Non-Invasive Benchmarking of Pulse Oximeters - An Empirical Approach

Procedures, Considerations and Limitations of Testing Health Sensor Platforms

Kenneth Aune Frisvold

Thesis submitted for the degree of
Master in Programming and Network
60 credits

Department of Informatics
Faculty of mathematics and natural sciences

UNIVERSITY OF OSLO

Autumn 2018

Non-Invasive Benchmarking of Pulse Oximeters - An Empirical Approach

Procedures, Considerations and Limitations of Testing Health Sensor Platforms

Kenneth Aune Frisvold

Non-Invasive Benchmarking of Pulse Oximeters - An Empirical Approach

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

Abstract

Available on the internet today, there exists a increasing amount of cheap gadgets and sensors that can be used for medical purposes. However, standardized methods for determining the quality of the sensors are often expensive and require special expertise. The potential high cost for testing and implementing these sensors into medical use is a obstacle for speeding up the diagnosis of well known and easy identifiable disorders such as obstructive sleep apnea(OSA). The traditional method of determining the quality of pulse oximeters includes for subjects to breath gas mixes, and analysis of blood drain from the subjects by a CO-oximeter. Instead, we design a non-invasive breathing script to guide subjects through a series of breath hold from functional residual capacity (FRC) while breathing room air. Then we compare the resulting SpO_2 values from the low-cost oximeter against a more expensive reference oximeter. In this thesis, we compare Cooking Hacks MySignals (CH) and BITalino pulse oximeters against NOX T3 Sleep Monitor (NOX) as the reference oximeter. We calculate the industry standard metric *accuracy* (A_{rms}), and perform a Bland-Altman analysis to find the *precision* (standard deviation of the difference) and *mean bias* (mean of the difference). In addition to the well known analysis method, we also perform a simple apnea detection analysis to decide the oximeters ability to detect the fall in arterial oxygen saturation (SaO_2) associated with sleep apnea, with NOX as the reference oximeter.

For CH, 3250 matched samples over a period of 19 minutes were obtained and paired with NOX for each from 10 test subjects. Results show that the accuracy is 1.34%, with NOX as the ground truth (or 3.34% including the NOX accuracy), in SpO_2 values between 100% and 70% (>90% of the values was spread between 95% and 100%). The mean precision of all subjects is 1.78, and combined results give a precision of 2.61. Mean bias is 0.14%. Further, of the total 79 desaturations recorded by CH, 88.7% is true positives, 15.2% is false positives, and 11.3% is false negatives. For BITalino, we were not able to determine the quality of the pulse oximeter. The collected data contained a perturbation pattern, affecting the signal, that we did not find the source of. Our research suggest that low-cost pulse oximeters might be suitable for detecting desaturation associated with sleep apnea, and it is possible to determine the quality of oximeter for such use by using the non-invasive methods mentioned in this thesis.

Acknowledgement

First and foremost, I would like to thank my supervisors, Professor Thomas Peter Plagemann and Stein Kristiansen, for all the advices and countless discussions helping me in this project. Their dedicated guidance made sure that I always stayed on the right path. I will also thank all the subjects volunteering for this project, you were a crucial part of the accomplishment of the research. And last, but not least, I will thank my friends and family, especially my girlfriend Ada and my two children Eila and Balder for all their support and patience throughout my studying years.

Contents

List of Figures

List of Tables

Part I

Introduction and background

Chapter 1

Introduction and motivation

1.1 Introduction

Health care monitoring has traditionally been reserved for hospitals and health clinics, i.e., places where the medical expertise is located. This situation has resulted from the earlier relatively high price of health sensors systems. Lately, the development of smartphone technology has enabled small sensing devices and sensors to be connected to portable computers. Simple health monitoring has been implemented for some years in apps on smartwatches and smartphones, which are using sensors such as pedometers or accelerometers to track or measure physical activity. Recently, however, a growing number of portable health sensor devices have emerged to record and measure the metrics used in diagnosing more advanced physical health conditions, such as the respiratory patterns or blood oxygen values used to detect sleep apnea. Such recordings can also be a good supplement for medical doctors, since they allow a fairly inexpensive monitoring of the patient at home. Proper software recordings can also serve as health safety monitors for individuals who are sick, physically disabled, elderly, etc.

