mean, a 95% confidence interval is created. The lines represent the limits of agreement, the

range in which 95% of the values lie. Beforehand maximum allowed differences have to be set up. This means that the limits of agreement coming from the data would have to be smaller than the threshold set up beforehand, to say that the devices are in agreement.

For the present study a ±5 bpm difference was set as the limit in the present study, in accordance with the Association for Advancement of Medical Instrumentation (AAMI) (2002). This means that the limits of agreement derived from the data, the mean ± 2 standard deviations, have to be smaller than ±5 bpm. Calculations were made per device, leading to the following research question:

> Is 95% of the heart rate data from wearables within 5 bpm of heart rate data from traditional research equipment when data is acquired in an ambulatory setting?

Method

Pilot

A pilot study with 17 participants was conducted to familiarize with the testing methods. The Empatica E4 as research wearable and the Biopac as traditional research equipment were used. From the experience of the pilot, there was chosen to use newer electrodes for the Biopac and another form of statistical analysis than correlations. The new electrodes would produce less noise in the ECG signal and correlations turned out to not be a valid measure of agreement between different devices. Further details about the pilot can be found in appendix A.

Participants

Noldus employees working at its headquarters were asked to participate in the study. 21 of them volunteered, of which 16 were male and 5 were female. Their age ranged between 21 and 56, M = 40.10, SD = 10.71. The participants held different positions in the company. The participant carried out their daily work while being measured, which consisted of desk work and meetings. Al participants gave their informed consent and the study was approved by the ethics committee of the University of Twente.

Materials

Physiological measurements. The participants were attached to electrodes for an ECG and wore four different wearables which all measured heart rate. All the wearables made use of PPG to indicate pulse rate. Noldus developed a software program which gathered the physiological information gathered by wearables with a sampling frequency of 1 Hz. The output consisted of a CSV file with a timestamp for all data. The use of this program meant that the raw data could be used, instead of data after unknown filtering as available from the manufacturer.

***Biopac MP150*.** The traditional research equipment used was the Biopac MP150. A wireless system was used with electrodes with wires attached to a small device that can be carried around, the Bionomadix (Biopac, 2014). This wireless device was connected to a recording device, which was stationary. Figure 5 shows the recording system and the wireless device. The electrodes were placed as recommended, the negative electrode on the right collarbone, the positive electrode on the lowest left rib and the ground electrode on the lowest right rib (Appendix B). The Biopac data was exported to Biopac Acqknowledge 4.4 software, from which it could be analyzed and then exported into various file types.

Figure 5. The Biopac system, consisting of recording equipment and a wireless measuring device (Biopac, 2014).

Microsoft band 2. The Microsoft Band 2 is the second smart watch from Microsoft and intended for consumer use. It has different biosensors, such as an optical heart rate sensor, 3-axis accelerometer, an ambient light sensor and a skin temperature sensor. Microsoft Band 2 is intended for everyday use and has sports functions, like a guided workout and a special function for golf. The Microsoft Band 2 is intended to be used with the Microsoft Health app for a smartphone. The Microsoft Band 2 is worn with the PPG sensor on the inside of the wrist.

Garmin Forerunner 235. The Garmin Forerunner 235 is a consumer smartwatch from Garmin. It is a running watch, so it displays heart rate zones and has GPS-functionality. Step count is also available. The Garmin Forerunner 235 is intended to be used with the smart phone app Garmin Connect.

Motorola Moto 360 2nd generation. The Motorola Moto 360 2nd generation is a smartwatch intended for everyday consumer use and looks like a normal watch. It can be adjusted to one's taste and has gender specific sizes. Next to the size, the display background, material of the band and color are customizable. The Motorola Moto 360 is a watch in the series of Android Wear.

Philips Elan prototype. The Philips Elan is a research wearable, still in a developmental state. Since Noldus IT and Philips are working together on a project about smartwatches, the prototype was made available. The Elan has several sensors and gives 22 text files after each session with physiological information. This ranges from heart rate to acceleration to respiration rate. Values were given with a timestamp and an indication of certitude of the value, a scale of quality of measurement. This scale ranged from 0 to 4, with 0 being the worst quality and 4 being the best quality. Philips would not disclose how this scale of quality was determined. Motorola also has a similar value, which gives a lower quality value when more movements are

detected, it is likely that Philips has a similar algorithm. There are two types of prototype, one for real time recording and one for analyzing afterwards. The one used for this project was the one not suitable for real time recording.

The Observer XT coding scheme. The participants all worked at Noldus and were familiar with the Observer XT. The Observer XT is a program used for logging behavior, real time or in a video afterwards. The participants were asked to code their activities during the experiment. The scheme was very simple, as to not disturb the participants in their tasks. The coding scheme can be found in appendix C.

Procedure

Ambulatory assessment was done at the office of Noldus Information Technology BV in Wageningen, the Netherlands. The heart rate of the participants was recorded while they were doing their normal daily tasks. The measuring took place on working days over a time period of 2.5 weeks. Each morning or afternoon was reserved for one participant. The experiment time varied, depending on the availability of the participant, but was on average 2 hours, with a mean of 123.39 minutes and a standard deviation of 40.41 minutes.

The participants were asked to put the wearables on their wrists and attach the electrodes to their upper body. The placement of the electrodes was tested by checking the ECG signal. The wearables were turned on and it was checked if the computer program received the data. The Elan had a light turning on every second to indicate recording. Participants wore two wearables on each arm, as can be seen in Figure 6. The placement varied per participant, to account for the influence of position. Appendix D shows the position of the wearables per participant. After checking if the devices worked as they are supposed to, the participants could go on with their work as usual. The researcher would check on the devices about every 30 minutes, to see if the connection between the devices and the computer was not lost. Sometimes the Bluetooth connection would fail, making the checks necessary.

Figure 6. A participant with the wearables.

Data acquisition

Biopac data. The raw data from the Biopac was prepared for analysis using Acqknowledge, a software program for physiological analysis from Biopac. The sampling frequency was set to 250 Hz to establish an ECG signal. All files were manually checked for noise, such as sudden movements, which influenced the ECG signal, and therefore the bpm. Markers were set to distinguish the disrupted data, to be able to discard this later on. Checking for noise was done manually, because when using an automatic filter, the data was not discarded, but smoothed. This means that it could not serve as ground truth. How exactly the noise was identified is described in appendix E. The bpm was calculated from the ECG with automatic settings for a human heart beat in a resting situation. The bpm was exported to an Excel file with one value per second. Since the data was recorded on a 250 Hz frequency, the last value given in a certain second was exported. The file with markers for noise was separately exported. A program was coded to use the markers to discard the values between these markers in the file with bpm values.

Garmin, Motorola & Microsoft Band. The data from these wearables was collected with a smartphone that had a Bluetooth connection to the wearable and a network connection to a computer. The computer program exported the data in Excel files, with values and a timestamp. No data from the wearables was altered or discarded, since no objective noise could be determined, due to the only available output being heart rate values and timestamps.

Philips. The wearables from Philips gave several txt files, from which only the heart rate file was used. This file gave the heart rate value in bpm, a timestamp and a value for the quality of measurement.

Data structuring

For each participant an Excel file was created with the collected data from all devices. This was done by combining all the different files as supplied by the different wearables, resulting in one file per participant. Each device had a column, every second a row. The next sheet of these files showed only the data between the recorded start and stop time, as indicated by the Observer. All values under 40 were removed, since it can be assumed that these values were artefacts. The values of the first minute from the wearables were discarded, to be sure that the devices were well attached and adjusted. A new file was made with the data of all participants together. In a new sheet, the data was averaged over 1 minute. This was done by using an averaging formula with the condition that at least 45 of the 60 values were present.

Data analysis

The different devices were compared with regard to agreement and quality of the data, such as missing data. The percentage of missing values was compared between the devices. The assessment of the agreement was done by using the Bland-Altman method (Bland & Altman, 1986). This method consists of plots of the differences between the devices against their mean. Three lines are plotted, one for the mean of the differences and two standard deviations above and under the mean. These two lines are the limits of agreement and refer to the interval in which 95% of the values lie. Beforehand a threshold is set, in this case ±5 bpm, in accordance with AAMI (2002). This threshold should be broader than the limits of agreement coming from the data in order to say that the devices are in agreement with each other.

The Bland-Altman method works best with normally distributed differences. When this is not the case and the differences are skewed, the data can be log transformed. When after the log transformation, the differences are still not normally distributed, the Bland-Altman method can still be used. This will not lead to acceptance of poor devices, since the limits of agreement tend to be broader in this case (Bland & Altman, 1986). Furthermore the percentages of the values outside the threshold were given.

For the quality of measurement from the Philips, a general linear model with repeated measures calculation was also used. This was done to identify the linear relationship, with the mean of differences as the dependent variable and the scale of quality of measurement as predictor. Paired sample t-tests with the mean difference per scale point were done to check if the differences between the Biopac and the Philips were significantly different on each scale point.

To check if the position of the wearables has influence on the differences between the wearables and the traditional research equipment, a paired samples t-test was done, with the mean difference per position. The pairs existed of all the possible combinations of the positions, since a linear relationship would be unlikely. The mean and standard deviation of the differences per position was given.

Results

Missing data

Almost all devices had missing data, due to different reasons. The percentage of missing data is shown per device and participant in Table 1.

Table 1.

Missing data per device, per participant in percentages. Empty cells represents no collected data from a certain device. Participant 1 has two rows, since there were two measurements.

Participant	Biopac	Microsoft	Motorola	Garmin	Philips
All data	9.99	14.32	20.01	97.04	3.77
1	25.16	1.45	18.42	99.85	
	11.65	1.32	43.46	96.65	13.76
2	9.02	44.40	45.31		0
3	59.00	12.78	37.77	94.73	0
4	11.50	5.13	50.05	98.54	13.67
5	2.17	0.70	4.68	94.92	0
6	4.91	23.92	56.54	97.96	0
7	8.78	27.72	4.00	99.86	
8	1.74				0
9	7.60				0
10	3.29	39.47	16.85	99.77	31.98
11	22.77	6.96	26.42	96.02	0
12	2.64	27.58	31.95	96.60	0
13	13.32	7.85	7.01	99.32	0
14	13.79	21.52	4.10		0
15	1.34	0.22	9.46	87.73	0
16	3.16	15.31	4.46	97.08	0
17	0.61	4.88	6.64	96.31	15.98
18	2.04	15.77	5.32	97.72	0
19	5.37	15.80	5.55	98.56	0
20	5.17	5.25	11.56	97.41	0
21	4.83	8.54	10.60	97.70	0

As can be seen, the missing data from the Garmin was exceptionally high. This meant that the Garmin data was not used in further calculations. It is likely that this was caused by the fact that the Garmin is not intended for real time Bluetooth communication. The missing data from the Biopac was in most cases due to the participant being too far away from the Biopac,

so the data could not be sent to the computer. Other reasons for missing data by the Biopac were movements when the electrodes were touched, resulting in noise in the ECG signal. The missing data from the Microsoft Band and Motorola was also due to the Bluetooth. The devices were not always able to send data each second, and the connection was sometimes disconnected. The missing data from Philips was due to an empty battery.

Microsoft Band 2 and Motorola synchronization

The synchronizing from the Microsoft Band and the Motorola with the Biopac did not go as intended. After the measurements it turned out that the program used for the communication to and from to wearables, applied the timestamp from the phone instead of the wearable to each value. Since the wearables were synchronized to the computer, but the phones were not, no certain conclusions can be drawn from these devices. The transmission of the information from the wearable to the phone via Bluetooth took on average 5 milliseconds, with rare outliers up to 0.3 seconds. For a measuring frequency of 1 Hz, this is negligible, since even if someone's heart rate would be 180 beats per minute, this would mean 1 missed beat at most due to the delay in sending. This is important, since the phone gave the time stamp at the moment of receiving, so the exact timestamp of the moment of measuring was lost. The phones and the computer were all set to network time, meaning that the displayed time should be the same. This does not mean that the time was the same, it was not verified during the measurements. After this came to light, which was after the data collection, the phone times were occasionally compared to the computer time and they sometimes differed up to 7 seconds. This means that even though all the devices were set to network time, there were differences. It could be that the devices only setting the time when turned on, and after that an internal clock is used in which the duration of an hour is slightly different than the actual duration of an hour. No set-off point in the data could be found by comparing peaks and slopes in the Biopac data and wearable data. All these factors led to the conclusion that the data as collected could not be used, since the synchronization had failed and it could not be verified by how much. This means that it was not possible to analyze the data on a 1 second basis, but on a 1 minute basis was possible as will be explained below.

Averaged data

Even if the data was not accurate enough to be analyzed per second, it was possible to use when averaged over 1 minute. It was possible to compare the data when averaged, because the synchronization was off by a maximum of a few seconds. This is negligible over 1 minute, since

heart rate values do not differ much per second. For the best comparison, the data from the Philips was also averaged over 1 minute. For each device a Bland-Altman plot was made, as can be seen in Figures 7, 8 and 9. The data was normally distributed, as shown in appendix F.

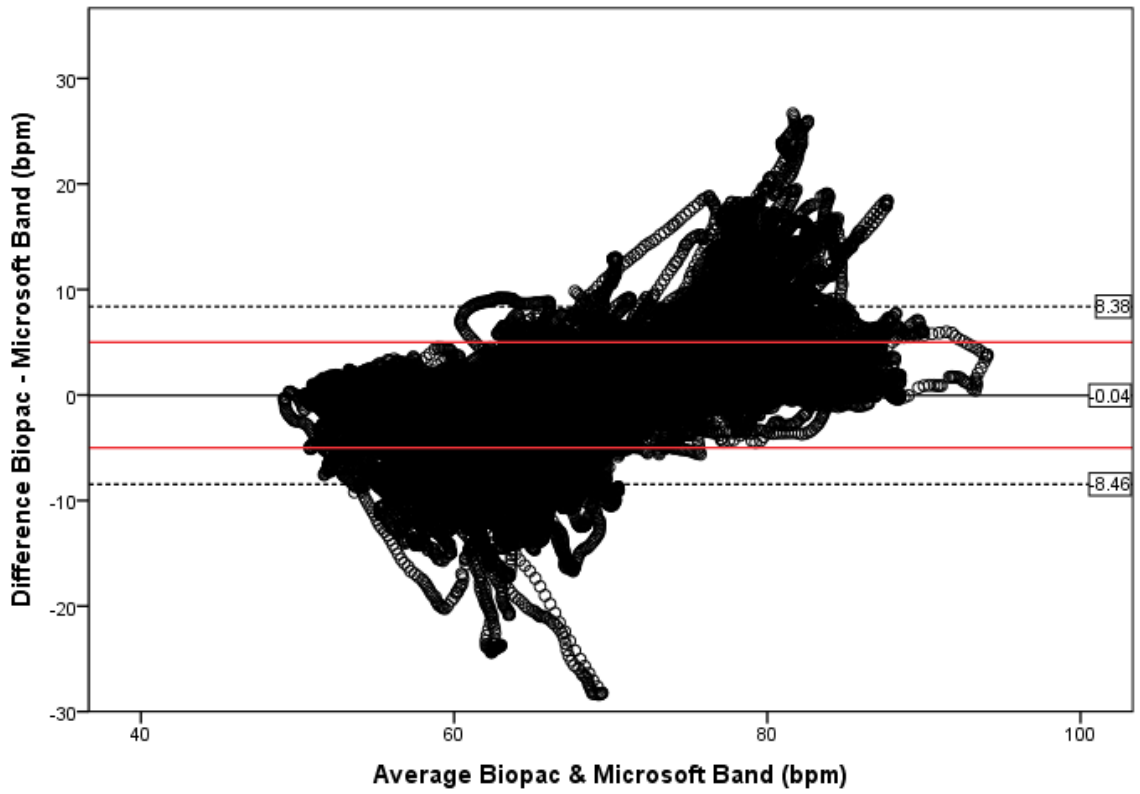


Figure 7. Bland–Altman plot of Biopac and Microsoft Band data, averaged over 1 minute. The black line represents the mean of the differences, and the dotted lines are the limits of agreement, two standard deviation from the mean. The red lines indicate the threshold of ±5 bpm.

The limits of agreement from the Biopac and Microsoft Band data ranged from -8.46 bpm, 95% CI [-8.50, -8.42] to 8.38 bpm, 95% CI [8.34, 8.42]. A total 16.81% of the values lay outside of the threshold of ±5 bpm.

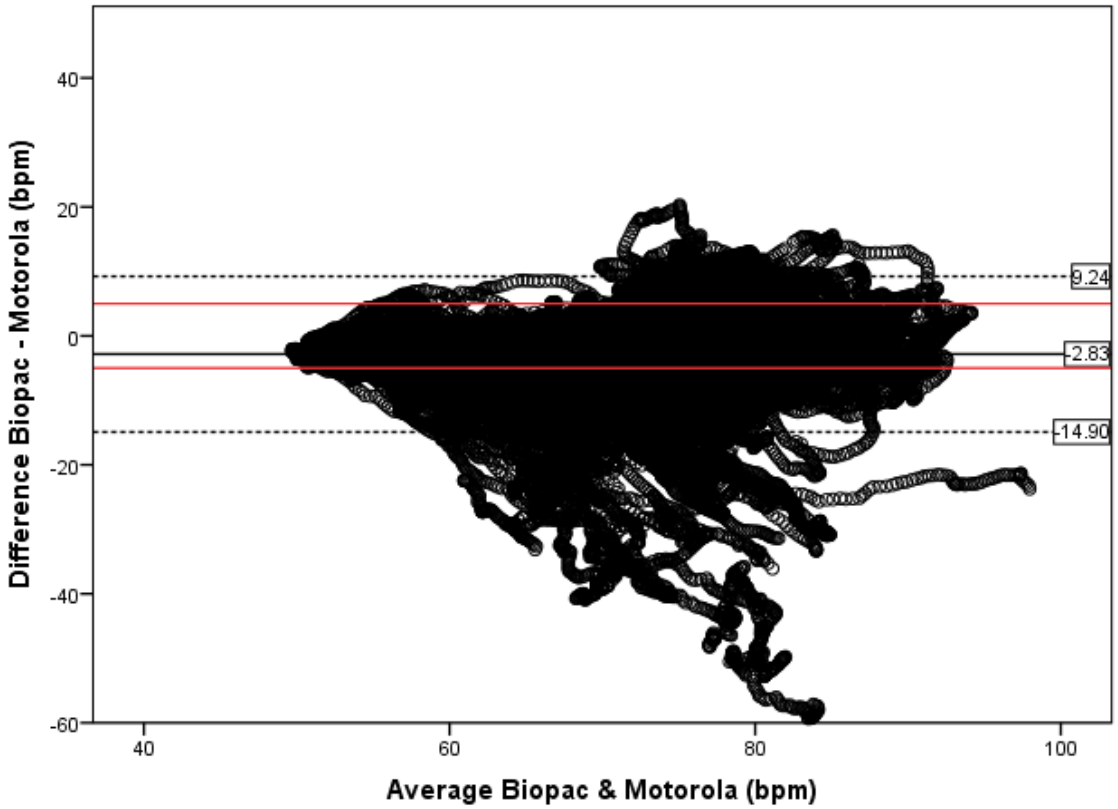


Figure 8. Bland–Altman plot of Biopac and Motorola data, averaged over 1 minute. The black line represents the mean of the differences, and the dotted lines are the limits of agreement, two standard deviation from the mean. The red lines indicate the threshold of ±5 bpm. The red lines indicate the threshold of ±5 bpm.

The limits of agreement from the Biopac and Motorola data ranged from -14.90 bpm, 95% CI [-14.96, -14.84] to 9.24 bpm, 95% CI [9.18, 9.30]. In total 22.14% of the values lay outside of the threshold of ±5 bpm.

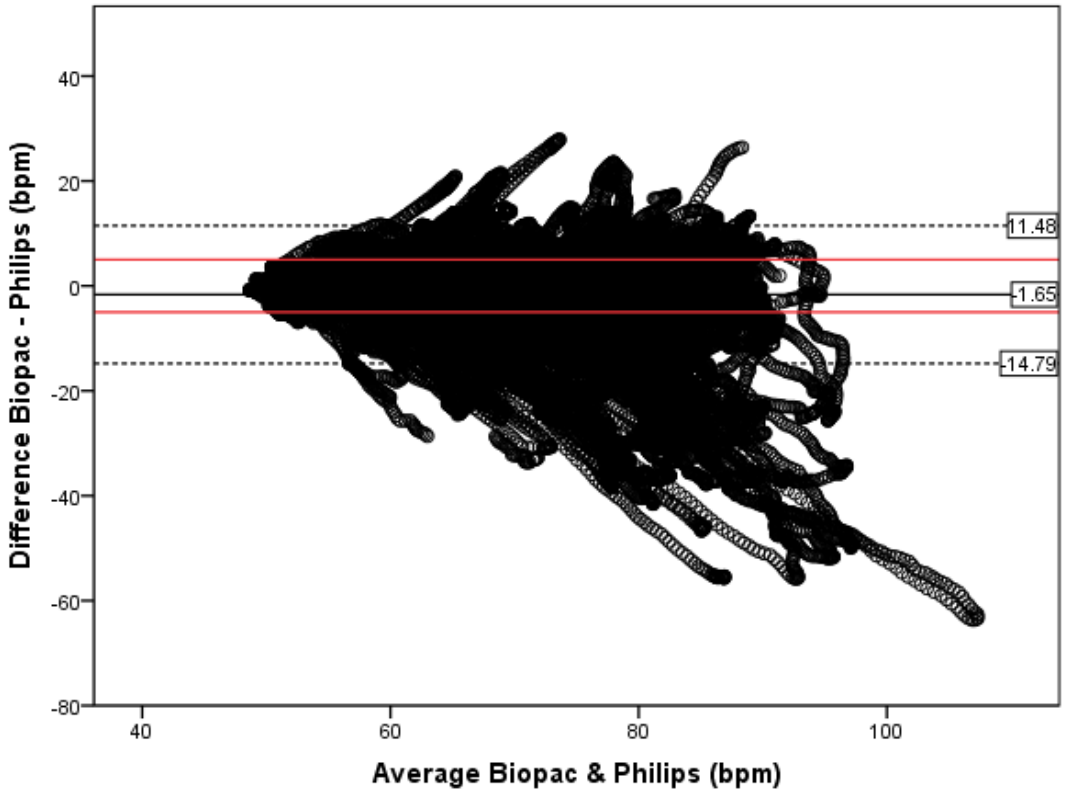


Figure 9. Bland-Altman plot of Biopac and Philips data, averaged over 1 minute. The black line represents the mean of the differences, and the dotted lines are the limits of agreement, two standard deviation from the mean. The red lines indicate the threshold of ±5 bpm. The red lines indicate the threshold of ±5 bpm.

The limits of agreement from the Biopac and Philips data ranged from -14.79 bpm, 95% CI [-14.85, -14.72] to 11.48 bpm, 95% CI [11.42, 11.55]. In total 17.01% of the values lay outside of the threshold of ±5 bpm.

As can be seen in Figure 7, the differences between the Biopac and the Microsoft Band ranged from negative differences with lower heart rate values to positive difference at higher values. This means that when the Biopac gave low heart rate values, the Microsoft Band gave even lower value and when the Biopac gave a high value, the Microsoft Band gave even higher values. The difference between the Biopac and the Microsoft Band was just as big with high and low values.

Another pattern occured with the Motorola and the Philips, as can be seen in Figure 8 and 9. When the Biopac gave higher values, the wearable gave lower values, whereas with lower values of the Biopac, the values of the wearable varied less.

Philips

The Philips data was also analyzed on a 1 second basis.

Bland-Altman. A Bland-Altman method was used to obtain the limits of agreement. The differences were normally distributed (appendix G). These limits ranged from -20.88, 95% CI [-20.91, -20.73] to 17.60 bpm, 95% CI [17.45, 17.63]. In total 44.27% of the values lay outside of the threshold of ±5 bpm. Figure 10 shows the Bland-Altman plot, displaying the difference between the Biopac and the Philip against the mean of the two. This visualizes how the differences were spread; when the heart rate was lower, the differences were smaller.

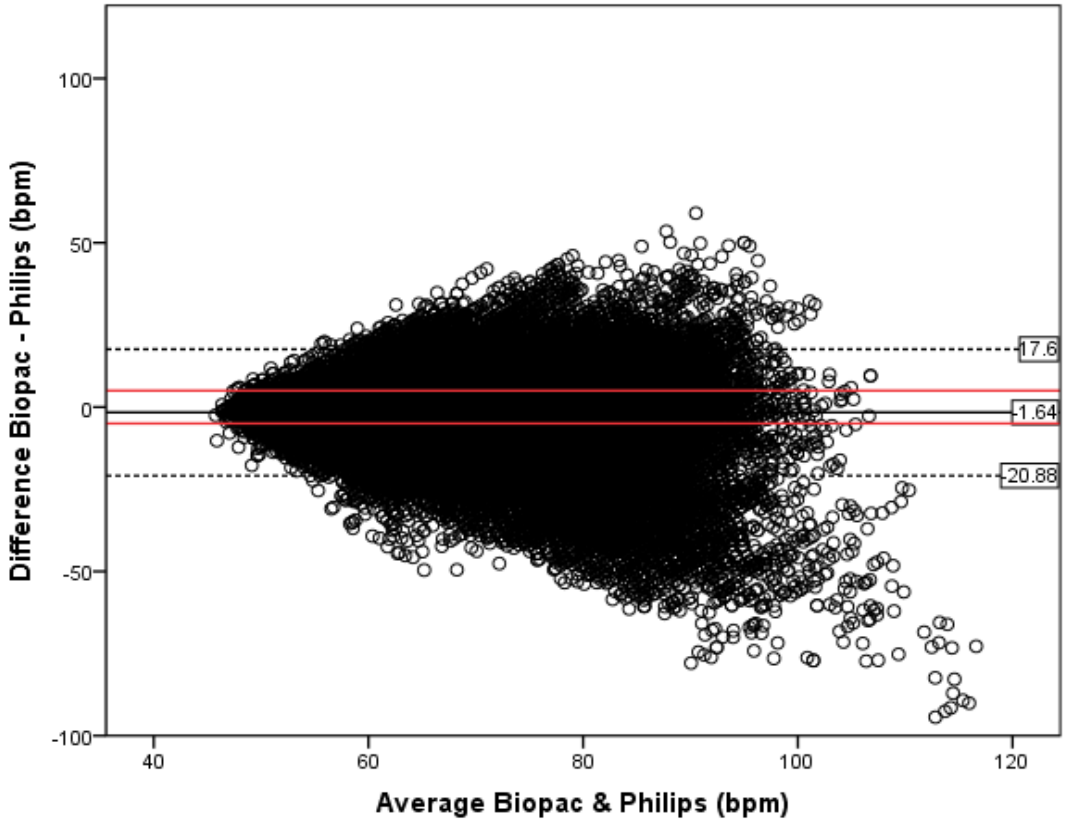


Figure 10. Bland-Altman plot of Biopac and Philips. The black line represents the mean of the differences, and the dotted lines are the limits of agreement, two standard deviation from the mean. The red lines indicate the threshold of ±5 bpm. The red lines indicate the threshold of ±5 bpm.

Differences between participants. The data was plotted per participant. These were checked for normal distributions, which can be found in Appendix H. Some of the participants (participant 9, 10, 12, 14 and 20) did not have normally distributed differences, the data from these participants was log transformed. None of the log transformed data was normally distributed, as displayed in Appendix I, so there was chosen to use the original data instead, in accordance with Bland and Altman (1986).

There were several patterns visible in the data per participant. Figure 11 shows some typical participants. It was clear that the agreement between the devices was better for some

participants than for others. Some participants had smaller limits of agreement in which the differences were divided proportionally, such as with participant 11. For other participants there was no clear relation between the devices, and the limits were broad (participant 2). Other patterns were when the Philips gave different values, while the Biopac gave more or less stable values (participant 9), or when the Philips gave low values, the Biopac gave varying high and low values (participant 6). The plots from all participants can be found in appendix J. The other wearables also displayed different patterns per participant.

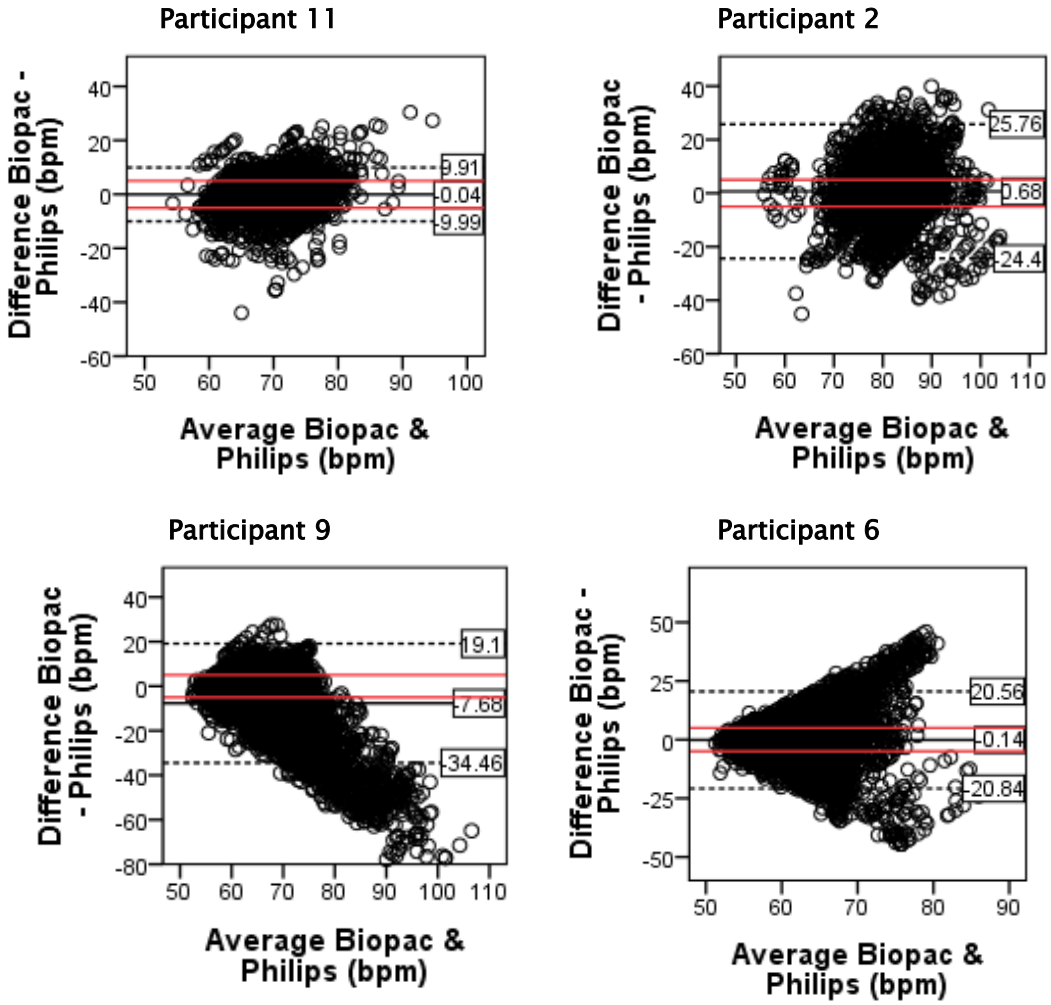


Figure 11. Bland-Altman plots showing the different patterns for agreement between the devices, each plot shows the data from one participant. The black line represents the mean of the differences, and the dotted lines are the limits of agreement, two standard deviation from the mean. The red lines indicate the threshold of ±5 bpm. The red lines indicate the threshold of ±5 bpm.

Effect of quality. The Philips had a scale ranging from 0 to 4 for the quality of the measurement for each value. Descriptive statistics of the differences can be found in Table 2. The relation of the differences between the Biopac and the Philips and the scale of quality of measurement was

linear, F (4, 14) = 20.16, p = 0.00. All the scales differed significantly (p < .02), except for the quality scales 0 and 1 (p = .09).

Table 2.

Mean and standard deviation of the differences between Biopac and Philips, arranged by quality of the measurement.

Position	Mean	Standard deviation
All data	-1.64	9.62
Quality =0	-15.12	19.39
Quality = 1	-11.08	15.10
Quality =2	-5.99	12.50
Quality = 3	-2.39	9.47
Quality = 4	.09	7.11

For the each of the quality scales a Bland-Altman analysis was done (appendix K). When the quality of measurement was higher, the limits of agreement were narrower, leading up to -14.21, 95% CI [-14.21, -14.06] to 14.31, 95% CI [14.24, 14.39] for the highest quality of measurement.

Effect of placement

Table 3 shows the descriptive statistics of the positions, but the values per position did not differ significantly (ranging from p = .07 to p =.88). This means that there were no significant differences in the positions of the wearables.

Table 3.

Mean and standard deviation of the differences between data from the wearables and the data from traditional research equipment arranged by different positions of the wearables.

Position	Mean	Standard deviation
All data	-1.29	3.27
Left wrist (closest to hand)	-.32	3.02
Left wrist (second)	.05	1.76
Right wrist (closest to hand)	-3.43	4.19
Right wrist (second)	-1.26	2.44

Discussion

Conclusions from the data

Limits of agreement. To be able to answer the research question, the limits of agreement found in the data between the Biopac and the wearables were compared to the threshold. On a 1 Hz sample rate, only the Philips Elan was sufficiently tested and the results were not within the acceptable limits. The limits of agreement from all wearables are too wide, even when averaged over 1 minute. This means that the research question, 'is the 95% of the heart rate data from wearables within 5 bpm of heart rate data from traditional research equipment in an ambulatory setting?' can be clearly answered with a 'no'. None of the tested wearables constantly gave values within 5 bpm of the Biopac when averaged over 1 minute. This means that when analyzed on a 1 second basis, the wearables would also not give values within ±5 bpm from the Biopac.

Quality. Quality in this setting means the foundation of the assumptions of the research. In this case, the quality was secured by two aspects, the placement of the wearables and the quality indication of the Philips. The placement of the wearables differed per participant, to adjust for potential placement effects. There were no significant differences between the positions of the wearables, meaning that the placement of the wearables did not have effect on the measurements. The Philips Elan gave an indication of quality of measurement with each heart rate value. The data of higher quality of measurement had smaller limits of agreement than the data of the lower quality of measurement. The agreement improved when the quality gets better, which implies that the measurement used as ground truth is a constant good value. If the quality measure was not actually indicating the quality no relation between the level of agreement and the quality measure would be expected.

Missing data. The Motorola had most missing data, followed by the Microsoft Band, the Biopac and then the Philips. The Motorola had double the amount of missing data compared to the Biopac, while the Microsoft Band lay in-between, but close to the percentage of the Biopac. Even though the wearables did not give significantly less missing values, this does not disprove the assumption that participants can be more mobile when using a wearable than with traditional research equipment without data loss (Wac & Tsiourti, 2014).

The missing data of the Biopac was due to the participant being too far from the Biopac or movement that caused disruptions in the ECG measurements. The Philips only had minimal missing data, which was all due to battery loss. The missing values of the commercial wearables

were mostly due to the Bluetooth connection, making it likely that this would be less when it was collected in another manner and if the data would be analyzed afterwards.

Implications

Other validation studies. The findings of the present study were in line with the results of Wang et al. (2016), who concluded that for accurate heart rate measurement, it is better to use a device that used electrodes on the chest. Stahl et al. (2016) and Spierer et al. (2015) claimed certain wearables to be accurate. Both stated that all wearables tested were in more or less of an agreement with research equipment. This conflicts with the present study, in which none of the tested wearables were deemed accurate enough. This could be explained by the different manners of data analysis. As de Vet et al. (2006) stated, only using reliability measurements without agreement measurements, often leads to wrong conclusions. Stahl et al. (2016) used the mean absolute percentage error (MAPE). Bland-Altman plots were used to illustrate the slope of the differences, but no threshold was stated. Their limits of agreement ranged from -9.7 bpm to 9.7 bpm for TomTom to -17.6 bpm to 25.8 bpm for Fitbit. Stahl et al. (2016) still deemed the wearables accurate, due to the MAPE calculations and a Tukey's range test. Spierer et al. (2015) used correlations and t-tests. None of these calculations are considered to be appropriate statistical methods for agreement between measurement devices (e.g. Ludbrook, 2002; de Vet et al., 2006; Köttner et al., 2011; Zaki et al., 2012). The MAPE is not appropriate since it has a bias towards the device which gives the lowest values (Tofallis, 2015). T-tests and Tukey's range test compare the mean of the different methods, not the values per second. Correlations give a measure of association, not of agreement (Bland & Altman, 1986).

The results of the Bland-Altman analysis of the present study were similar to the results of Stahl et al. (2016), but led to different conclusions. This could be due to different thresholds. If judged by the Bland-Altman analysis with the same threshold as the present study, the wearables from Stahl et al. (2016) would also be classified as non-accurate. This means that, none of the wearables are able to measure heart rate to an accuracy of ±5 bpm. This is assuming that with the different devices tested by Stahl et al. (2016) and the present study, a broad selection of recent wearables was included. Because they use similar PPG technology, this makes it likely that other wearables using similar technology and in the same price-class would also fail to give results within the limits of ±5 bpm.

Other validation studies of heart rate measurement used other thresholds for the Bland-Altman analysis, for example ±6.5 bpm (Kornowski et al., 2003), ±11 bpm (Gatti, Scheider & Migliaccio, 2014) and ±0.6 bpm (Radespiel-Tröger, Rauh, Mahlke & Mück-Weymann, 2003).

Gatti et al. (2014) made a statement that their threshold was based on medicine and sports science. The other articles gave no arguments for their chosen limits. The limits between the studies differed a lot and lacked justification for the set limits, so another approach was chosen for the present study. The limits should not be too broad, since the absolute value is important for research purposes. With an acceptable limit of ±11, the real value could be within a range of 22 bpm. Limits that are too narrow will result in devices being rejected when this might not be necessary. A limit of ±0.6 is unfeasible with wearables with an output of rounded numbers. ±5 bpm seemed not too broad and not too narrow, meaning that this was deemed a good limit in which there was a good trade-off of the chance of wearables being falsely accepted or rejected.

Different use cases call for different acceptable limits. In consumer use, the acceptable limits could be broader, since an approximate value is enough to be able to determine for example a heart rate zone during running and users might be more interested in the slope than the absolute value (Tholander & Nylander, 2015). With the results of the present study no conclusions can be drawn about whether wearables could be used when the change of heart rate is more important than the absolute heart rate value.

Research implications. The present study was done in an ambulatory setting, with wearables aimed at the consumer market and the focus on absolute values. That means that the conclusions cannot be generalized outside these settings. For studies in which absolute heart rate values are important, this means that at least the tested wearables are not a viable alternative for traditional research equipment. As stated above, it is likely that the current generation of wearables is not suitable for research in regard to absolute heart rate values. This means that studies focusing on absolute heart rate using these types of wearables might have come to different conclusions when traditional research equipment was used. For other types of research, when the absolute heart rate data is not analyzed, consumer wearables could still be suitable. Examples are a study of user comfort or usability of wearables or a study designing the visualization of real time physiological data. It would also be possible to use heart rate measurement from wearables in a study comparing the base line, when a subject is at rest, with elevated levels as an indicator of exercise. In such a study there would not be made use of absolute heart rate values, but the elevation of the heart rate, together with data from an accelerometer for example, could be used as an indication of the exercise.

With the results of the present study, there can be no conclusions about the accuracy of the wearables on the market developed for research, such as the Empatica E4. It is possible that these type of wearable have better technology, leading to values more in agreement with

traditional research equipment. The present study also does not provide conclusions about wearable use for consumers not focusing on absolute heart rate values.

Consumer implications. With the present study there are no conclusions about the use of wearables for consumers, since it is not clear to which extent consumer rely on absolute heart rate. If a wearable is used for activity tracking, the training might be adjusted to heart rate or heart rate zones (Tholander & Nylander, 2015), but there was no information available on how heavily a consumer relies on absolute heart rate values.

When looking to buy a wearable, there is much information available on independent review websites (e.g. Wareable (www.wareable.com)). As stated before, these kind of websites often test wearables on their accuracy without specifying what that means. Since most of these tests are done on one participant, this might not be generalizable to a broader population. The present study revealed big differences between participants, so depending on the person who tested the wearables, the conclusion could vary. If the person who tested the wearables found good agreement between the devices, it would not necessarily mean that this would work for prospective buyers too. The wearable would be promoted as accurate, while this may not be the case for many other wearers. The same could happen the other way around, if the person testing the devices finds low agreement between devices, for example due to skin color (Schäfer & Vagedes, 2013), the wearable would be classified as inaccurate, while for others it may work well. Even though some websites warn that this might happen, not all consumers might be aware of this. All in all, users of wearables should be cautious in relying only on a wearable for their heart rate. This is important, since one of the potential dangers of wearable use is over-relying on the given data (van Dijk et al., 2015).

Recommendations

Limitations of the study. The present study could have been improved by better synchronization. The values of two of the wearables could only be analyzed on a 1 minute basis. This was the major flaw in carrying out the present study, since it was planned to analyze all the wearables on a 1 second basis. In this case, it did not influence the possibility to answer the research question, since the wearables did not meet the standards on a 1 minute basis and thus it can be concluded that the wearables would not meet the standards on a 1 second basis. For the sake of methodology it would have been better to be able to do analysis on a 1 second basis, since that is the frequency the wearables will most likely be used at.

Another shortcoming of the present study was that no use was made of the coded behavior. This was due to unforeseen errors in the gathering of the logs, making it unreadable

in the Observer. This information could potentially have explained some of the variation and missing data.

Not all wearables gathered the data in the same way in the present study. This was due to the Philips software for real time measuring not working in the first weeks of the experiment. By the time everything worked as it was supposed to, it was not feasible to implement the real time recording device into the experiment. The Biopac, Microsoft, Garmin and Motorola devices all measured real time, while the Elan saved all the data within the device, to be exported afterwards. This is one of the reasons the Philips was the device with the least missing data. Furthermore, there was no replication of the devices, meaning that one device was used without establishing that this one device was representative for all the devices in the series (Biopac MP 150, Microsoft Band 2, Garmin Forerunner 235, Motorola Moto 360 2nd gen, Philips Elan).

Due to the use of the program for gathering the data for the consumer wearables, there was made use of the raw data of the wearables. This means that when the data was gathered the way the manufacturers intended, more filtering would be applied and there would have been less outliers. This could explain some of the discrepancies between the results of the present study and the manufacturer's claims about the agreement of the wearable with traditional research equipment. There was chosen to gather the data this way to fit the use case, since researchers require access to the raw data instead of data that has been transformed in an unknown way. Other reasons are that when using this program, the data would not be stored at the manufacturer and some use cases required real time access to the data.

It would have been better to have used another statistical analysis next to the Bland-Altman method, but there was not sufficient time available to carry out further data analysis. By carrying out alternative calculations, such as the least product regression analysis or the concordance correlation coefficient, the slope would have been takin into account next to the absolute values. The use of a concordance correlation coefficient would lead to a conclusion if use of wearables is feasible in situations where change in values is generally more important than absolute values. This is important, since not only for consumer use, but also for research purposes, the change in heart rate might be as important as the absolute heart rate. This is for example the case in a study measuring stress. Since the present study was focused on absolute heart rate values, the use of another method would be of added value, but not necessary to answer the research question.

Recommendations. For future validation research it is recommended to use the same way of measuring for optimal comparison, meaning that all wearables measure real time, or all

wearables do not. This will reduce the amount of different factors between the wearables that could have influence on the measurement. If the choice for real time measurement is made, it is important to use a tool appropriate for synchronization and monitor the synchronization frequently, to prevent later issues. When not measuring real time, it is as important to have correct synchronization, since the data will be collected independent of the other devices. The choice for either of these measurement methods, should depend on the goal of the research. If the end goal for the use of wearables calls for information real time, such as an adaptive training, it is best to test the wearables real time.

Before starting a study, it would be good to do a pilot test to see if the chosen wearables are suitable for real time recording by testing the Bluetooth connection. Another recommendation is to follow the guidelines for reporting reliability and agreement studies from Köttner et al. (2011), which provides guidelines for each part of a research paper. It is also important to use appropriate statistical analysis, such as the Bland-Altman method or the least products regression analysis (Ludbrook, 2002). A concordance correlation coefficient is also an appropriate statistical method for evaluating reproducibility (Lin, 1989, Schäfer & Vagedes, 2013). By combining multiple statistical calculations, agreement and reliability parameters can be assessed (de Vet et al., 2006), leading to a more complete image of the reproducibility of the device.

It would be good to further research the different personal characteristics influencing accurate pulse rate measurement. The agreement between the wearables and the traditional research equipment was different per participant. No statistical analyses about personal characteristics, such as skin color could be done, since the subgroups would have been too small. The fact is that there were differences in the agreement per participant. For some participants the wearables were in more agreement with the traditional research equipment than with others. In future research it would be good to test a larger heterogeneous group participants and analyze the effects of characteristics such as skin color, age, hairiness and fat percentage. The differences between the devices should also be investigated further, since for different participants different wearables were more in agreement with the traditional research equipment.

In general when doing research using wearables, there might be ethical concerns which are not as important when using traditional research equipment. The data gathered with wearables is not always stored in a secured place, mostly the manufacturer's website, making it ambiguous who has access to the data (Ryan, 2016). This is an important consideration when selecting a wearable for independent research. When using a commercial device in research,

the data might be accessible to third parties too. In case of research which makes use of imposed or exploited self-tracking, the researcher might be the third party using the data without the users knowing. This is a sensitive issue, due to privacy of the user on one hand and the opportunities of big data on the other hand. Manufacturers of wearables should be transparent with their users about who has access to their data and why (Fort et al., 2016). When users are aware why and how their data is used, they might be more content with sharing their data. The availability of big data could give insight in all kinds of behavior and habits. This could contribute to the improvement of products for the users, for example when Strava's heat maps are used to improve the most used bicycle tracks. Research in big data could also contribute to general knowledge about habits, for example when Jawbone examined sleeping patterns all over the world and found that people sleep on average least in Tokyo and most in Melbourne (Wilt, 2014). All in all, the implication of availability of big data is twofold, on one hand there is no clear statement on the privacy for users (Fort et al., 2016), but on the other hand big data could give interesting insights.

Another recommendation in general is to not use lower cost wearables to measure absolute heart rate in research until they are sufficiently validated. Until then it is better to use traditional research equipment, which is also available in wireless options, or as an alternative a validated chest strap, for example the Polar H7 (Polar, 2012). This may not be the most ideal option, but does meet some of the advantages of wearables, such as participant mobility. To be able to use absolute heart rate data from wearables in research, further development of the pulse rate technology is needed, as well as validation studies with correct statistical analyses.