Projects exist that take advantage of this "revolution" in the low price and mobility of the new market of health sensor platforms, which we introduce below. Many of these sensors are not certified for clinical use. Measuring tools are therefore needed to determine the quality of these sensors and their value for the intended use. The CESAR project aims to develop a tool for the diagnosis of obstructive sleep apnea (OSA). OSA is a sleep disorder caused by partial or complete blockage of the respiratory passage. The gold standard for sleep studies and diagnosis of OSA is by an overnight sleep study known as polysomnography (PSG). In PSG the patient is attached to a various number of sensors by medical personnel or specialists, and stays overnight at a laboratory to have his or her sleep recorded. Afterwards the result, called a polysomnogram, is analysed by medical personnel who score sleep apnea events and the degree of severity of the disorder. Because of the nature of the study, the process of diagnosing a patient for OSA through traditional PSG is fairly expensive, and it may also be experienced as intrusive.

Nevertheless, the consequences of remaining undiagnosed may lead to both mental and physical illnesses. While there are known negative health consequences of sleep disorders (from subtle consequences such as sleepiness and decrement in mood and quality of life, to the more harmful hypoxia, cardiac dysfunctions or death), estimates show that most occurrences remain undiagnosed and that the prevalence is increasing [33]. In Norway it is estimated that about one in six persons suffer from the sleep disorder[21], and indications are that as many as 70-80% of those affected remain undiagnosed[37]

The growing number of health sensors, which vary in both price and quality, raises the question of their value in the monitoring and diagnosis of patients and disorders. If possible, the use of more inexpensive sensors would also lower the threshold for implementing more use of home monitoring in the health sector.

1.2 Problem Description

The CESAR project aims to improve home monitoring and diagnosis with the use of low-cost sensors. Patients are monitored at home by a private market pulse oximeter, unattended by health personnel. The records of the night sleep may reveal abnormalities in their sleeping patterns. The doctor of a patient could potentially identify a sleep disorder based on the recorded data, but only if they are of sufficient quality. If high-quality data are analyzed, the doctor has a good foundation to evaluate whether the patient should be referred to a specialist in sleep studies for further diagnosis.

Pulse oximeters for private markets are often very inexpensive in contrast to medical-grade oximeters. They are increasing in number, and doctors, patients, researchers or developers can buy them at a lower cost. However, the low price introduces a question of whether the sensor is suitable for use in a medical setting. Even though the manufacturers often specify the quality of their sensors according to international recommendations, we can assume, if not otherwise stated in its documentation, that a sensor is probably not clinically certified or tested by an independent actor. Testing in a professional laboratory is expensive, and the methods are often intrusive. The industry standard for measuring the quality of pulse oximeters is generally to use CO-oximetry or other spectral analysis of blood drain, therefore requiring medical attendance and expertise.

As just described, to expand our knowledge of the quality of inexpensive oximeters, we either have to request a potentially expensive laboratory study or implement invasive testing procedures requiring medical equipment and personnel. As a result, from initially being a low-cost sensor, testing the quality might raise the expenses to a total where the solution is no longer a low-cost, first-step alternative to more standardized diagnosis tools. The intended low threshold for buying inexpensive medical equipment (e.g., for use in the home monitoring of patients) is therefore

undermined by the total cost of the implementation.

It is possible to use an oximeter without a quality check, although omitting a third-party examination of medical equipment might lead to unfortunate consequences. The use of equipment with inadequate quality assurance might give wrongful indications about a person's physical condition. Results from such sensors may falsely support or disprove medical health assumptions, thus causing unnecessary expenses from either an extended sleep study or, even worse, false conclusions found that no further diagnosis is needed.

Based on the challenges described, therefore, this thesis addresses quality testing of pulse oximeters in the setting of apnea detection, using a non-invasive method and with no additional equipment or medical supervision needed. However, even though we focus on oximeters in this thesis, other sensors might be used in combination with them. In addition to the sensors, our computer science lab is equipped with other common technological devices, such as computers and smartphones. Accordingly, by limiting the need for resources, and by only including the equipment mentioned, we contribute in lowering the financial expenses and limit or obviate the need for medical expertise to evaluate the quality of oximeters. With this strategy, we hopefully also lower the threshold for buying, testing and using inexpensive physiological sensors.