Conclusions

Wearable technology is not yet as far developed as one might wish. The heart rate measurements taken from consumer wearables do not fulfill the guidelines for heart rate accuracy as stated by the AAMI (2002). At the moment, it is not feasible to use heart rate measurements from those wearables in research when the use of absolute heart rate values is important.

References

Acharya, R., Kannathal, N., Sing, O.W., Ping, L.Y., & Chua, T.L. (2004). Heart rate analysis in normal subjects of various age groups. *BioMedical Engineering OnLine, 3*, 24-31. doi:10.1186/1475-925X-3-24

Allen, J., (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement, 28,* (3), 1-39. doi:10.1088/0967-3334/28/3/R01

American Heart Association (2015). Target heart rates. Retrieved from http://www.heart.org/HEARTORG/HealthyLiving/PhysicalActivity/FitnessBasics/Target-Heart-Rates_UCM_434341_Article.jsp#.V-J2RyiLQdU

Association for the Advancement of Medical Instrumentation (2002). Cardiac monitors, heart rate meters, and alarms. Retrieved from http://www.pauljbennett.com/pbennett/work/ec13/ec13.pdf

Atkinson, G., & Nevill, A.M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine, 26,* (4), 217-238. doi: 10.2165/00007256-199826040-00002

Berntson, G.G., Quigley, K.S., & Lozano, D. (2007). Cardiovascular psychophysiology. In J.T Cacioppo, L.G. Tassinary & G.G. Berntson (Eds.), *Handbook of Psychophysiology* (159-181) New York, NY: Cambridge University Press.

Biopac (2014). Bionomadix: Physiology where, when, and how you want it. Retrieved from https://www.biopac.com/product/bionomadix-2ch-wireless-emg-amplifier/

Bland, J.M., & Altman D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, 327,* (8476), 307-310. doi:10.1016/S0140-6736(86)90837-8

Chuah, H.W.S., Rauschnabel, P.A., Krey, N., Nguyen, B., Ramayah, T., & Lade, S. (2016). Wearable technologies: The role of usefulness and visibility in smartwatch adoption. *Computers in Human Behavior, 65,* 276–284. doi:10.1016/j.chb.2016.07.047

de Vet, H.C.W., Terwee, C.B., Knol, D.L., & Bouter, L.M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology 59,* (10), 1033–1039. doi:10.1016/j.jclinepi.2005.10.015

Fahrenberg, J., Myrtek, M., Pawlik, K., & Perrez, M. (2007). Ambulatory assessment – monitoring behavior in daily life settings. A behavioral-scientific challenge for

psychology. *European Journal of Psychological Assessment, 23, (4),* 206-213. doi:10.1027/1015-5759.23.4.206

Ferguson, T., Rowlands, A.V., Olds, T., & Maher, C. (2015). The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: A cross-sectional study. *International Journal of Behavioral Nutrition and Physical Activity, 12,* (42). doi:10.1186/s12966-015-0201-9

Fort, T.L., Raymond, A.H., & Shackelford, S.J. (2016). The angel on your shoulder: Prompting employees to do the right thing through the use of wearables. *Northwestern Journal of Technology and Intellectual Property, 14,* (2), 139-170. Retrieved from http://scholarlycommons.law.northwestern.edu/njtip/vol14/iss2/1/

Gatti, U.C., Scheider, S., & Migliaccio, G.C. (2014). Physiological condition monitoring of construction workers. *Automation in Construction, 44,* 227-233. doi: 10.1016/j.autcon.2014.04.013

Kalat, J.W. (2007). *Biological Psychology.* Belmont, CA: Wadsworth.

Kornowski, R., Zlochiver, S., Botzer, L., Tirosh, R., Abboud, S., & Misan, S. (2003). Validation of vital signs recorded via a new telecare system. *Journal of Telemedicine and Telecare, 9,* (6), 328-333. doi: 10.1258/135763303771005225

Köttner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B.J., Hróbjartsson, A., … Streiner, D.L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *International Journal of Nursing Studies, 48,* (6), 661–671. doi:10.1016/j.ijnurstu.2011.01.016

Kreibig, S.D. (2010). Autonomic nervous system activity in emotion: a review. *Biological Psychology, 84,* 394–421. doi:10.1016/j.biopsycho.2010.03.010

Lemay, M., Bertschi, M., Sola, J., Renevey, P, Parak, J., & Korhonen, I. (2014). Application of optical heart rate monitoring. In E. Sazonov & M.R. Neuman (Eds.), *Wearable Sensors: Fundamentals, implementation and applications* (105-192). San Diego: Elsevier.

Lin, L.I.K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45,* (1), 255-268. doi:10.2307/2532051

Ludbrook, J. (2002). Statistical techniques for comparing measurers and methods of measurement: a critical review. *Clinical and Experimental Pharmacology and Physiology, 29,* (7), 527-536. doi:10.1046/j.1440-1681.2002.03686.x

Lupton, D. (2014, August 27) *Self-tracking modes: Reflexive self-monitoring and data practices*. Paper presented at: 'Imminent citizenships: Personhood and identity politics in the informatic age', ANU, Canberra.

McBride, G.B. (2005). A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. *NIWA Client Report.* Retrieved from: https://www.medcalc.org/download/pdf/McBride2005.pdf

Nederlandse Vereniging voor Cardiologie. *ECG.* Retrieved from http://www.hartwijzer.nl/ECG.php

Noldus Information Technology BV (2016). *Requirement analysis wearables.* Wageningen, the Netherlands, L.E.A. Teekens.

Patel, M.S., Asch, D.A., & Volpp, K.G. (2015) Wearable devices as facilitators, not drivers, of health behavior change. *Journal of American Medical Association, 315,* 459-460. doi:10.1001/jama.2014.14781.

Poh, M.Z., Swenson, N.C., & Picard, R.W. (2010). A wearables sensor for unobtrusive, long-term assessment of electrodermal activity. *IEE Transactions on Biomedical Engineering, 57,* (5), 1243-1252. doi:10.1109/TBME.2009.2038487

Radespiel-Tröger, M., Rauh, R., Mahlke, C., & Mück-Weymann, M. (2003). Agreement of two different methods for measurement of heart rate variability. *Clinical Autonomic Research, 13,* (2), 99-102. doi: 10.1007/s10286-003-0085-7

Ryan, L. (2016). Navigating ethics in the big data democracy. In L. Ryan (Ed.), *The Visual Imperative* (61-84). Elsevier.

Sawh, M. (2016). Apple Watch is best for measuring heart rate from the wrist says study. Retrieved from http://www.wareable.com/apple/apple-watch-is-best-for-heart-rate-monitoring-3357?utm_content=buffer504a2&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

Schäfer, A., & Vagedes, J. (2013). How accurate is pulse rate variability as an estimate of heart rate variability?: A review on studies comparing photoplethysmographic technology with an electrocardiogram. *International Journal of Cardiology, 166,* (1), 15–29. doi: 10.1016/j.ijcard.2012.03.119

Schwartz, B., & Baca, A. (2016). Wearables and apps – Modern diagnostic frameworks for health promotion through sport. *Deutsche Zeitschrift für Sportmedizin, 67,* (6), 131-136. doi:10.5960/dzsm.2016.237

Spierer, D.K., Rosen, Z., Litman, L.L., & Fujii, K. (2015). Validation of photoplethysmography as a method to detect heart rate during rest and exercise. *Journal of Medical Engineering & Technology, 39,* (5), 264-271. doi:10.3109/03091902.2015.1047536

Stahl, S.E., An, H.S., Dinkel, D.M., Noble, J.M., Lee, J.M. (2006). How accurate are the wrist-based heart rate monitors during walking and running activities? Are they accurate enough? *BMJ Open Sport & Exercise Medicine, 2,* (1), 1-7. doi:10.1136/bmjsem-2015-000106

Swan, M. (2012). Sensor Mania! The Internet of things, wearable computing, objective metrics, and the quantified self 2.0. *Journal of Sensors and Actuator Networks, 1, (3),* 217–253. doi:10.3390/jsan1030217

Swift, D.L., Dover, S.E., Nevels, T.R., Solar, C.A., Brophy, P.M., Hall, T.R., … Lutes, L.D. (2015) The intervention composed of aerobic training and non-exercise physical activity (I-CAN) study: Rationale, design and methods. *Contemporary Clinical Trials, 45,* 435–442. doi:10.1016/j.cct.2015.11.005

Tholander, J., & Nylander, S. (2015, April 18-23). *Snot, sweat, pain, mud, and snow: Performance and experience in the use of sports watches.* Paper presented at CHI, Seoul, Korea. doi:10.1145/2702123.2702482

Thompson, W.R. (2016). Worldwide survey of fitness trends for 2016. *ASCM's Health & Fitness Journal, 19,* (6). 9-18. doi:10.1249/FIT.0000000000000164

Tiedemann, A., Hassett, L., & Sherrington, C. (2015). A novel approach to the issue of physical inactivity in older age. *Preventive Medicine Reports, 2,* 595–597. doi:10.1016/j.pmedr.2015.07.008

Tofallis, C. (2015). A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society, 6,6* (8), 1352-1362. doi:10.1057/jors.2014.103

Trull, T.J., & Ebner-Priemer, U. (2012). Ambulatory assessment. *Annual reviews Clinical Psychology, 9,* 151-176. doi:10.1146/annurev-clinpsy-050212-185510

van Dijk, E.T., Beute, F., Westerink, J.H.D.M., & Ijsselstein, W.A. (2015, April 18-23). *Unintended effects of self-tracking.* Paper presented at CHI, Seoul, South-Korea.

Wac, K., & Tsiourti, C. (2014). Ambulatory assessment of affect: survey of sensor systems for monitoring of autonomic nervous systems activation in emotion. *IEEE Transactions on Affective Computing, 5,* 251-272. doi:10.1109/TAFFC.2014.2332157

Wang, R., Blackburn, G., Desai, M., Phelan, D., Gillinov, L., Houghtaling, P., & Gillinov, M. (2016). Accuracy of wrist-worn heart rate monitors. *JAMA Cardiology.* doi:10.1001/jamacardio.2016.3340

Welk, G.J., Schaben, J.A., & Morrow, J.R. Jr. (2004). Reliability of accelerometry-based activity monitors: A generalizability study. *Medicine and Science in Sports and Exercise, 36,* (9), 1637-1645.

Wilhelm, F.H., Pfaltz, M.C., & Grossman, P. (2006). Continuous electronic data capture of physiology, behavior and experience in real life: towards ecological momentary assessment of emotion. *Interacting with Computers, 18, (2),* 171-186. doi:10.1016/j.intcom.2005.07.001

Wilt, B. (2014, August 15). In the city that we love [web log comment]. Retrieved from https://jawbone.com/blog/jawbone-up-data-by-city/

Zaki, R., Bulgiba, A., Ismail, R., & Ismail, N.A. (2012). Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: A systematic review. *PLoS ONE, 7,* (5), e37908. doi:10.1371/journal.pone.0037908

Appendices

Appendix A. The pilot study

Introduction

The goal of the experiment was to have an indication of the variation between subjects. For example, wearables are believed to be sensitive to the amount of melatonin in the skin (Spierer, Rosen, Litman & Fujii, 2015), or the fit of the device could be different for men and women (Wac & Tsiourti, 2014). In this pilot, a heterogeneous group of participants were tested, to see if and how their characteristics influence the results. This experiment focused on the differences in participants relevant for the wearable, meaning the difference in measurements, not the differences in the values. The aim was to get a handle on the amount and diversity of participants needed for the actual experiment.

Method

Participants

Colleagues from the Wageningen office were asked to participate in the experiment. If they wanted to participate, they could chose a timeslot. Fifty timeslots were available, but a minimum of fifteen participants was accounted for. Seventeen persons volunteered to participate, thirteen of which were male, and four were female. Their age varied between 21 and 56, with a mean of 38 and a standard deviation of 12.

Apparatus

Biopac. The participants had three electrodes attached to them from the Biopac. The Biopac was used as ground truth in this experiment, since it's one of the most used devices for psychophysiological measurements. The electrodes were placed as recommended by the manufacturer, with the negative electrode on the right collarbone, the negative electrode on the left lowest rib and the ground electrode on the right lowest rib. A photo plethysmography (PPG) sensor was used on the right forefinger. Electrodes for electrodermal activity were placed on the palm of the non-dominant hand. One was placed on the palm just before the thumb and the other on the palm on the outer side.

Empatica E4. The Empatica E4 is a new wearable device that aims to measure heart rate, heart rate variability, EDA, movement and temperature. Empatica wants to introduce wearables to the research market and claims the E4 to be accurate enough for psychophysiological research. The E4 is placed on the wrist and has different sensors at the

back of the top of the device. The participants wore the E4 on the wrist of their non-dominant hand.

Procedure

Participants were given an informed consent. Each participant was asked to place their wrist on a grey paper where a photo was taken. The participant was connected to the Biopac and the E4. They had to sit calmly for five minutes and walk around for two minutes.

Analysis

EDA. The raw values of EDA were used for visually checking the similarity between the devices. The frequency of SCRs were used for further calculation of similarity.

HR. The beats per minute calculated by the devices, were used as values for calculations about the similarity.

Participant characteristics. Different characteristics of the participants, such as age, sex, skin color, wrist size and amount of fat on the wrist, hair on the wrist and use of lotion on the wrist could influence the measurements. Sex, age and skin color were the most important of these variables. To have an objective measurement of skin color, the wrist of all the participants was photographed while placed on a grey background. Since the background of these photos is all the same, the skin color can be objectively measured. The skin color was rated on a scale from 1 to 5, based on standard deviations of the RGB color spectrum of all participants. Participants were also rated on a scale from 1 to 3 for the amount of fat and hair on the wrist. This was done by visual estimation. The participants were asked for age and use of lotion.

Data analysis. Pearson's correlation coefficients were calculated, to see to which extent the different devices match in their derived measures. The calculations were done per participant, so an estimation could be given about the variance between participants.

Results

Data preparation for analysis

The data was set up in two different forms, one with in each column the data of a specific participant with a specific device, and the other based on time, in which all the data was in the same column.

Heart rate. The data from the Biopac had a lot of noise, mostly due to movements, and as was revealed later, old electrodes. This meant that the data was not as good as expected, resulting in a derived heart rate varying from 30 beats per minute to, in some cases, 240 beats per minute. These outliers were deemed unrealistic, with regards to the activities of

the participants. Since movement was a big factor in creating the noise, only the values measured when the participant was sitting were used. To further solve this, the derived beats per minute data, without outliers were used. This meant that any value above or below two standard deviations from the mean was discarded. This resulted in realistic values for most participants. The data from the PPG from participant 3 was excluded, because even with these precautions, the standard deviation was still 19.31 and there were values below 30 bpm and above 140 bpm. The ECG data from participant 11 was also excluded, because and the standard deviation was 16.96 and there were many values above 130 bpm, when the previous value was around 75 bpm.

Electrodermal activity. To calculate the amount of SCRs, a low-pass filter was used. The amplitude criterion for the present study was .01 µS and the criterion for the speed changes .000009 µS. the minimal gap between peaks was 700 ms. The SCRs were all in a pre-defined window of 1-3 to 1-5 seconds, with a amplitude of .01 to .05 µS. The Biopac electrodes were not properly attached by participant 12 and 14, resulting in long periods of 0.00 values. This data was not discarded, since for the calculation with the SCRs, it was not of a big influence. This is because the SCRs still occur by participant 12 and by participant 14, it happened in a period that the E4 gave little to no SCRs. The data of participant 15 was discarded, because the synchronization did not work.

Heart rate

Similarity between devices. The similarity of the devices was compared by visual estimation and by use of Pearson's r. Figure 1 represents the visually estimated best similarity of the devices. Figure 1 shows the values of participant 5, in which clearly can be seen that the values vary more when measured by the Biopac, due to automatic algorithms in the E4.

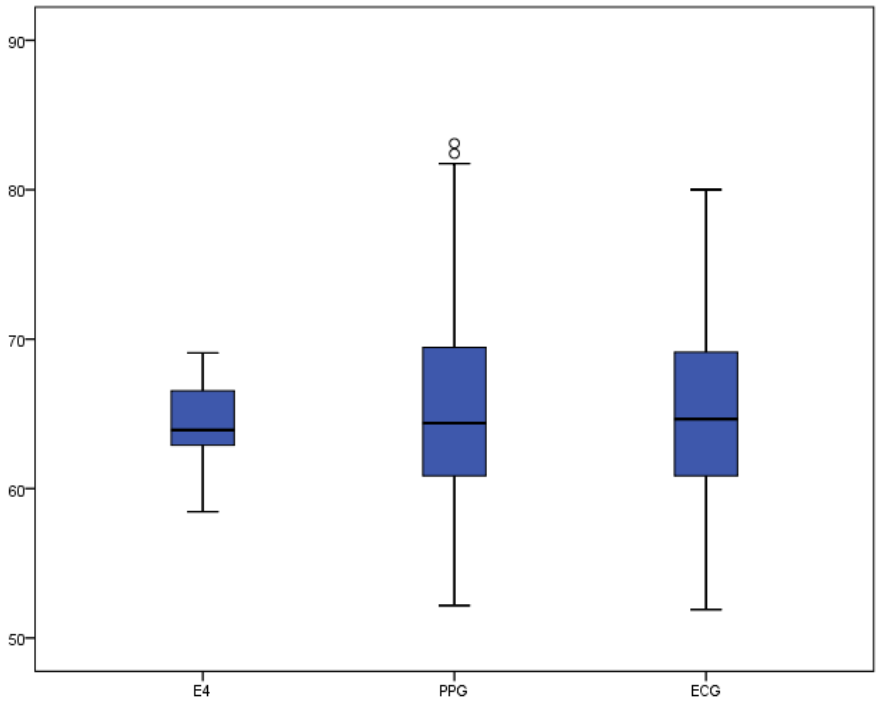


Figure 1. Beats per minute of participant 5 by three different measurements.

The data of the worst similarity, as estimated visually, can be seen in figure 2. These are the values of participant 8, in which a difference is clear between the data from the E4 and from the Biopac (PPG and ECG).

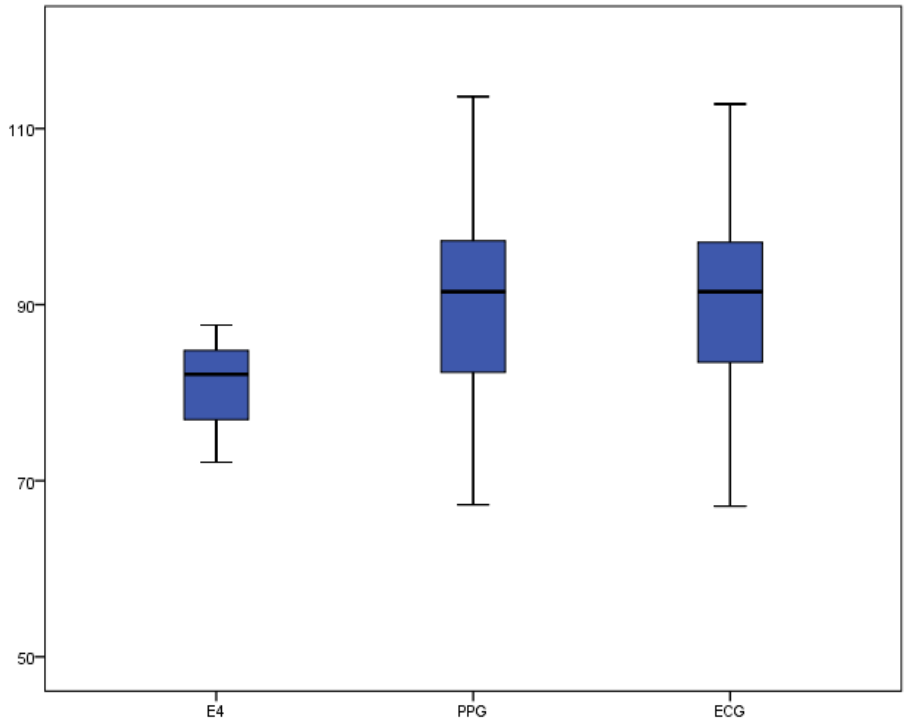


Figure 2. Beats per minute of participant 8 by three different measurements.

To have a statistical measurement of similarity, Pearson's r was calculated between the E4 and the two measures of the Biopac. The values and significance level are displayed per participant in table 1.

Table 1.

Correlations between the E4 and the Biopac heart rate data.

* indicates statistical significant data.

	E4 - PPG Biopac		E4 - ECG Biopac	
	Pearson's r	significance	Pearson's r	significance
Participant 1	-.08	.22	-.04	.50
Participant 2	.14*	.03	-.46*	.00
Participant 3			.02	.65
Participant 4	.09	.11	.01	.82
Participant 5	.14*	.01	.14*	.01
Participant 6	-.03	.73	.00	1
Participant 7	.33*	.00	.50*	.00
Participant 8	.09*	.10	.10	.07
Participant 9	.17*	.00	-.16*	.00
Participant 10	.04	.62	.08	.35
Participant 11	.20*	.00		
Participant 12	.37*	.00	.12	.09
Participant 13	.11*	.03	.28*	.00
Participant 14	.39*	.00	.42*	.00
Participant 15	.26*	.00	.36*	.00
Participant 16	-.30*	.00	-.31*	.00
Participant 17	.20*	.00	.27*	.00

Variance. A table was set up with descriptive statistics of the variance in values per participant, as displayed in table 2. As can be seen, the mean heart rate differs from participant to participant. What is most striking in this table, is the differences in similarity between measurements. For some participants, such as 2, 4, 5, 7, 10, 15, and 17, the mean and standard deviation are very similar across the measurements. However, for other participants there are clear differences, such as participant 3, 8, and 16, where the mean heart rate differs almost or more than ten beats per minute.

Table 2.

Mean and standard deviations of beats per minute per participant

	E4		PPG		ECG	
	Mean HR	SD	Mean HR	SD	Mean HR	SD
Participant 1	68.38	2.36	69.56	6.10	74.00	12.51
Participant 2	69.98	3.06	69.49	3.20	68.24	1.39
Participant 3	61.97	7.48			54.70	4.27
Participant 4	74.84	6.21	73.81	6.99	73.89	6.11
Participant 5	64.36	2.57	65.56	6.45	65.38	6.16
Participant 6	75.22	8.81	72.42	6.00	72.22	5.61
Participant 7	66.57	6.56	65.33	10.51	67.47	16.46
Participant 8	80.75	4.55	90.55	9.82	90.53	9.41
Participant 9	74.71	3.42	76.01	7.94	77.21	12.90
Participant 10	64.35	4.38	63.85	3.63	64.38	4.13
Participant 11	86.01	4.54	83.18	10.04		
Participant 12	71.40	2.99	69.85	5.05	68.61	9.49
Participant 13	61.11	7.04	56.22	9.90	58.72	4.19
Participant 14	81.19	8.55	78.63	4.52	78.32	5.43
Participant 15	77.26	4.43	76.14	7.02	76.43	5.73
Participant 16	69.23	19.53	56.69	7.37	55.78	5.72
Participant 17	73.94	2.80	73.41	5.43	73.43	4.74

Electrodermal activity

Similarity between devices. The comparison between devices was done by using the raw EDA data and the amount of SCRs.

Raw EDA data. The raw data of the EDA values were set up in a double Y-axis graph, so the relative changes could be compared. The similarity between devices differed per participant, as can be seen in figure 4 and 5. Figure 4 displays the data from participant 6, where the values had the same relative changes. In figure 5 the values of participant 3 are displayed, but the different measures look disparate.

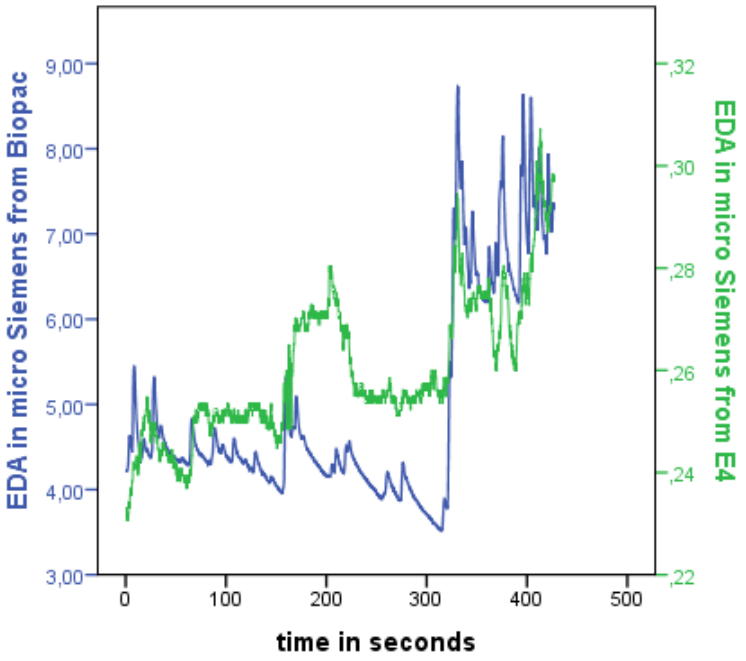


Figure 4. EDA values of participant 6 in micro Siemens by the E4 and the Biopac, relative over time.

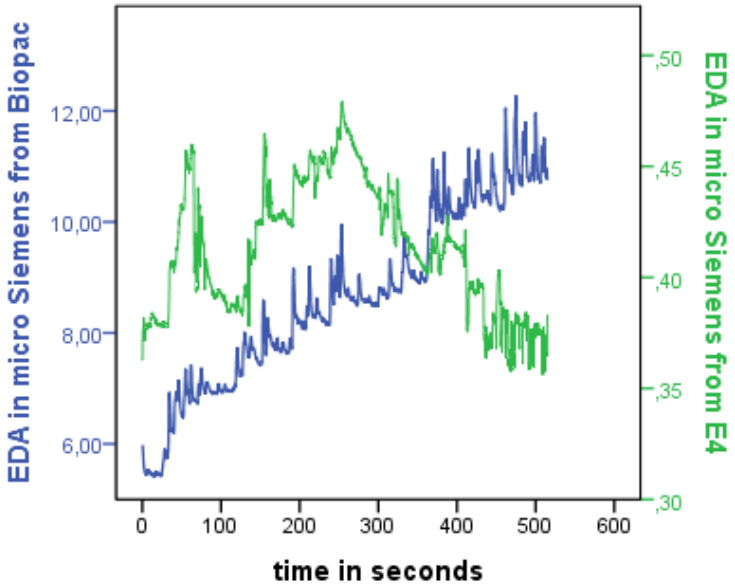


Figure 5. EDA values of participant 3 in micro Siemens by the E4 and the Biopac, relative over time.

Pearson's correlation coefficients for the devices were calculated, to have a measure of similarity. All correlations were significant, but few were above .6, which could be described as an adequate correlation. This was the case for participant 1, 5, 6, 7, 8, 9, 11, and 12. Participant 12 had a negative correlation, but this is not so relevant, it still indicates that the relation between the devices has an adequate correlation, only this relation is reverse compared to the others. Participant 5, 7, 8, and 9 had a strong correlation.

Table 3.

Pearson's correlation coefficients for the EDA values from the E4 and the Biopac.

** indicates statistical significant data.*

	E4-Biopac	
	Pearson's r	*significance*
Participant 1	.65*	.00
Participant 2	.46*	.00
Participant 3	-.16*	.00
Participant 4	.38*	.00
Participant 5	.98*	.00
Participant 6	.71*	.00
Participant 7	.84*	.00
Participant 8	.87*	.00
Participant 9	.89*	.00
Participant 10	.07*	.00
Participant 11	.70*	.00
Participant 12	-.65*	.00
Participant 13	.17*	.00
Participant 14	.35*	.00
Participant 16	.15*	.00
Participant 17	.58*	.00

SCRs. The devices were also compared in the amount of SCRs they detected. In table 2 the values for the amount of SCRs can be found sorted by device. As can be seen, the values were not close to each other, which could be due to differences in sensitivity. Of interest is that the relative differences due to activity were sometimes comparable, such as with participant 3, 4, 6, 8, 9, and 11.

Table 4.

Amount of SCRs per participant, split by device and activity.

	Sitting		Walking		Total	
	E4	Biopac	E4	Biopac	E4	Biopac
Participant 1	29	44	0	37	29	80
Participant 2	7	68	5	28	12	96
Participant 3	38	75	8	14	46	89
Participant 4	11	49	4	23	15	72
Participant 5	2	89	2	15	4	104
Participant 6	17	35	7	14	24	49
Participant 7	31	66	8	22	39	88
Participant 8	51	98	17	35	68	133
Participant 9	5	54	1	17	6	71

Participant 10	8	59	10	40	18	99
Participant 11	59	105	28	57	87	162
Participant 12	16	65	8	116	24	181
Participant 13	29	76	19	27	48	103
Participant 14	20	85	38	62	58	147
Participant 16	3	37	26	61	29	98
Participant 17	1	36	0	50	1	86

Influences of personal characteristics. To check for the individual differences between participants, a generalized estimating equation was run with amount of skin conductance responses as the dependent variable, participant number as subject variable, time and device as within subject variable, task as random factor and the characteristics skin color, amount of hair on wrist, width of wrist, gender and age as covariates. The only significant effect found was from task ($\chi = 24.97$, $\alpha < .00$). The unstandardized residuals were plotted and were normally distributed.

Discussion

As a pilot, the experiment was very useful, meaning that several factors that could be of influence were discovered before the actual research started. One of the most important findings was the unreliability of the data used as the gold standard, the Biopac. This was due to old electrodes and movements that caused friction on the electrodes. To solve this in following research, new electrodes were ordered and there was chosen to do ambulatory research, in which participants do not have to move much. To have better data, the ECG signal has to be cleaned, before using the derived beats per minute values. In this experiment, the used values from the Biopac for heart rate were not derived from a clean ECG, meaning that they may be less valid. Since this is the standard the E4 is compared to, absolute conclusions about the agreement between the devices cannot be drawn from this experiment only.

There seem to be differences between participants. For some participants the values are very similar in both devices, while for other participants there is no clear connection. This could mean that there are individual differences in play, but no significant factors were found for the electrodermal activity data.

Since it was decided to do ambulatory research for the coming experiment, it would be inconvenient for the participants to have electrodes placed on their wrist. That could hinder their ability to type for example. The E4 is also one of the few wearables with EDA measurement, so for the coming experiment, only heart rate measurement will be used.

References

Spierer, D.K., Rosen, Z., Litman, L.L., & Fujii, K. (2015). Validation of photoplethysmography as a method to detect heart rate during rest and exercise. *Journal of Medical Engineering & Technology, 39,* (5), 264-271. doi:10.3109/03091902.2015.1047536

Wac, K., & Tsiourti, C. (2014). Ambulatory assessment of affect: survey of sensor systems for monitoring of autonomic nervous systems activation in emotion. *IEEE transactions on affective computing, 5,* 251-272. doi:10.1109/TAFFC.2014.2332157

Appendix B. placement of the elctrodes

An example picture of the correct placement of the electrodes for the ECG. Only three electrodes were used, 1 for the negative electrode, 2 for the ground electrode and 3 for the positive electrode.

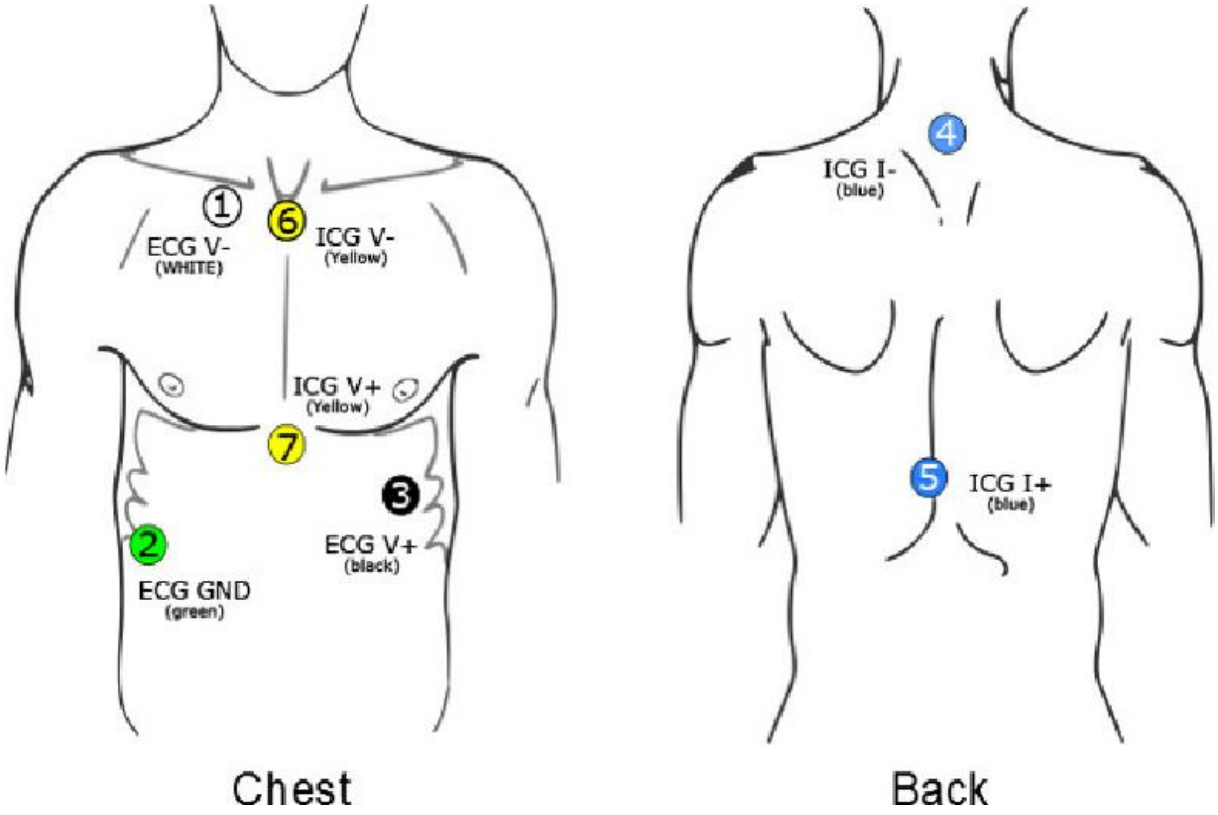


Figure obtained from van Dijk et al. (2013).

Van Dijk, A.E., van Lien, R., van Eijsden, M., Gemke, R.J.B.J., Vrijkotte, T., & de Geus, E.J.C. (2013). Measuring cardiac autonomic nervous system (ANS) activity in children. *Journal of Visualized Experiments, 74,* (74). doi:10.3791/50073

Appendix C. Coding scheme from the Observer

Independent variables

- Participant number
- Age
- Gender

Observer XT coding scheme

Activity

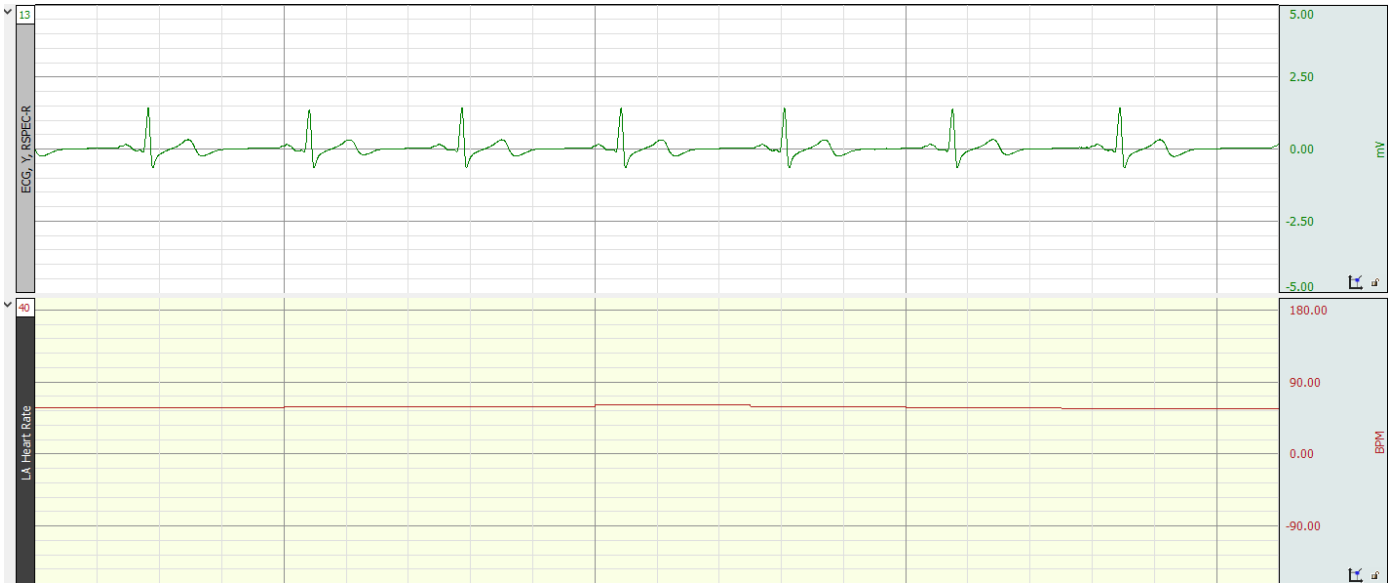
- Working at desk
- Meeting
- Leaving desk
- Leaving floor
- Return to desk
- Other

Appendix D. Position of the wearables per participant

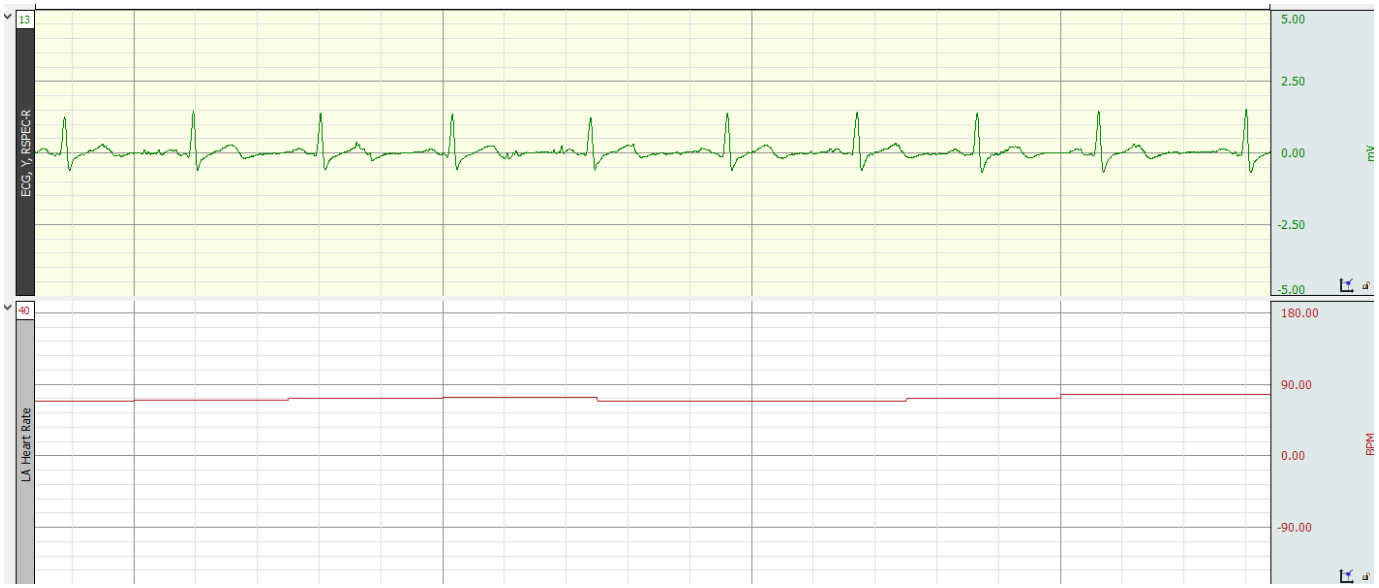
Participant	left closest to hand	second left	right closest to hand	second right
1	Motorola	Microsoft Band	Garmin	Philips
2	Microsoft Band	Motorola	Philips	Garmin
3	Garmin	Philips	Motorola	Microsoft Band
4	Microsoft Band	Garmin	Motorola	Philips
5	Motorola	Microsoft Band	Garmin	Philips
6	Philips	Motorola	Microsoft Band	Garmin
7	Garmin	Philips	Motorola	Microsoft Band
8	Microsoft Band	Garmin	Philips	Motorola
9	Motorola	Microsoft Band	Philips	Garmin
10	Philips	Motorola	Microsoft Band	Garmin
11	Garmin	Philips	Motorola	Microsoft Band
12	Garmin	Microsoft Band	Philips	Motorola
13	Motorola	Microsoft Band	Garmin	Philips
14	Philips	Motorola	Garmin	Microsoft Band
15	Garmin	Philips	Motorola	Microsoft Band
16	Microsoft Band	Garmin	Philips	Motorola
17	Motorola	Microsoft Band	Garmin	Philips
18	Philips	Motorola	Microsoft Band	Garmin
19	Garmin	Philips	Motorola	Microsoft Band
20	Microsoft Band	Garmin	Philips	Motorola
21	Motorola	Microsoft Band	Garmin	Philips

Appendix E. Description of discarding noisy data in ECG signal

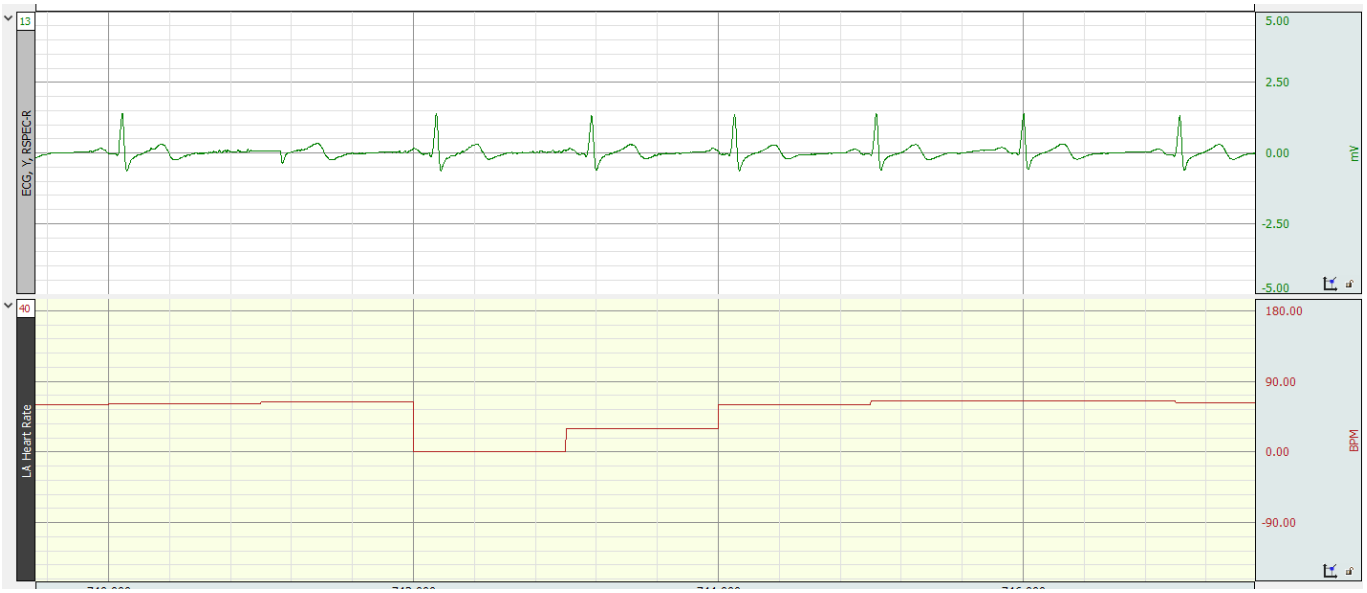
Most important feature to look into is the possibility to distinguish the normal ECG line, the upper, green, line. This should be as expected, with the regular peaks and valleys. When it is disrupted, only mark the data as good when the BPM line, the lower red line, is at a realistic value. Examples are below.