1.3 Claims

The main work of this thesis is the design and evaluation of the use of a noninvasive benchmarking protocol, i.e., a testing procedure, to test the quality of pulse oximeters. Our work has been developed as an inexpensive alternative to the industry's standardized testing methods, which require medical attendance. Therefore, our protocol is an easy-to-follow, step-by-step manual that can be used as a guide when benchmarking pulse oximeters, with no need for medical attention or equipment. It includes the fundamental considerations and precautions for implementation, in addition to the specific testing procedures.

This paper also covers the implementation of the researched benchmarking protocol, and we complete a quality study of pulse oximeters from the mentioned BITalino and Cooking Hacks. Results from experiments are analysed, and we calculate the accuracy (root mean square error), precision (standard deviation of the difference), mean bias (mean of the difference) and limits of agreement of the oximeters. We also provide statistical data of their ability to correctly identify the desaturations recorded by the reference pulse oximeter.

1.4 Approach

As our research depends on technology that measures physiological factors, this paper contains a survey of the Health Sensor Platform domain. We also present the process of pulse oximetry and the theory and diagnosis

of sleep apnea. In addition to surveying the health sensor and sleep apnea domains, our literature review is rooted in these three points:

- What are the standardized methods for establishing the quality of pulse oximeters, and are there alternatives?
- Are there any challenges in recording the brief physiological event of apnea with a inexpensive pulse oximeter?
- What dependencies determine the success grade of our benchmarking protocol, and how can we improve it?

To address the first point, we examine both accepted standards and related work on quality testing of pulse oximeters. This investigation provides us with knowledge of possible alternatives to standardized methods. For the second point, we discuss pulse oximetry as a technology, and the possible challenges attached to an oximeter's capability to measure desaturations associated with sleep apnea correctly. Last, in addition to exploring sleep apnea, we look into papers that investigate the use of awake apnea simulation as a method to achieve desaturations. This step also gives us an impression of what to expect from different levels of oxygen saturation.

As mentioned, we have three sensor kits at our disposal, from Nox Medical, Cooking Hacks, and BITalino. A medical-grade home monitoring set, Nox Medical is used as the reference monitor. Our testing procedures start with exploring the abilities of the platforms mentioned empirically in a series of preliminary experiments. We determine methods for data acquisition and synchronization, inspect the sensor data and discuss the need for data filtering and processing. For BITalino, related work in the CESAR project has already developed and implemented a data acquisition method, and the quality of the data is therefore the main topic of discussion. The platform from Cooking Hacks, MySignals, is new to the CESAR project, and it is therefore explored more thoroughly than the other two platforms. We examine the relevant code from documentation in detail, and write and implement new code that fits our purpose better. Then, based on our findings and the background material, we identify requirements and design the benchmarking protocol. The evaluation of our research is accomplished through a series of experiments. Introduced not merely to determine the quality of the pulse oximeters, the tests also provide data we use when evaluating our protocol. The experiments are therefore divided into two parts. First, we run two tests with different test subjects, and then we evaluate if the benchmarking protocol is an object for optimizations. The second part continues with the full set of experiments, using the now- updated protocol. We summarize experimental results and analyze characteristics in data, in addition to providing a final quality statement about the oximeters tested. Conclusions about our non-medical, non-invasive benchmarking protocol are evaluated against our expectations and goals.

1.5 Paper Organization

After the introduction in Chapter 1(including problem statement, claims and approach), our research is presented in Chapter 2 with a general introduction of the platforms and sensors available to us. There we also look into technologies and terms used in this paper. Then, in Chapter 3, we cover the basics of pulse oximetry technology, including physiological processes and challenges. We also discuss different methods for analysis of sensor data. The last chapter in the first part of this paper, Chapter 4, covers the taxonomy and diagnosis of obstructive sleep apnea. As an important step for our design of the benchmarking protocol, we investigate breath-held apneas.

Part II includes the design and implementation of our benchmarking protocol. First, in Chapter 5, we conduct a series of preliminary experiments to establish methods of data acquisition, and investigate the issue of data quality. There we also test different methods of synchronization, and their possible consequences. In Chapter 6 we, include the requirements for our benchmarking protocol and define the limitations and scope of our tool. Next, methods for determining its quality are established. We also define the purpose and methods of the benchmarking protocol. The design of our benchmarking tool is presented in Chapter 7, as well as the considerations, preparations and guidance pertinent to the benchmarking process. Chapter 8 contains the implementations of the scripts needed in the benchmarking process.

Evaluation is the topic of Part III. Chapter 9 contains the experimental results. Information is included about the test population and synchronization, and the quality of the pulse oximeters from BITalino and Cooking Hacks are determined. In addition to calculating the values from our defined metrics, we determine the ability of the pulse oximeter from CH to record desaturations in Chapter 10, which includes an investigation into the classification failures. And last, Chapter 11 consists of discussions about the design of the benchmarking protocol, and how well suited our benchmarking method is for determining the quality of pulse oximeters.

Part IV begins with a contributions summary (Chapter 12). Chapter 13 outlines several research challenges we faced, while Chapter 14 suggests relevant topics for future investigations. The last components of our research are Appendixes, where A contains source code location, B work on the Cooking Hacks platform, and C important documents for our benchmarking method.

Chapter 2

Sensor Technologies

As we will learn in this thesis, portable sensor devices now exist for monitoring patients outside of hospitals and laboratories. These small, portable devices can be useful as an initial investigation into a person's health, before referral to a specialist for further diagnosis. The more expensive devices are often certified for medical use by physicians and other specialists. Dentists today often use them to diagnose problems such as sleep apnea before installing oral appliances. An example of such device is the NOX T3 from Nox Medical[27]. Even though the device is less costly than those used at a clinic, and certainly more portable, it is still fairly expensive to purchase. Lately, an increasing number of inexpensive devices have emerged that include many different sensors. In contrast to the NOX T3, most of them are not certified for medical use or verified by a third party. Instead, these are development kits, which can be used to monitor physiological processes. We consider in our research both types of device as *health sensor platforms*, and we provide definitions later in this chapter. An example of an inexpensive health sensor platform would be the BITalino [6]. Apart from the price, the difference between the two is their purpose. NOX T3 is designed to perform sleep studies, and BITalino is, as mentioned, a development kit. In this chapter, we explore the subject of health sensor platforms, and further examine the ones mentioned.

In the next section, we establish a common understanding of terms and technologies introduced in this paper. In Section 2.2, we dive into the different health sensor platforms. Various types of sensors and their use is described in Section 2.3

2.1 Terms and Technologies

In this section, we go through the terms and technologies used later in this paper.

Gold Standard and Ground Truth: In medicine, the "gold standard" refers to the method proven to be the best practice for measurement, such as monitoring physiological processes. The "ground truth" is the reality of the situation; it is what actually happens. Then, the gold standard is the one method out of all methods that give the most accurate estimation of

the ground truth. New methods are often tested, or calibrated, against the gold standard.

Arduino and MCU: Arduino[1] is an open-source electronic platform that serves as a controlling and processing unit for sensors or boards. It contains a *microcontroller(MCU)*, which is a lightweight internal computer, with processing, storing, interfacing and communication capabilities. The Arduino platform also offers its own Integrated Development Environment(IDE), Application Programming Interfaces(API), and both wired and wireless connectivity abilities.

API: A common task for APIs is that they enable communication between software or hardware components. For instance, the Arduino boards have a set of C/C++ functions used for communications, defined in the language reference list. In our research, we use APIs to setup the devices and collect data.

Android and Bluetooth: Android[2] is an open-source operating system(OS) mainly used in mobile devices and smartphones. The advantage with Android, is that users can install own applications without any form for approval. Bluetooth is a short-range wireless protocol for the exchange of data between mobile devices. It is one of the most used communication technologies today, together with WiFi. We learn more about Bluetooth below.

Bluetooth Low Energy

Bluetooth Low Energy (BLE) is a version of Bluetooth designed especially for smart devices, and is a lightweight subset of classic Bluetooth [46]. Because of the low energy use, BLE is often implemented in small devices, such as portable devices and sensors. The basic characteristics of BLE technology and its use are explained below. *Generic Access Profile (GAP)* is what defines and controls how to connect with the BLE unit. There are two types of devices, the peripheral (slave) and central (master). To simplify, we can think of these two as the sender and receiver of sensor data, respectively. GAP handles the advertising payload of the device, which can be either advertising data or scan response. Both payloads contain up to 31 bytes of data: the only difference is that the scan response payload can contain additional information about the peripheral device. The scan response payload is a result of a scan response request from the central device. After the central unit and the peripheral unit establish contact, the devices can then start to exchange data. The peripheral unit cannot then connect to other central units until connection is broken.