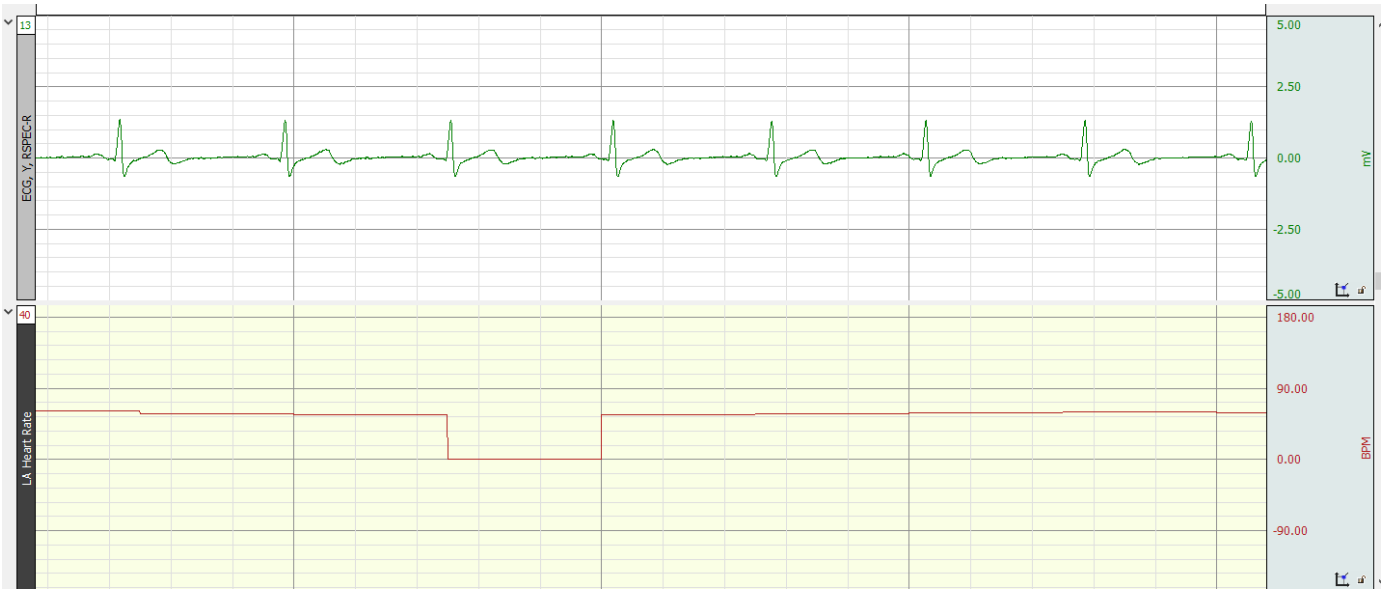
Perfect example



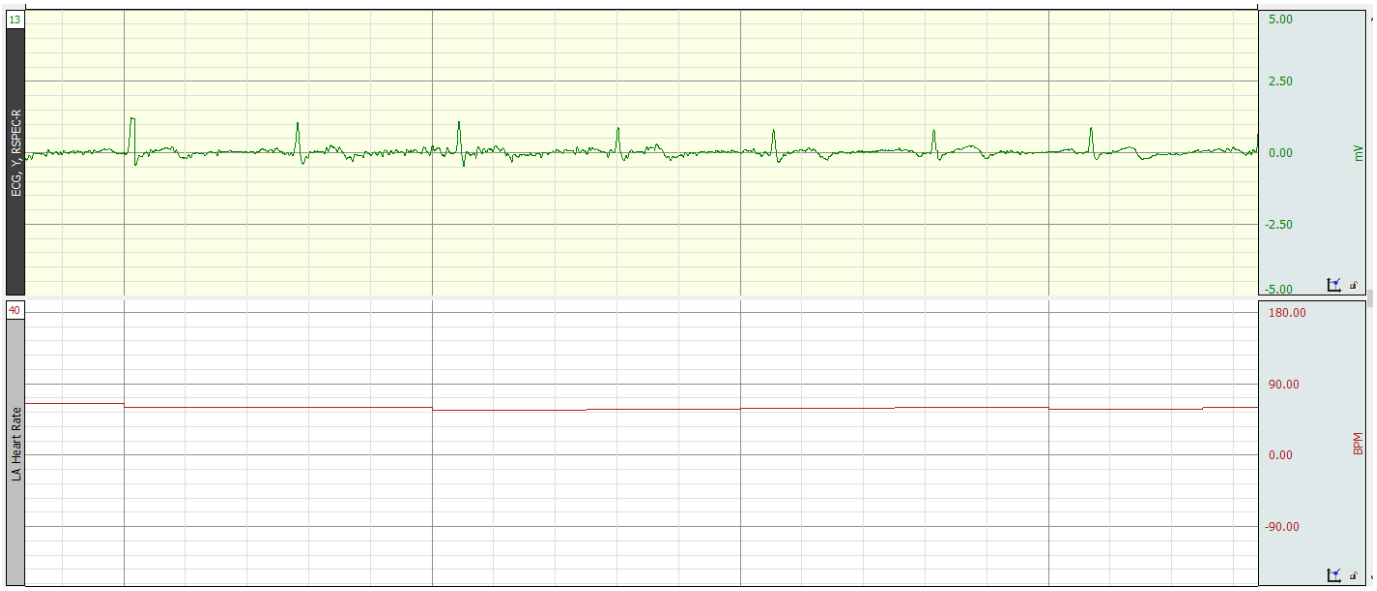
Kept, because R peaks are clearly distinct.



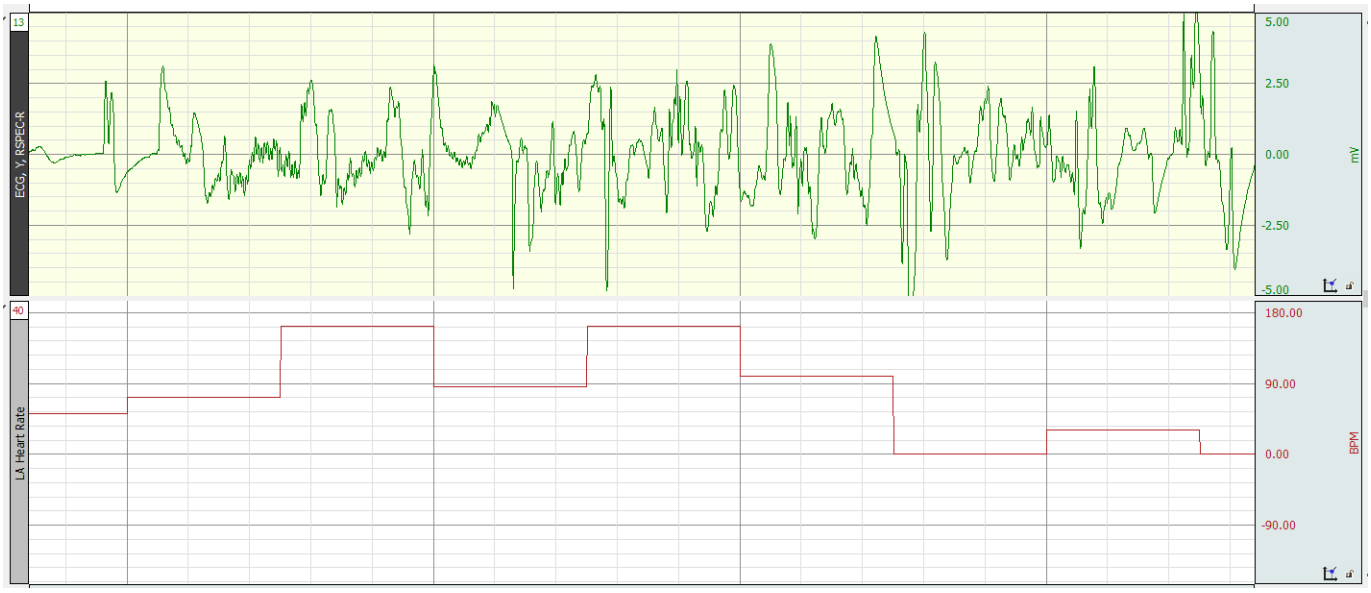
Kept, because no clear disruptions in ECG line. 0-values were removed after.



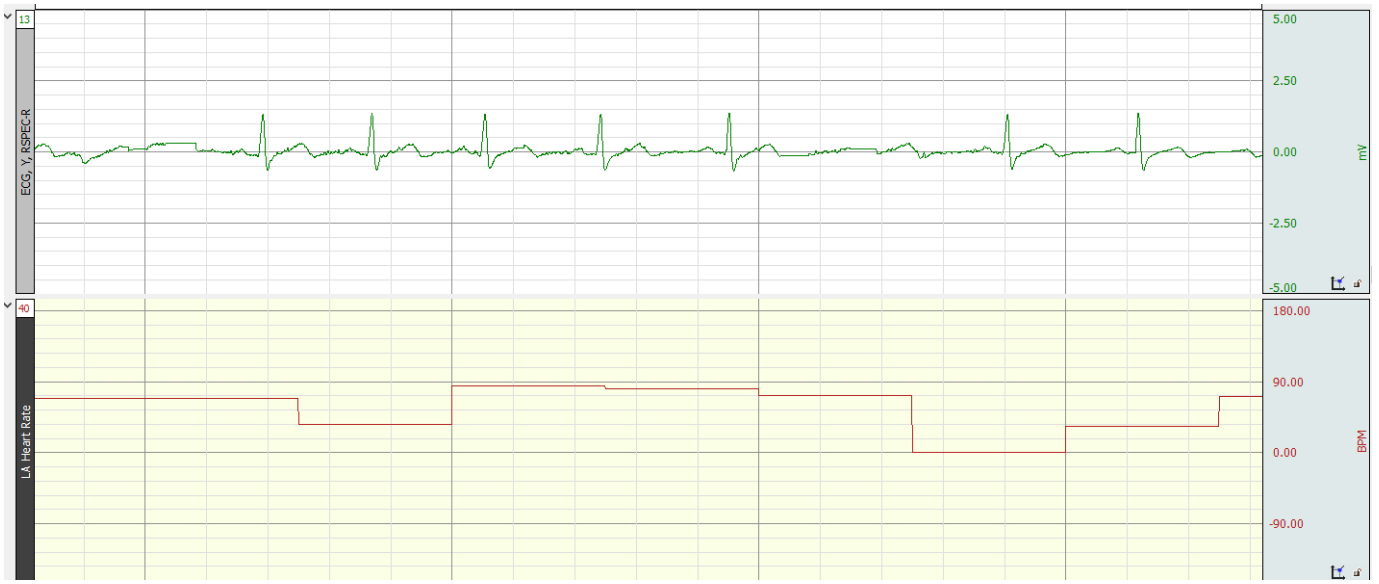
Kept, because no clear disruptions in ECG line. 0-values were removed after.



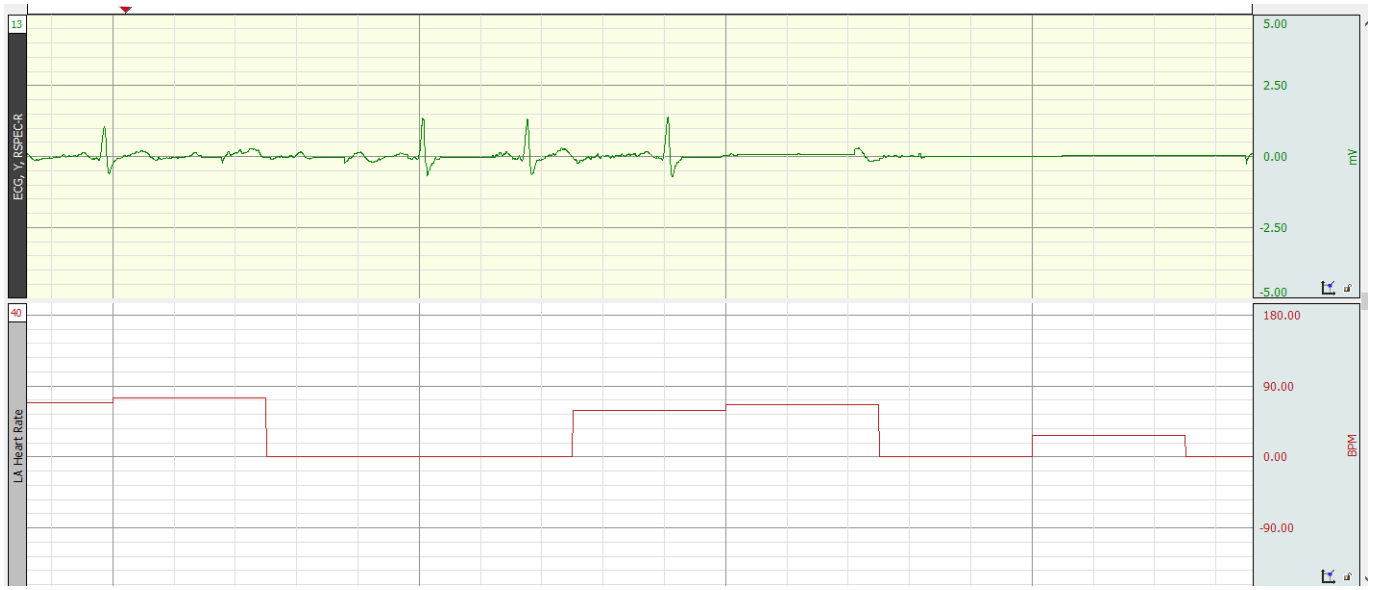
Kept, but only because all the data from this participant has this noise in it. The ECG line is still clearly distinguisable. If this type of line occurs before/after loss of data/more disruption, it is discarded. If it seems to interfere with BPM, or looks like it could interfere BPM (by having high peaks in noise) it is discarded



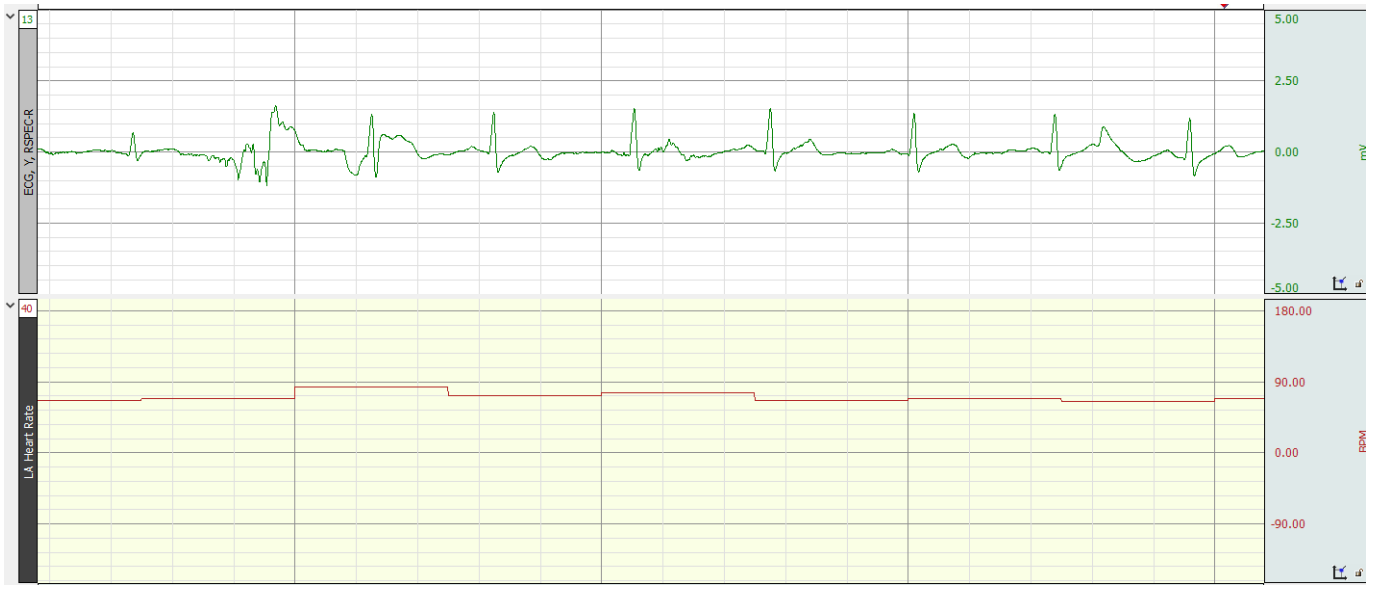
Discarded, since the ECG line is clearly disturbed.



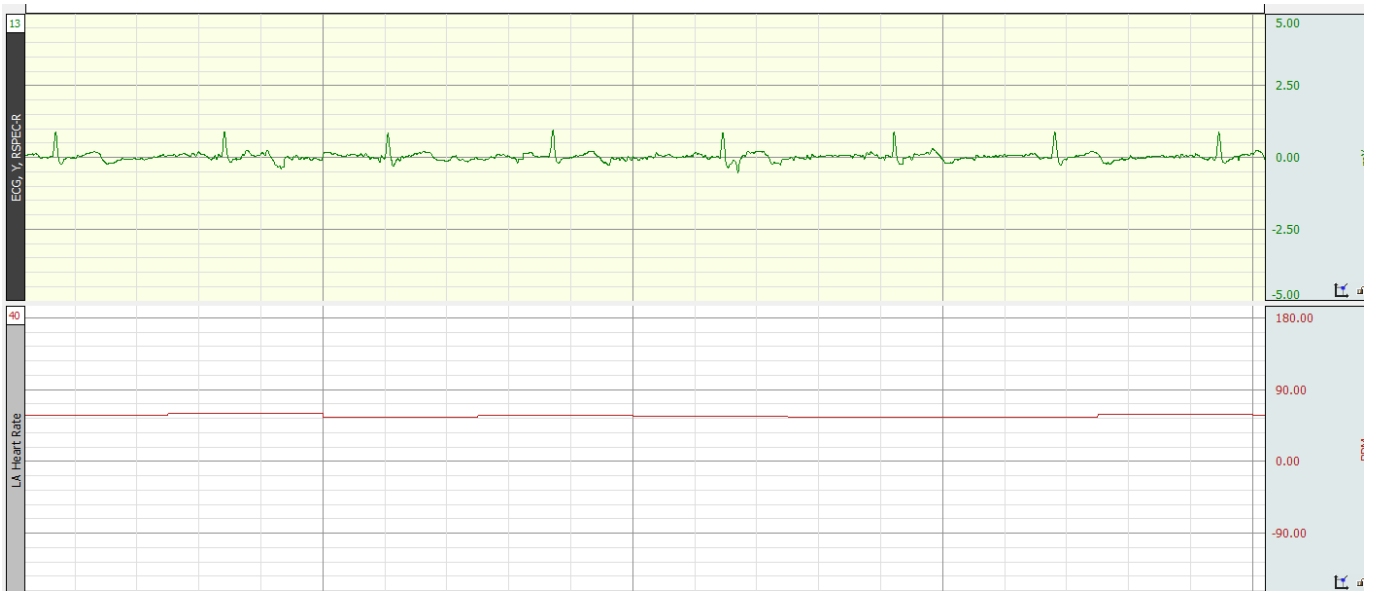
Discarded, because some disruptions in ECG line cause unreliable values for BPM.



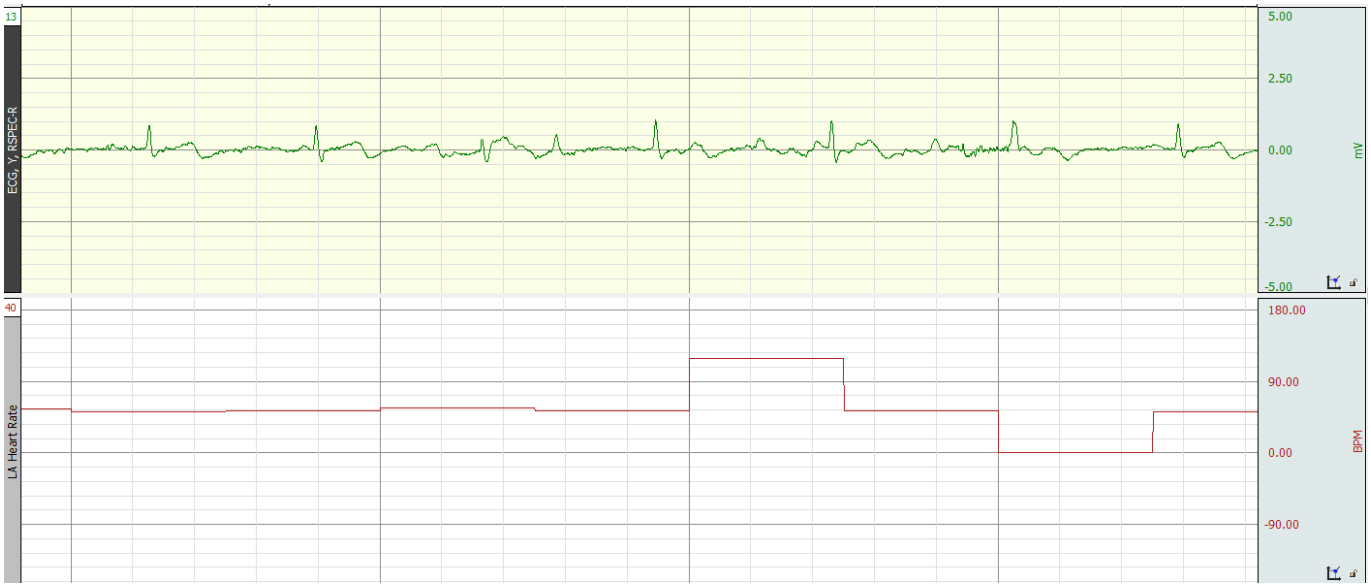
Discarded, because participant is out of reach of the Biopac, as can be seen by minimal change in ECG line at the end – and afterwards.



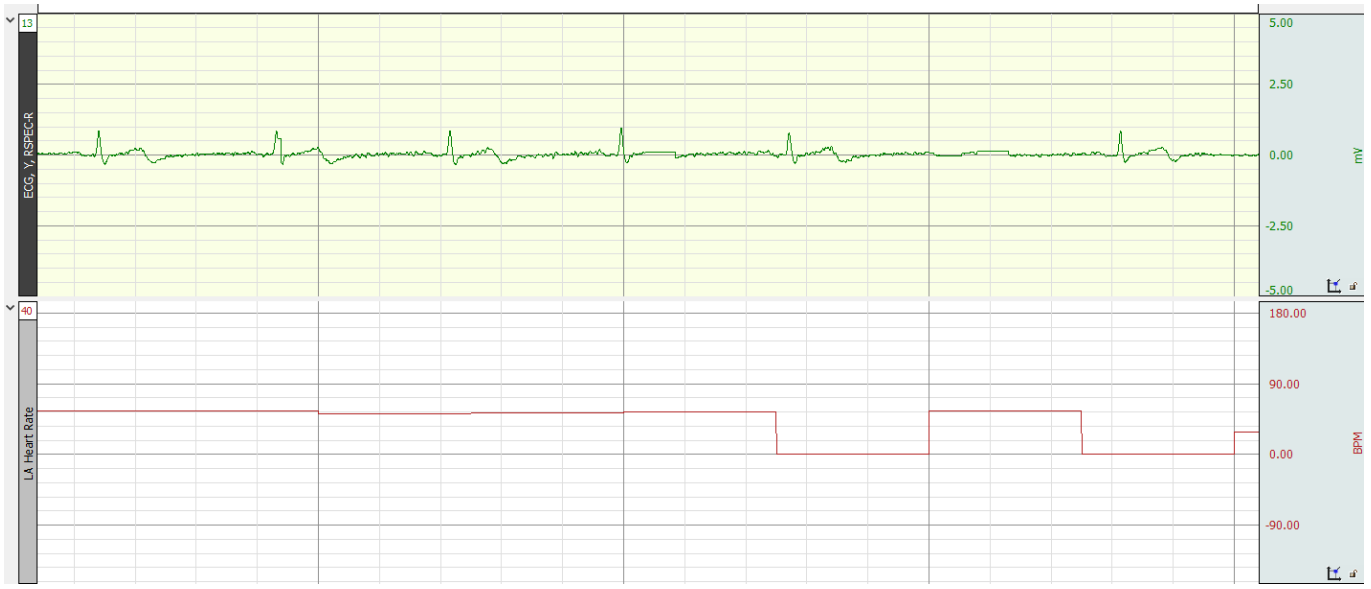
Discarded, because of noise at the start and is still not normal at the end of this screen.



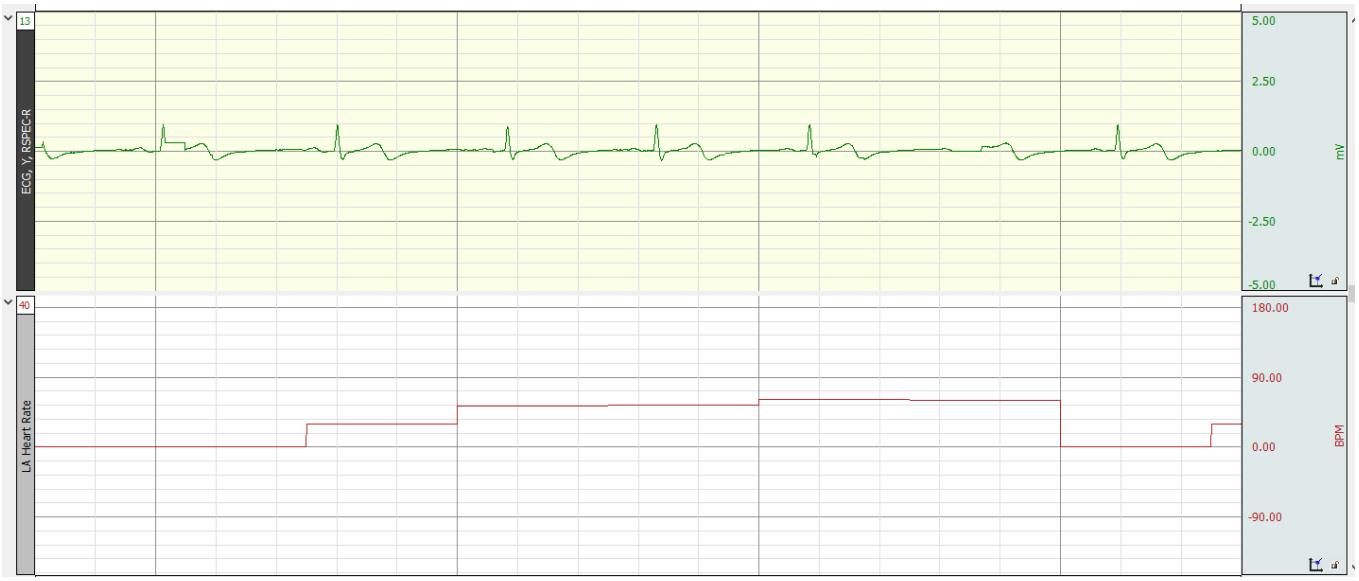
Discarded, since noise is interrupting the ECG line.