The Generic Attribute Profile (GATT) defines the way the BLE units transfer data. GATT can be seen as a server/client relationship. A GATT

[1] http://www.arduino.cc
[2] http://www.android.com

transactions contains *profile*, *services* and *characteristics*, which can be stored in the master device. A profile can be seen as a frame for the data transaction, and it involves standardized profiles (e.g. pulse oximetry). Each profile consists of one or more services that are used to break data into chunks of data, called characteristics. Both services and characteristics distinguish themself by *universally unique identifiers* (UUID).

2.2 Health Sensor Platforms

In this paper, we use the term *Health Sensor Platform* (HSP) to refer to devices with sensors that record physiological data, and most of these platforms usually come with software. To determine what is HSP's, and to exclude other sensor devices, we now provide a definition. Since to the best of our knowledge no such definition exists in the literature, we propose the following: *A Health Sensor Platform is a combination of hardware and software, including APIs, that provides digital data from health sensors*
Therefore, an HSP is the hardware we use to connect to sensors, along with the accompanying software or APIs that enable extraction of data with health value. While all the platforms in this paper can provide us with the physiological health data of a person or a patient, the XeThru platform we describe in Section 2.2.5 is also used in many other nonmedical tasks, such as position or movement detection. Its sensors lack the specific purpose of measuring physiological signals. With the definition above in hand, we can therefore argue that the XeThru is not an HSP. The next sections explore the classification of sensors, comparing those used in this essay in detail.

2.2.1 Type of sensors

In this paper we use the term health sensor (further defined in Section 2.3) to cover all devices that can measuring a person's physiological processes. These devices differ in complexity and placement. As the goal of our thesis is to develop an noninvasive method for testing pulse oximeters, it would be useful to investigate the different levels of sensor invasiveness. First, we can examine the Oxford Living Dictionary's definition of invasive of medical procedures as "involving introduction of instruments or other objects into the body or body cavities" [13].

A sensor's degree of invasiveness can therefore be determined by whether it is inside, going into, or completely outside the body. We can also see invasiveness in relation to intrusiveness. For example, we can assume most people would regard an invasive sensor requiring access to the arteries for blood drains as intrusive, as both the procedure and the environment of operation, could be experienced as unpleasant.

Da Silva et al. define categories of hardware devices (health sensors) used to monitor the health condition patients or subjects [42]. We use their definitions to classify and discuss the different types of health sensors.

- In-the-person - Covers implantable health sensors such as pacemakers; often involves an operation followed by a hospitalization. The

location of this type of sensor is mostly invasive, and the implementation process is often experienced as highly intrusive.

- On-the-person - Covers stationary and ambulatory devices. Often large and used at clinics and hospitals, stationary devices record signals from a person through wires to a stationary recording/processing unit. Ambulatory devices can be used with portable units such as a smart clock or "necklace" that is implemented either in the unit or connected through wires or Bluetooth. These kind of health sensors we define as non-invasive, as they are not inside the body. However, they are attached to the person, and are therefore experienced as intrusive. Devices in this category are the platforms NOX T3 and BITalino mentioned earlier.

- Off-the-person - Covers devices with contact-based sensors. Instead of being worn on or within the person, the sensors are implemented in everyday life gadgets such as a gaming control or a keyboard. Other devices covered by this category are those that never touch the user or patient. The presence of these kind of sensors may not experienced by the monitored person at all; therefore, they are both non-invasive and experienced as non-intrusive. An example of a platform in this category is XeThru, which uses a contactless radar sensor.

2.2.2 BITalino

BITalino profiles itself as a low-cost, do-it-yourself toolkit, that can be used for developing health care applications. However, the platform works out of the box, and real-time data streams can be visualized by using their free software, OpenSignals. As the data are also made available through APIs in different programming languages, you have the ability to write your own software and to stream the recorded data (e.g., to an Android device). The price of the various kits ranges from 150 to 200€, including all sensors and cables. It is the least expensive of the platforms mentioned in this paper.

Technical Description

Shown in Figure 2.1 are the board components of the kit used in this thesis: the (r)evolution Plugged Kit BLE. It comes with about 10 sensors and actuators, cables, technology blocks and a battery. It is also possible to connect one's own developed sensors.