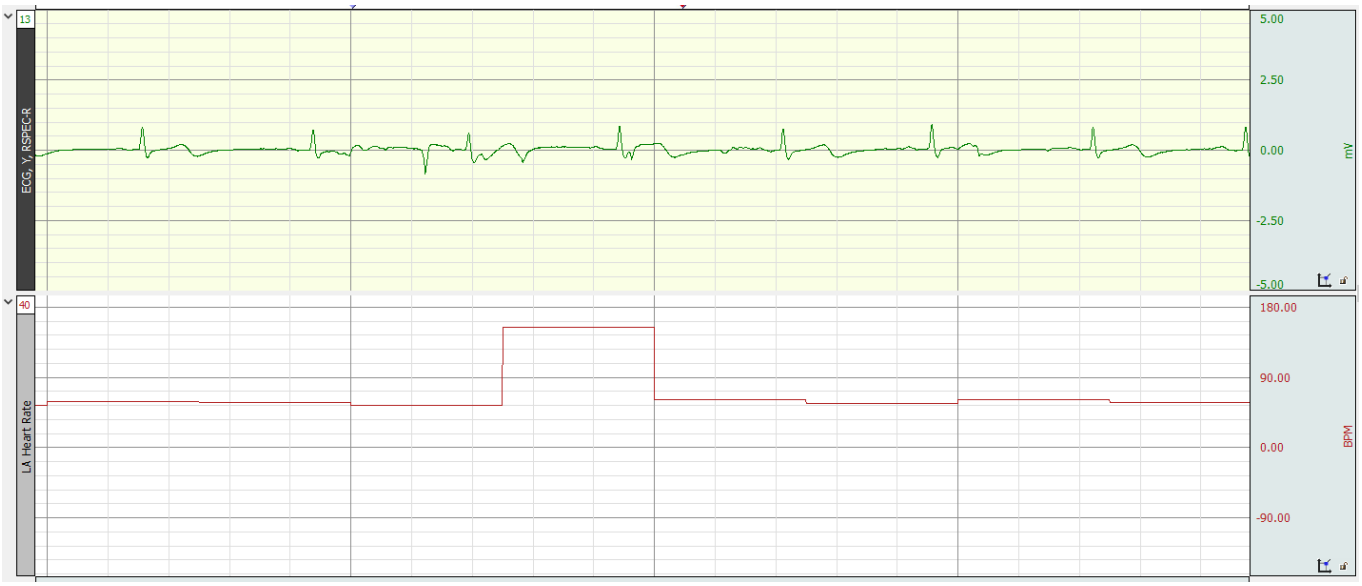
Discarded, ECG line disrupted



Discarded, ECG is not clear



Discarded, ECG line not clear, lots of influence on BPM

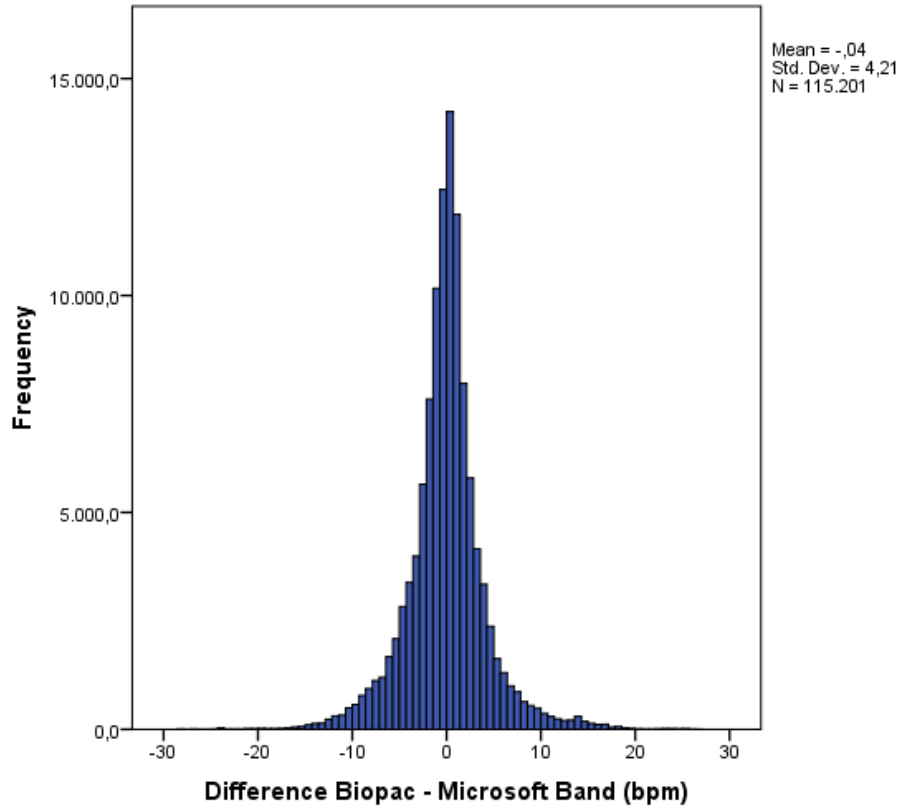


Discarded, because more R-peaks are calculated than there should be, as can be seen in BPM

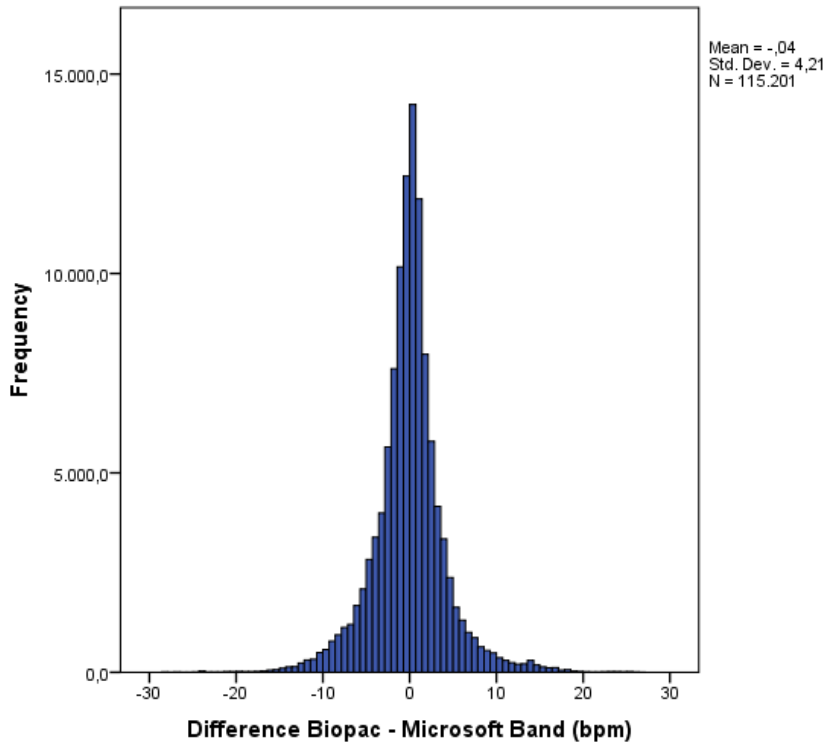
Appendix F. Plots to check for normal distributions of the differences for the averaged data

Normal distribution of the differences of the averaged data, as a prerequisite of the Bland-Altman analysis.

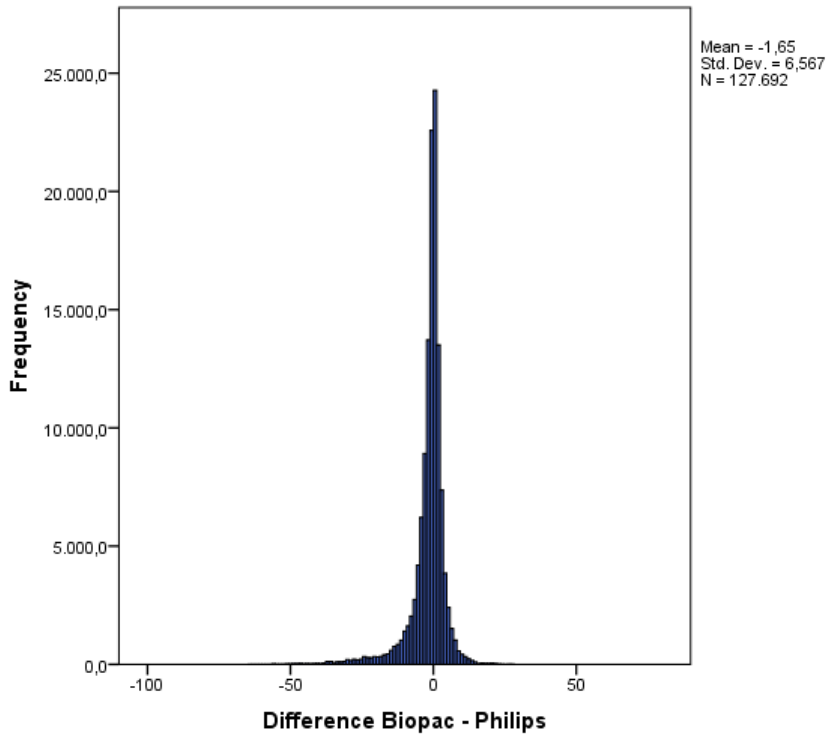
Microsoft Band



Motorola

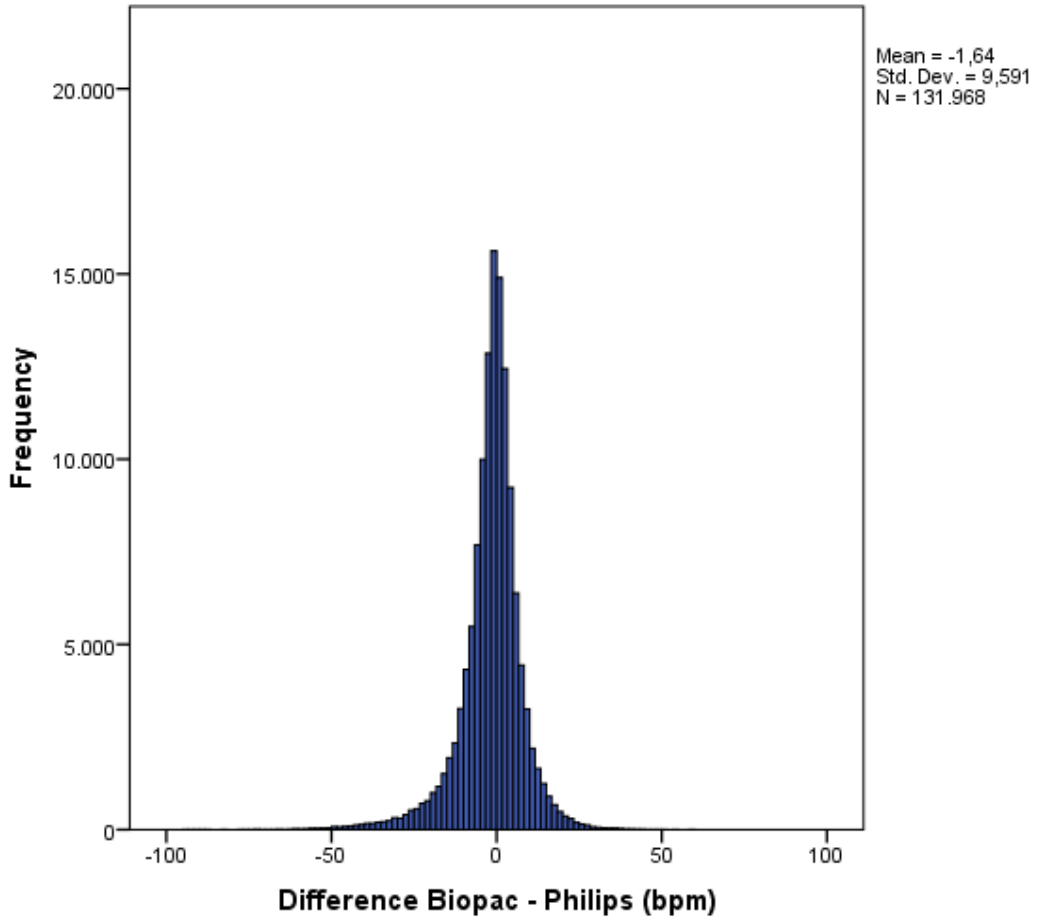


Philips



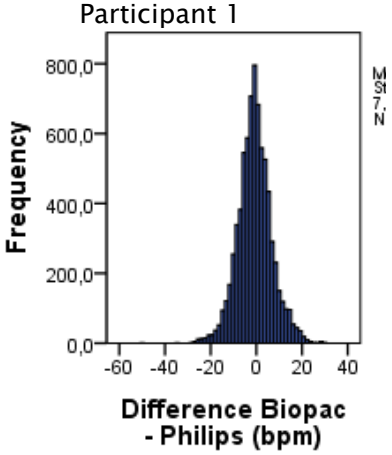
Appendix G. Plots to check for normal distribution of differences for Philips

Differences Biopac and Philips from the non-averaged data.

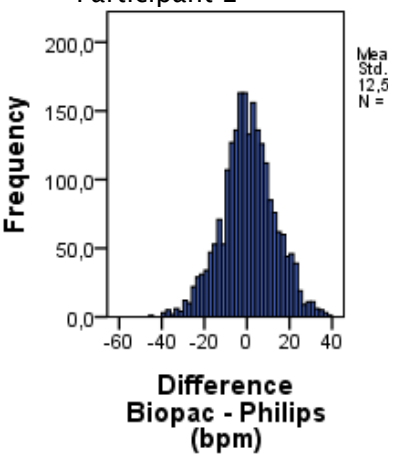


Appendix H. Plots to check normal distributions differences per participant

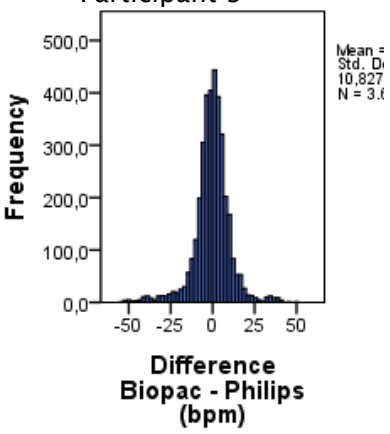
Participant 1



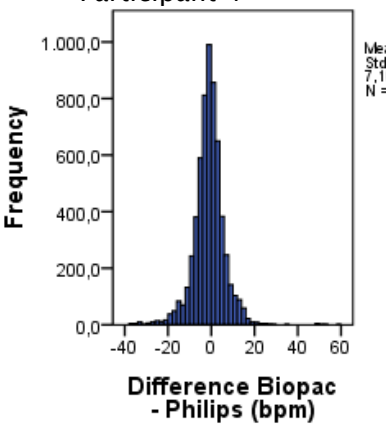
Participant 2



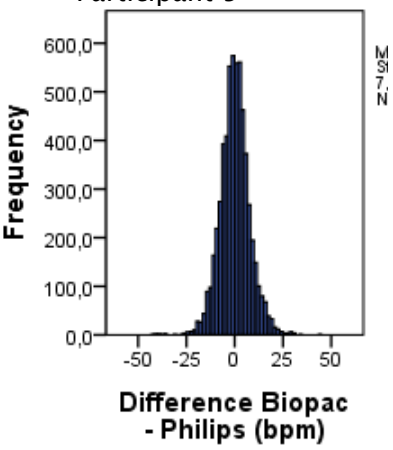
Participant 3



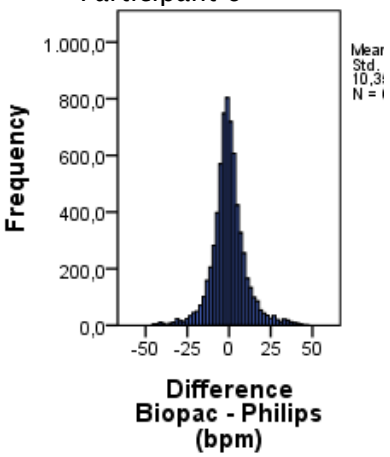
Participant 4



Participant 5



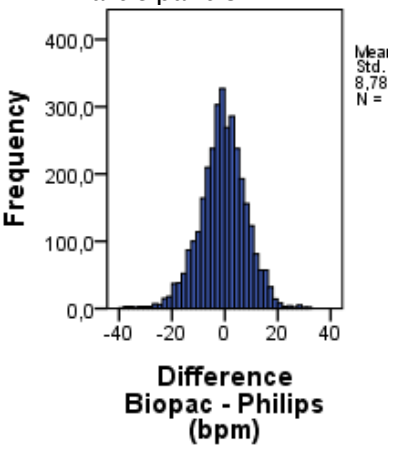
Participant 6



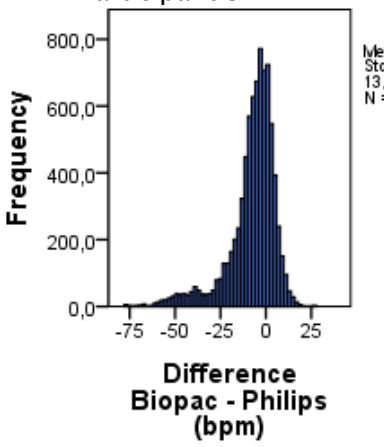
Participant 7

No Philips data

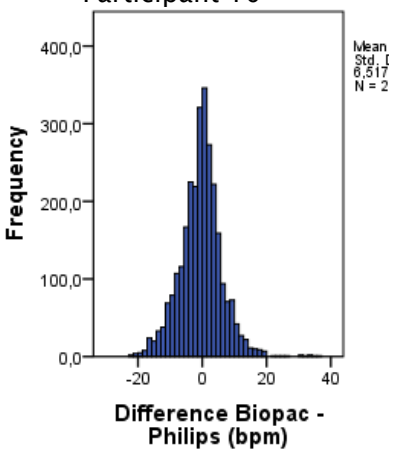
Participant 8



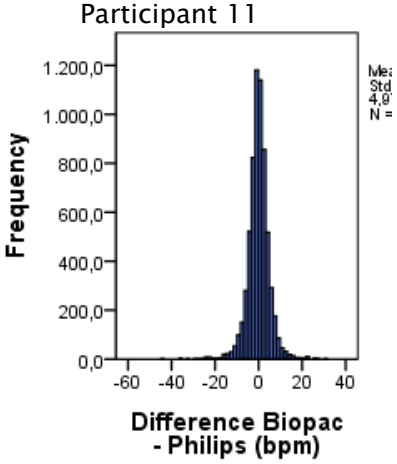
Participant 9



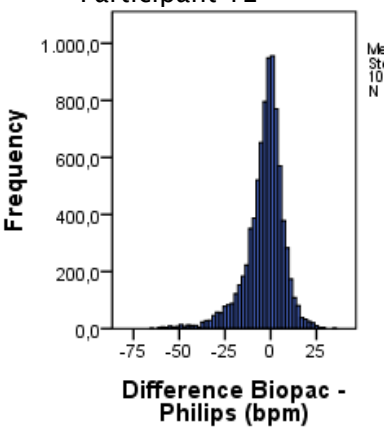
Participant 10



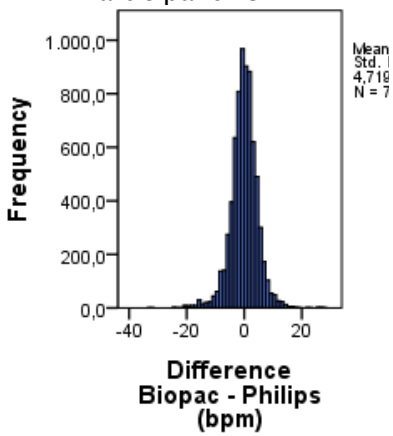
Participant 11



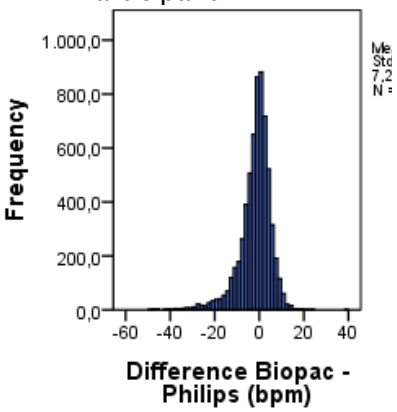
Participant 12



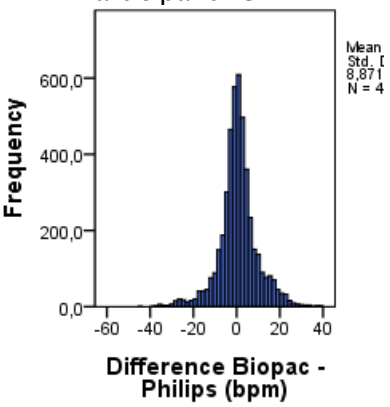
Participant 13



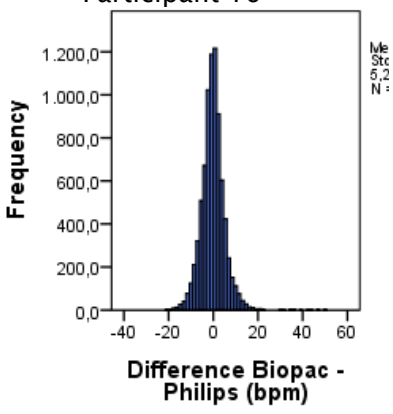
Participant 14



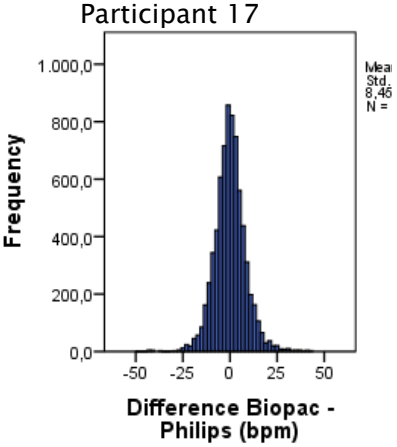
Participant 15



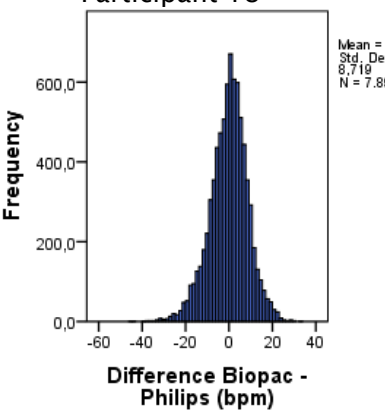
Participant 16



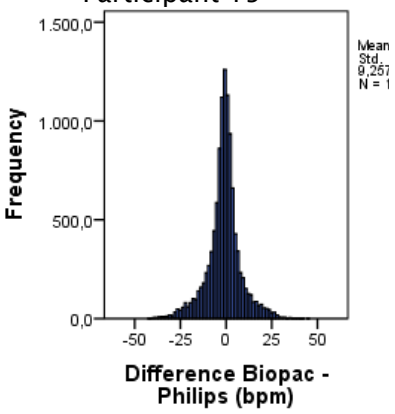
Participant 17



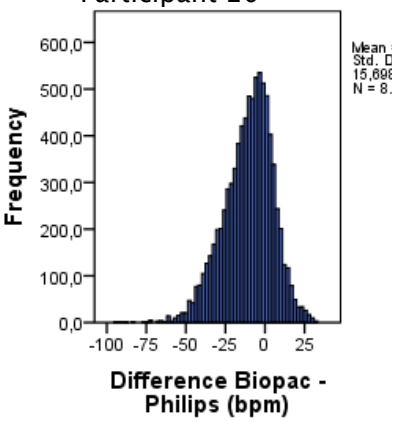
Participant 18



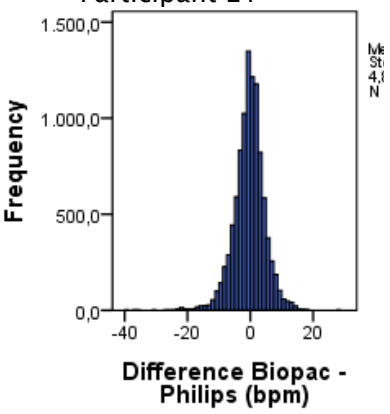
Participant 19



Participant 20



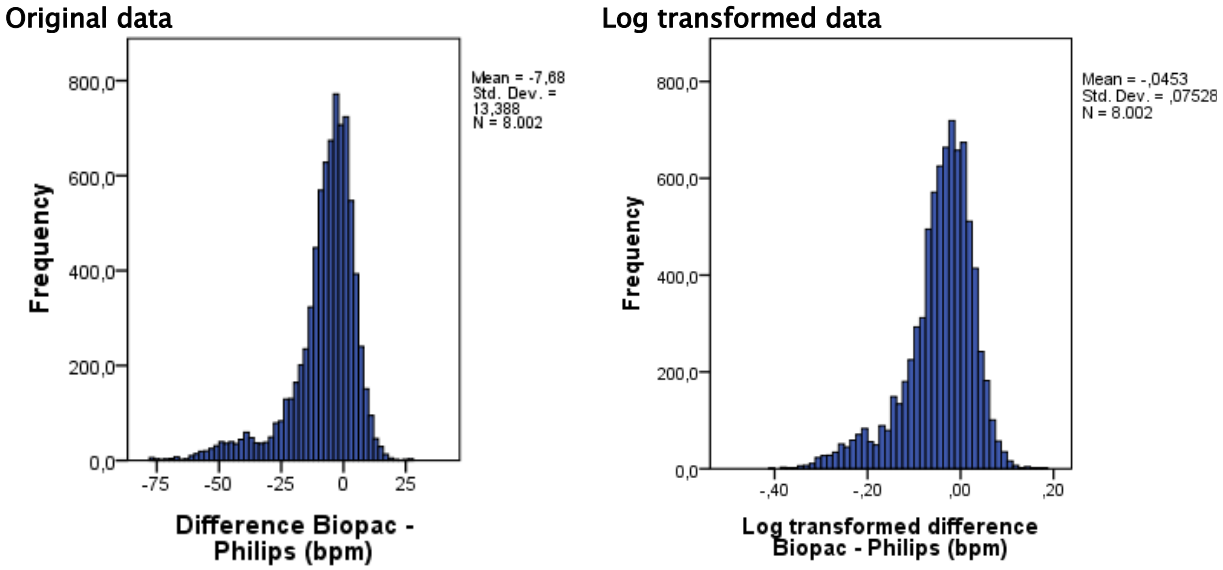
Participant 21



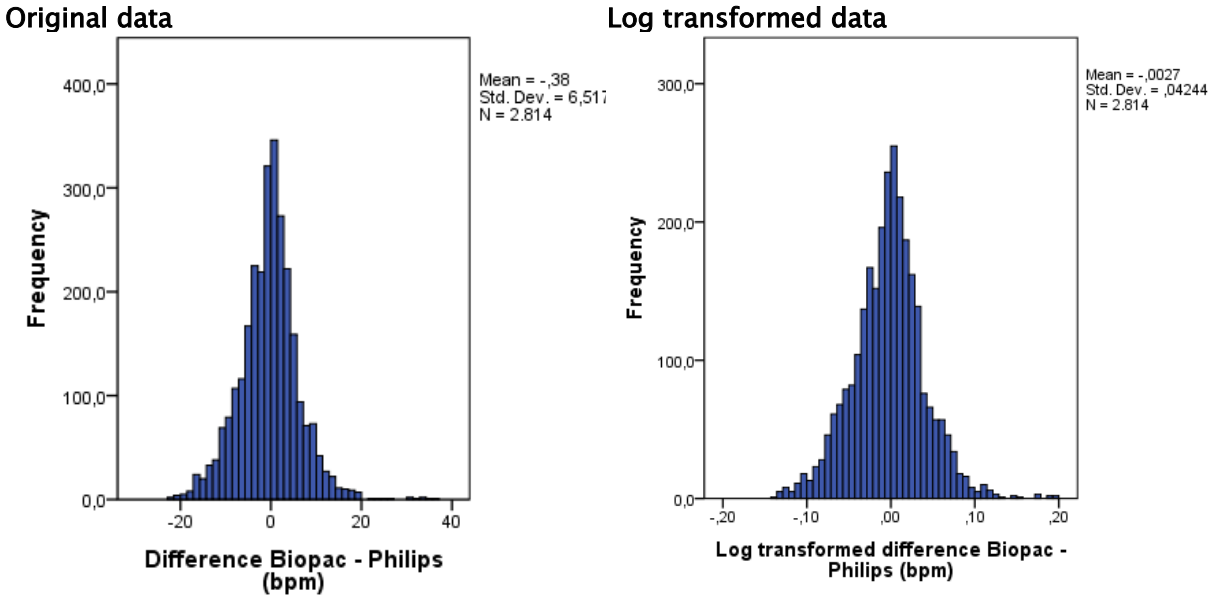
Appendix I. Distributions of original and log transformed Biopac-Philips differences per participant

Histograms to show the distribution of the original and log transformed differences.