The device does not have storing capabilities. It is therefore necessary to connect it to a computer(e.g., laptop or smartphone) to extract or visualize the data. The previously mentioned OpenSignals software includes both an interface to connect to the BITalino board and options for visualization of recorded data. Even though the kit contains multiple sensors, it does not include a pulse oximeter. Instead, a separate sensor is obtained, the Contec CMS-50+. The pulse oximeter also comes with a cable designed to fit the BITalino boards. The cost of this oximeter at a BITalino shop is 165€, about

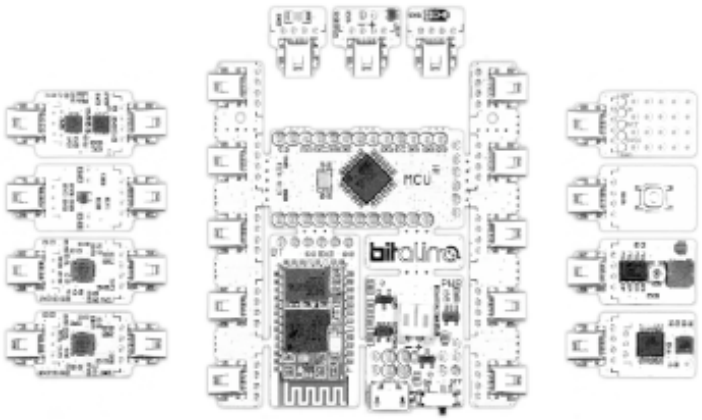

Figure 2.1: BITalino (r)evolution plugged kit

the same as the complete Plugged kit, and its functionality is not dependent on the board. It can record up to 24 hours of data, and configuring is set up with the help of its LED screen and function button. It is also possible to set alarms according to defined conditions.

Instead of using the OpenSignals software, we use a data acquisition tool developed earlier by Svein-Petter Gjøby [18]. This work uses Android apps to send commands to the BITalino, storing the data received from the sensors either on a phone or on an external database.

2.2.3 Cooking Hacks

Cooking Hacks is also a fairly low cost platform providing tools for developing health care applications and products. The kit used in this paper, MySignals HW BLE Complete, costs 1,350€.

In contrast to BITalino, MySignals, is more a sensor and technology board that serves as an interface for the sensors to be connected. It also includes a WiFi module, a Bluetooth Low Energy module, and a module to connect a TFT screen. Then the MySignals board is connected either to an Arduino, a Raspberry Pi or a Waspnote, all of which are computing devices It is possible to get a pre-programmed, all-in-one device, the MySignals SW box, which is ready to use with the MySignals App. The second version, MySignals HW, is a development version that is programmable without pre-installed software. We have the MySignals HW v2 that uses Arduino as a computing device. The kit is visible in Figure 2.2, with the MySignals Board to the left and Arduino to the right.

Technical Description

The platform provides a wide variety of both connectable analog and wireless BLE sensors and hardware, and it is possible to connect your own sensors. The kit includes a power connection; however, a battery may be connected to the Arduino board.

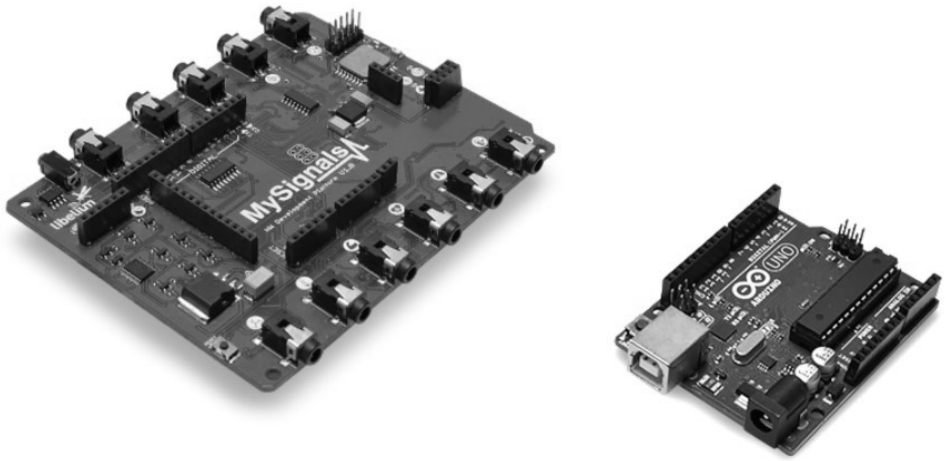

Figure 2.2: MySignals HW and Arduino Uno

The pulse oximeter from Cooking Hacks as a wireless BLE device, with no display or other buttons than an on button. Its values are available through visualization on the TFT screen, as seen in Figure 2.3, or through data collection in an external device. Neither the BLE profile nor any other documentation about the internals of the oximeter are provided by the manufacturers. However, through their forum they report the accuracy as $\pm$2% between 80 and 100%, and $\pm$3% between 70 and 79%