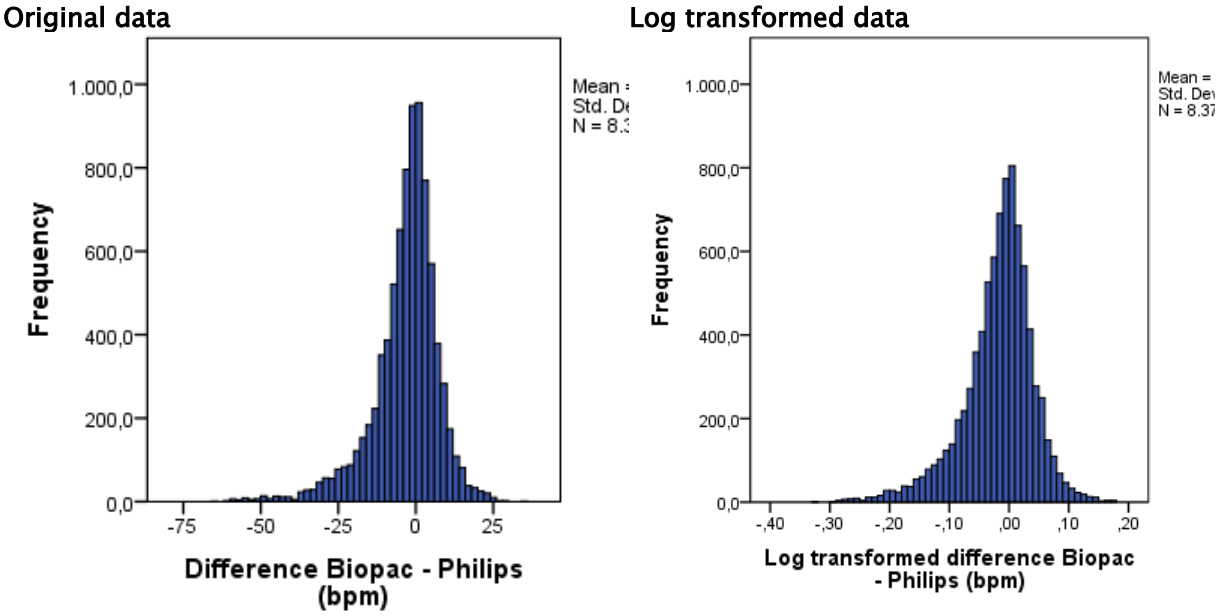
Participant 9



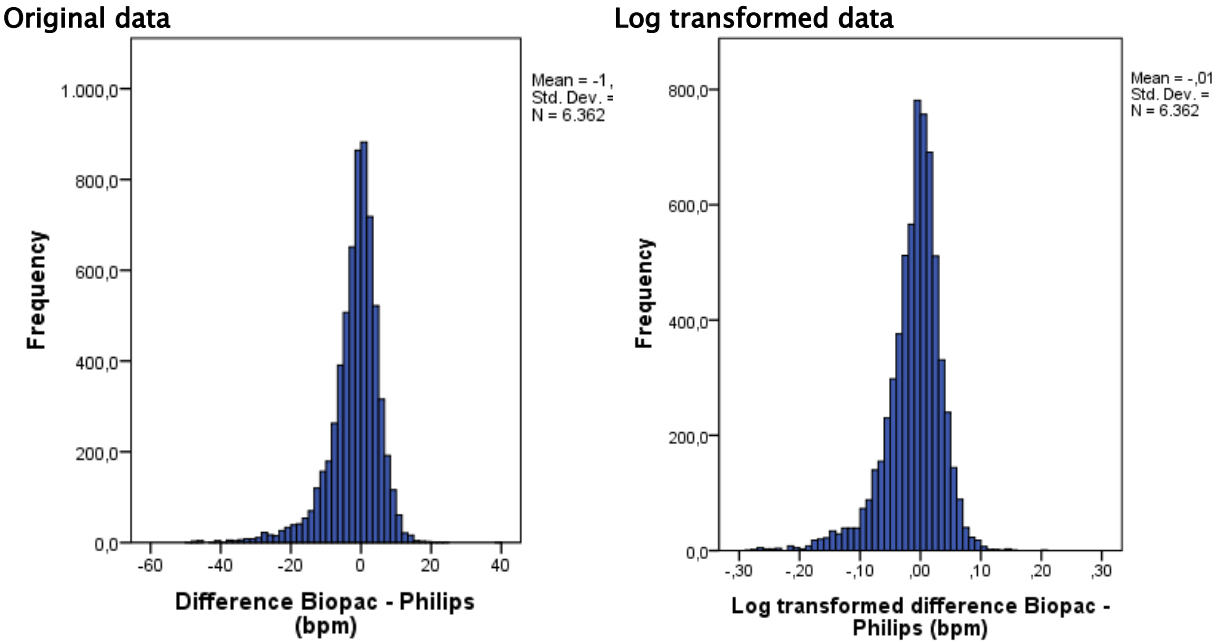
Participant 10



Participant 12

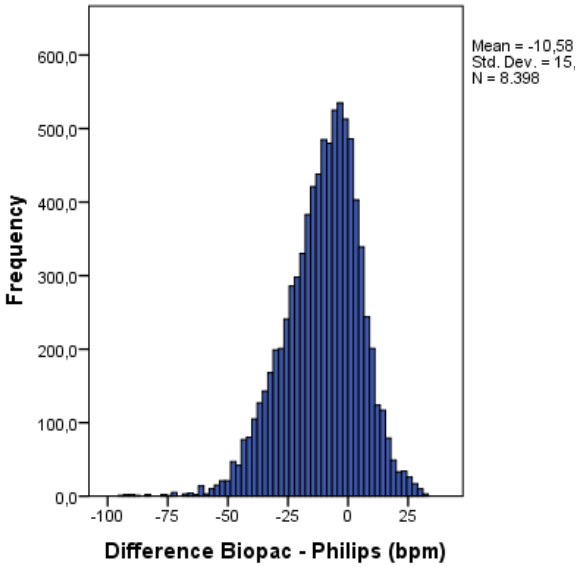


Participant 14

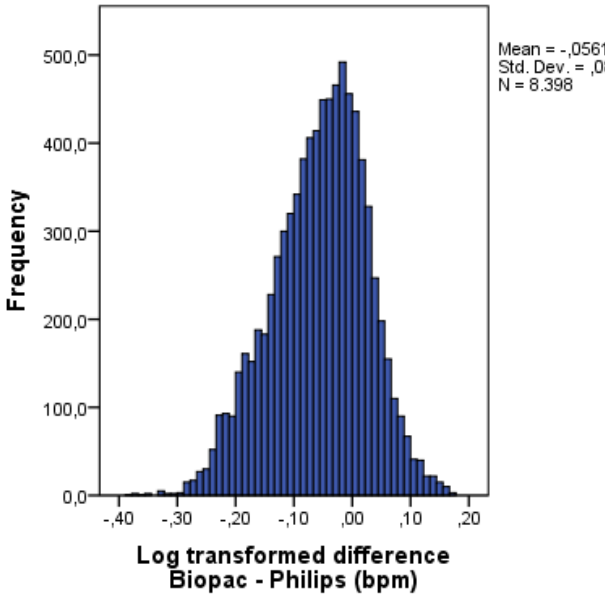


Participant 20

Original data

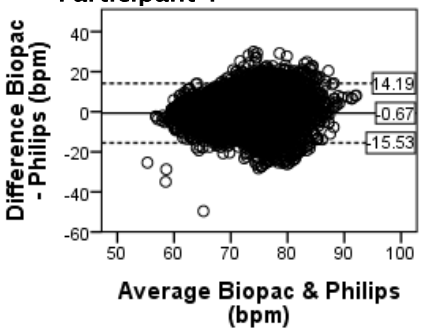


Log transformed data

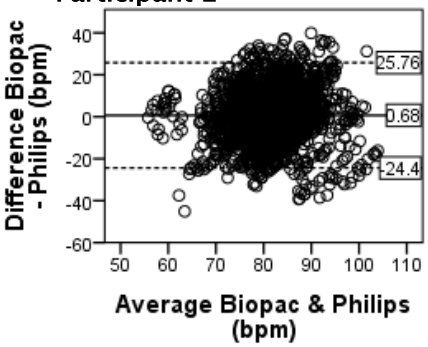


Appendix J. Bland-Altman plots per participant of the Philips data

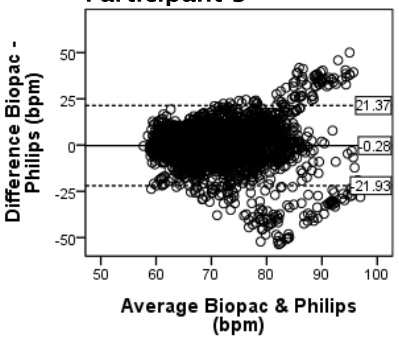
Participant 1



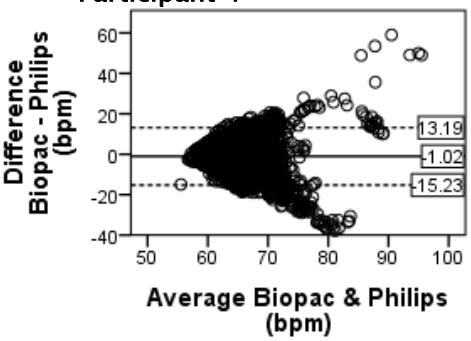
Participant 2



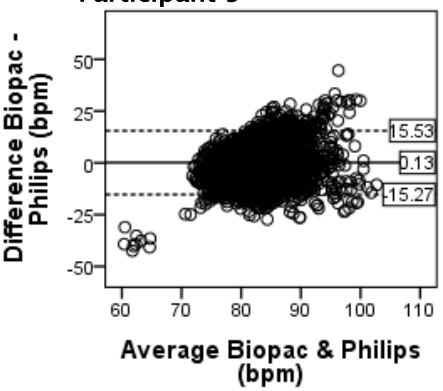
Participant 3



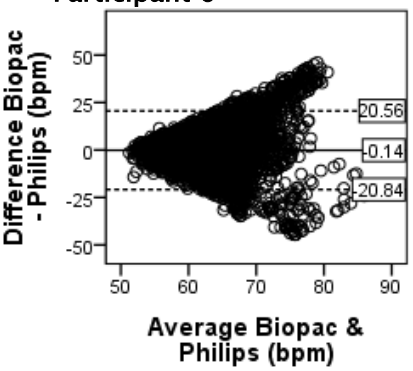
Participant 4



Participant 5



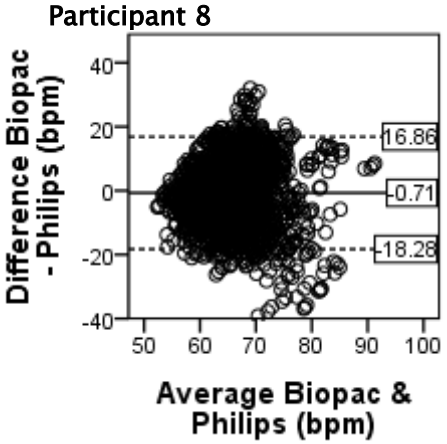
Participant 6



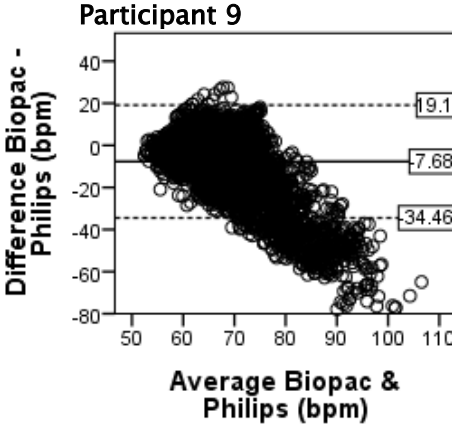
Participant 7

No Philips data

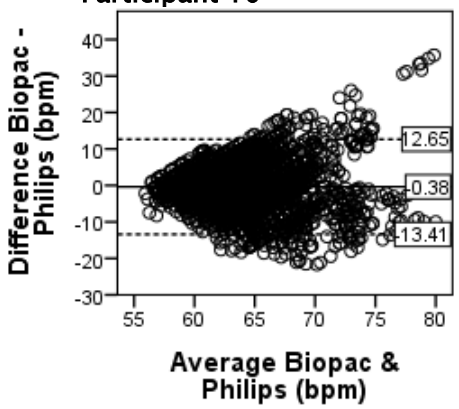
Participant 8



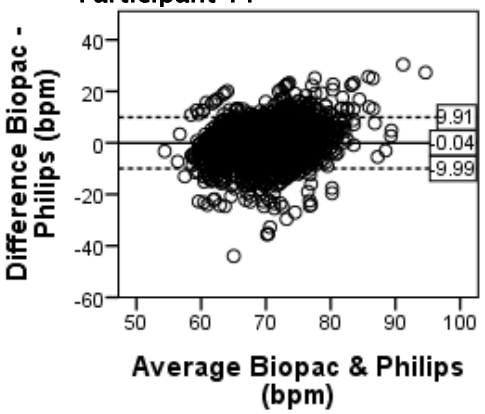
Participant 9



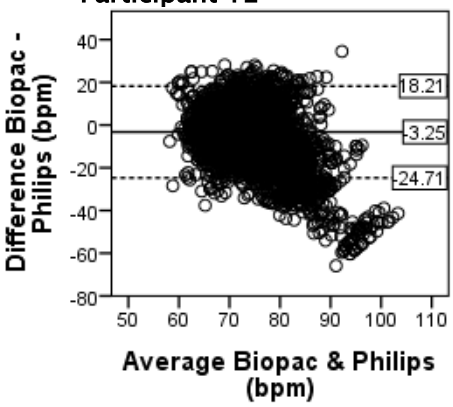
Participant 10



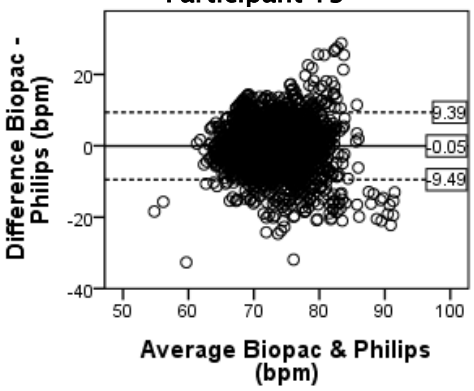
Participant 11



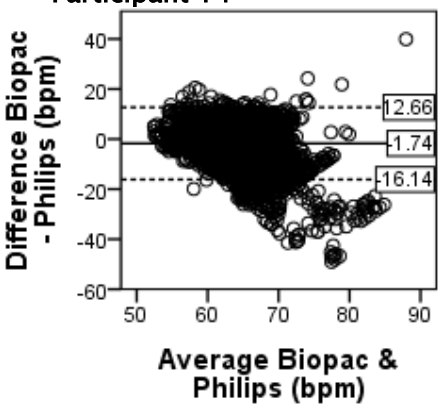
Participant 12



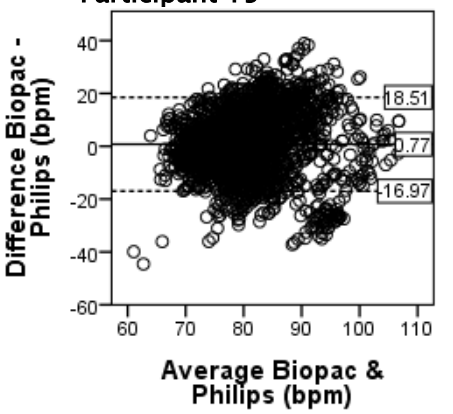
Participant 13



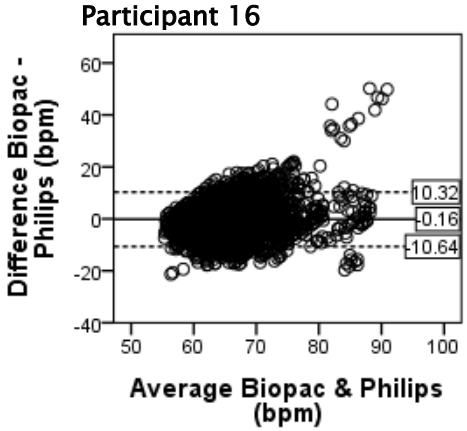
Participant 14



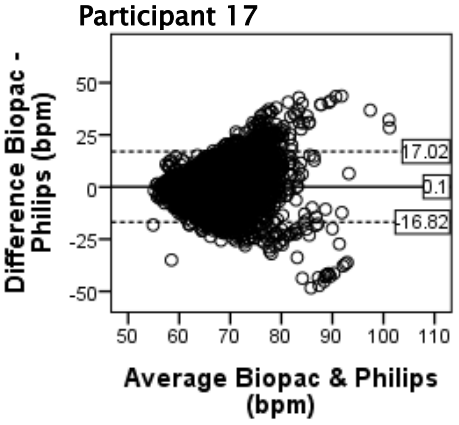
Participant 15



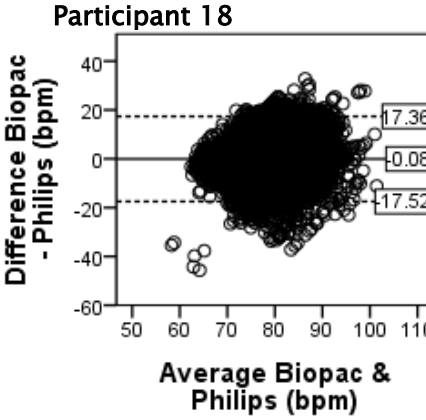
Participant 16



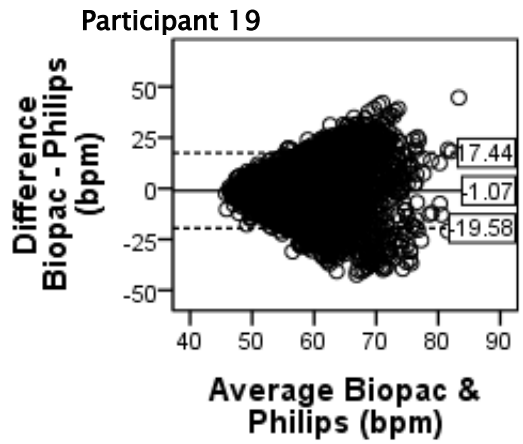
Participant 17



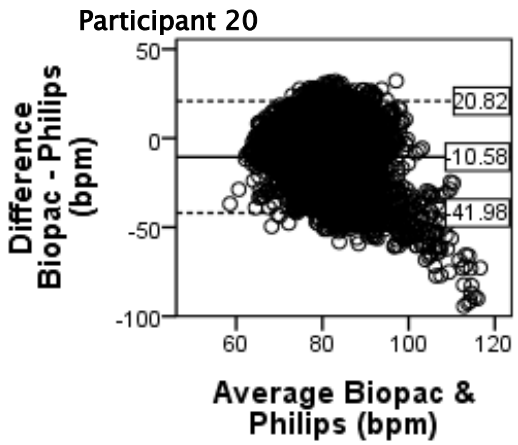
Participant 18



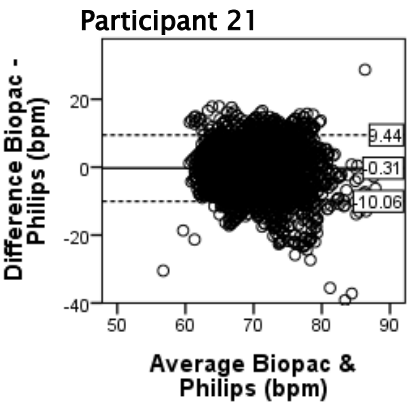
Participant 19



Participant 20



Participant 21

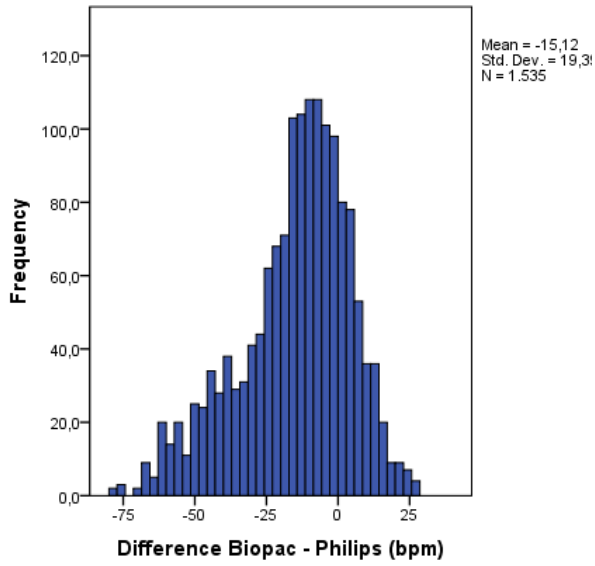


Appendix K. Bland-Altman plots per quality scale

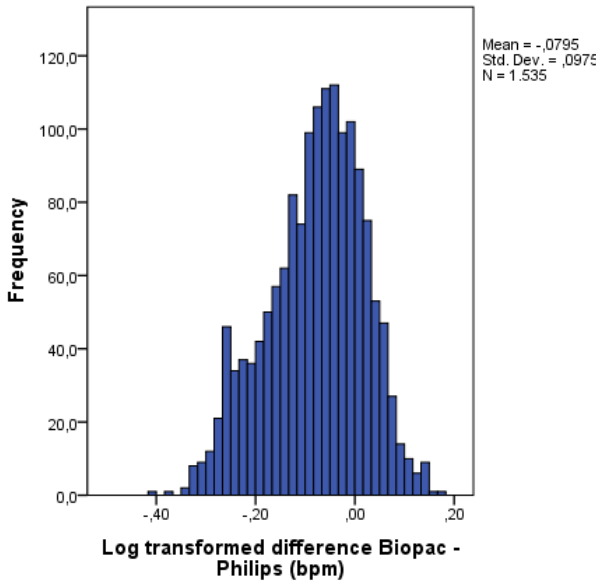
Normal distribution of the differences and Bland-Altman plots of Biopac and Philips data.

Quality = 0

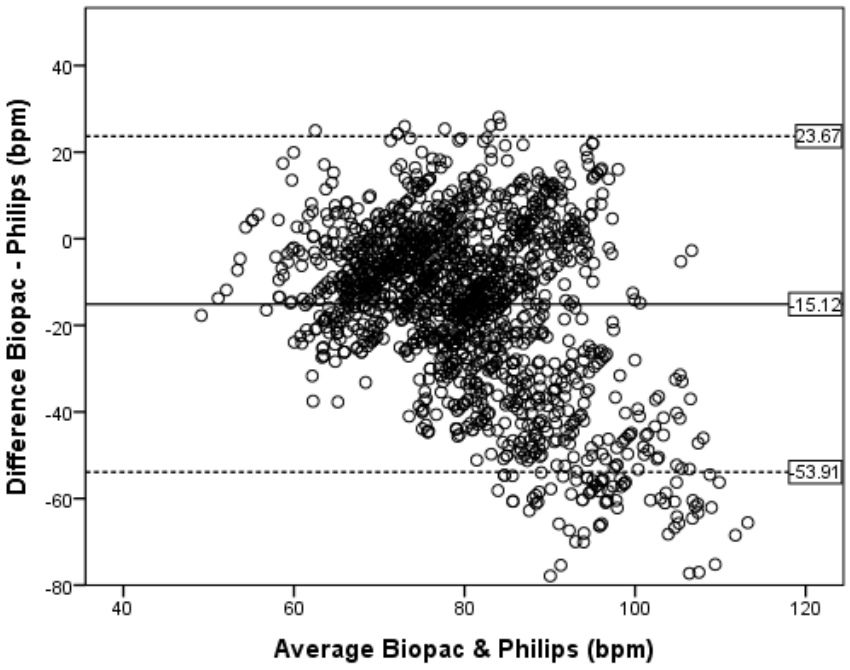
Original data



Log transformed data

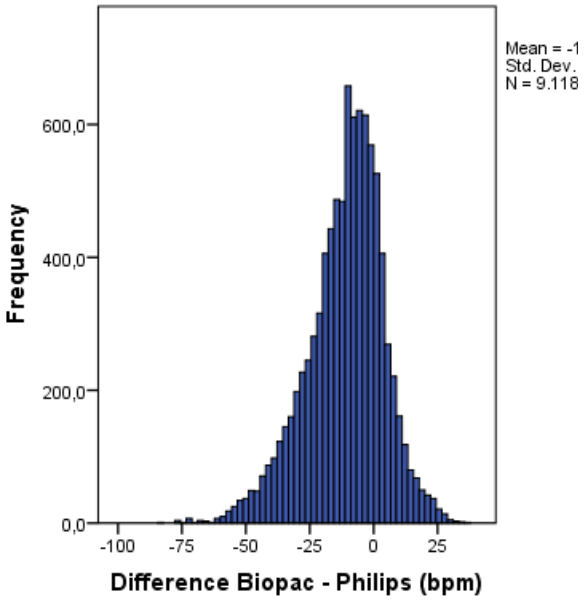


The differences are also not normally distributed with the log transformed data, so the original data is used for the Bland-Altman plot.

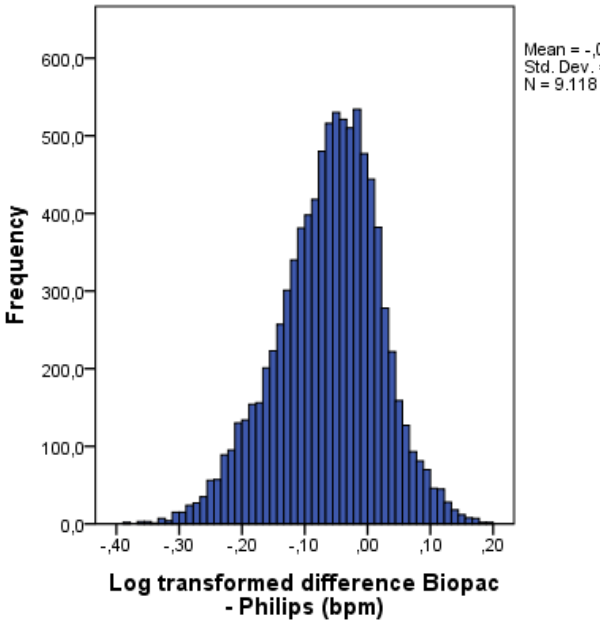


Quality = 1

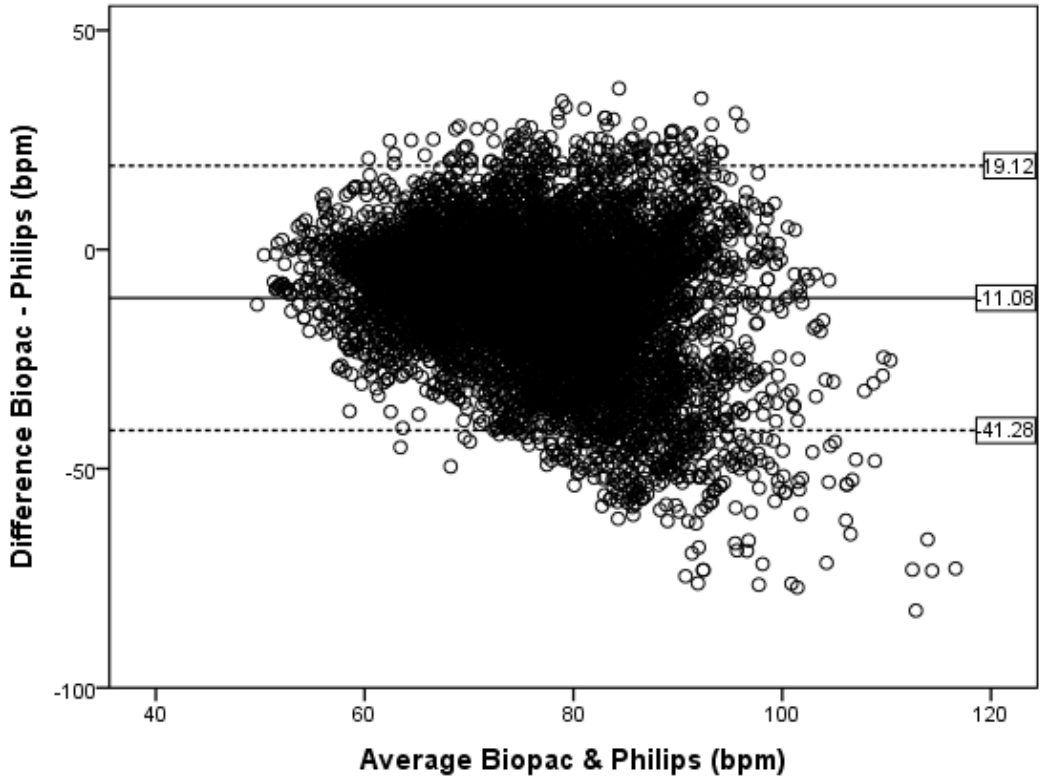
Original data



Log transformed data

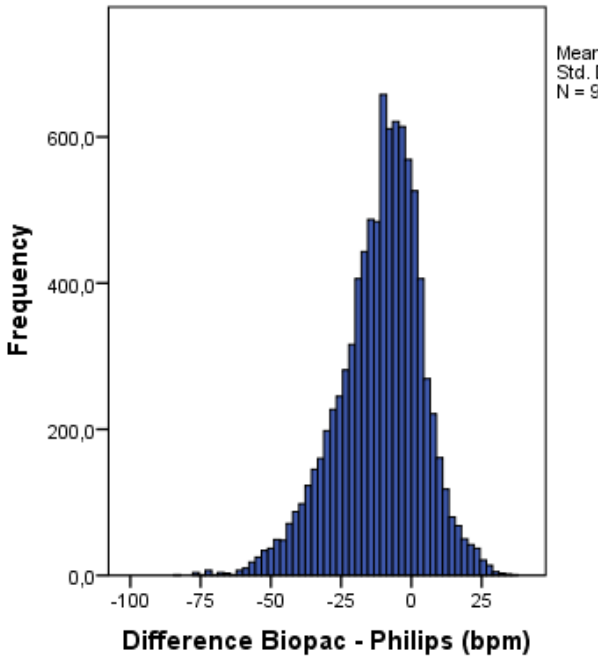


The differences are not normally distributed with the log transformed data, they are skewed to the left, so the original data is used to make the Bland-Altman plot.

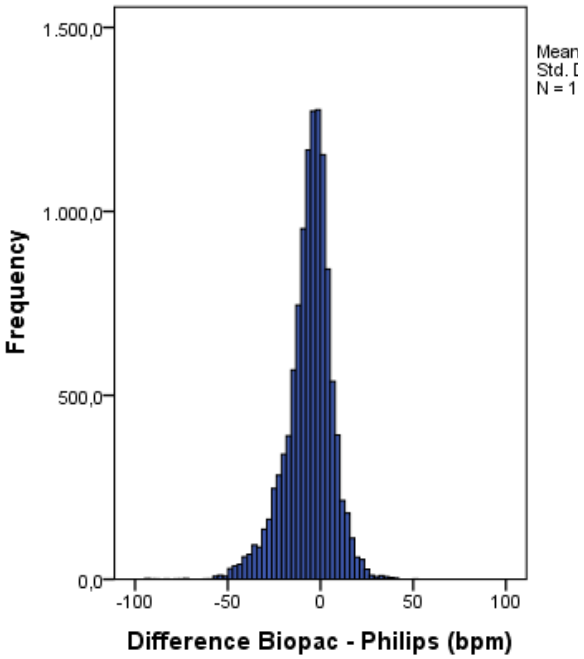


Quality = 2

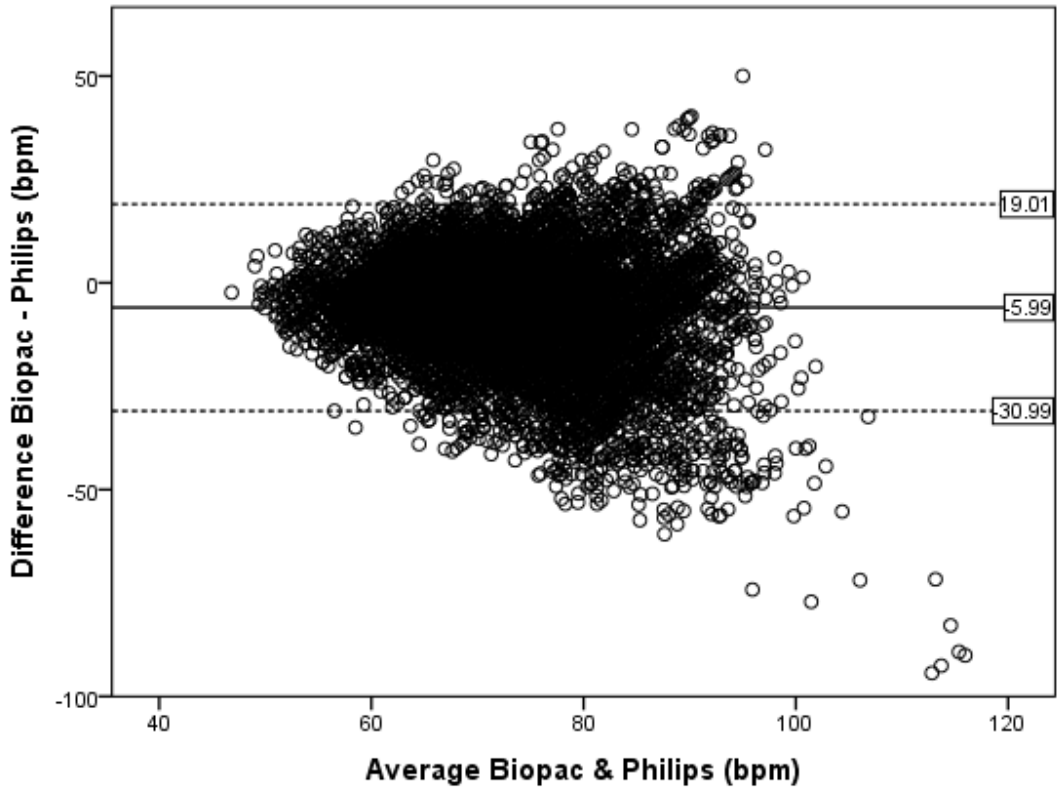
Original data



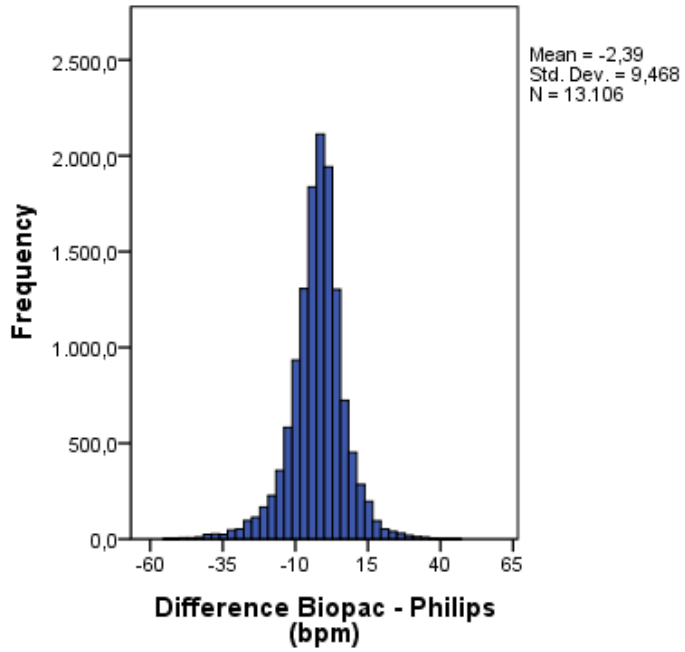
Log transformed data



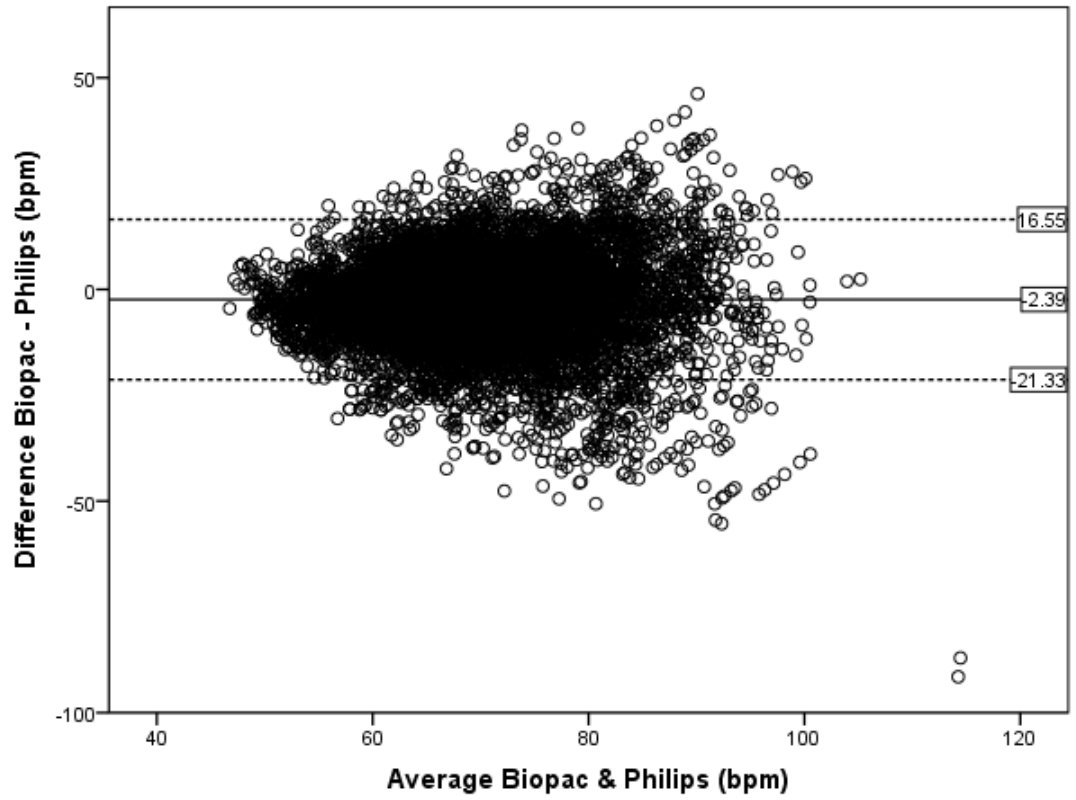
The differences are not normally distributed with the log transformed data, they are skewed to the left, so the original data is used to make the Bland-Altman plot.



Quality = 3

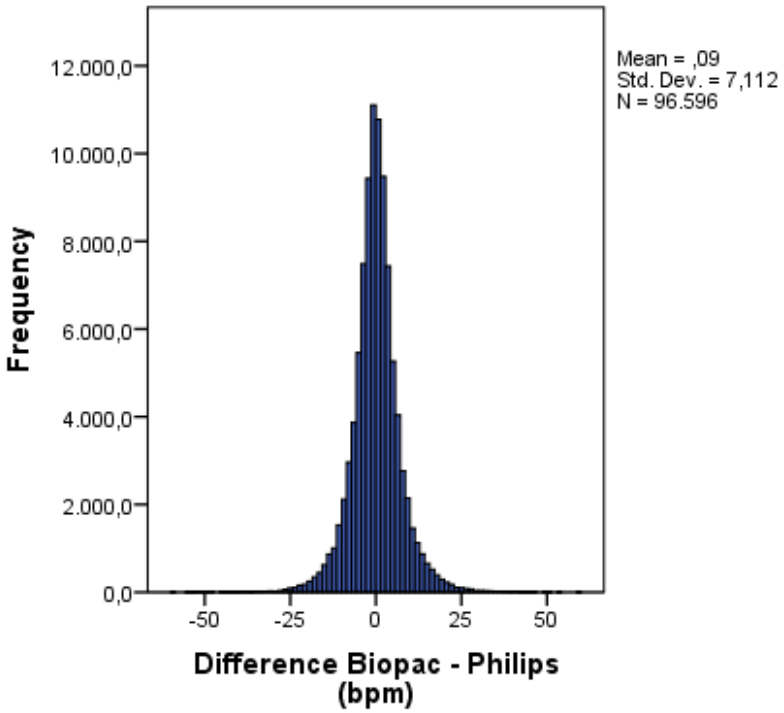


The differences are normally distributed.



Quality = 4

The differences are normally distributed.



The differences are normally distributed.