When using Arduino as processing board, MySignals can be programmed with the Arduino IDE on a computer. The documentation [19] contains code examples with many possibilities for collecting the data. These are explored further later in this thesis. It is worth noting that the documentation contains a capabilities overview and basic instructions; however, it lacks deeper technical explanations about how sensors operate, such as the internal protocol of the pulse oximeter. This platform is unique in that it can store data in the Libelium Cloud and give authorized access to it from a remote location.

2.2.4 NOX T3 Sleep Monitor

The NOX T3 Sleep Monitor from Nox Medical [27] is a portable home monitor device for sleep diagnosis. It is the most expensive of the platforms in this thesis, and in contrast to the two above, this one is medically graded. The price is more than 5,000€ .It is ready to use, with pre-programmed features for monitoring physiological signals. It is strictly portable, and the components seem very robust (whereas the two platforms mentioned above have open circuits etc.). It comes with a complete sleep analysis and diagnosis tool, Noxturnal, for analyzing the data produced.

Technical Description

The device consists of a central recording unit(Figure 2.4), a pulse oximeter unit, along with other sensors commonly used in sleep studies (more in

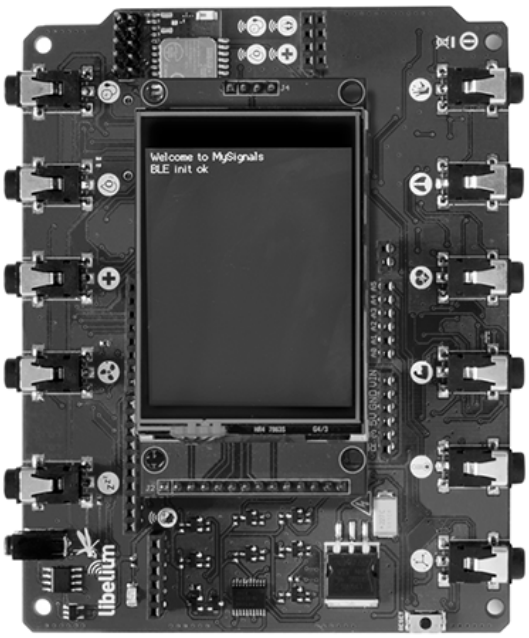

Figure 2.3: MySignals HW with TFT screen

Figure 2.4: NOX T3 Sleep Monitor[27]

Figure 2.5: NOX T3's Pulse Oximeter[27]

Section 4.3. All sensors are connected to the central unit, which both controls the recordings and stores the data from the sensors.

The pulse oximeter, Nonin WristOx2, is a wireless Bluetooth device that is automatically linked to the central device when the user inserts a finger into the clip. The oximeter has a separate finger clip (with the sensor) and computing device, although, they are linked via a short cable. The computing device has a screen that displays live data of SpO_2 values, pulse and technical information, while the clip is of flexible plastic for improved comfort. The accuracy of this oximeter is labeled in the technical description to be 2% between 100 and 70%. This pulse oximeter is displayed in Figure 2.5.

In order to collect data, the central unit is connected to a computer with a USB cable. To our knowledge, the only way to set up the device for recording, and to extract data afterwards, is by using the Noxturnal software(Figure 2.6). However, it is possible to extract raw sensor data from each channel using copy/paste. Otherwise, the data from a monitoring session are analyzed with the software, which includes common scores from classification systems in sleep studies.

2.2.5 Other Platforms

In our research, the time frame limited our possibility of exploring more platforms than those described earlier in this chapter. However, as an example of contactless sensor technologies, we can inspect the XeThu from Novelda.

As mentioned earlier in this essay, the XeThru stands out as an off-the-person platform, while the others are of the on-the-person type. It uses an ultra wide-band impulse radar as its sensor device, and the main uses are presence detection, respiration and sleep monitoring. Their software, Module Connector, is available for most operating systems, and is used for communication with the XeThru devices. Due to this platform's sensor

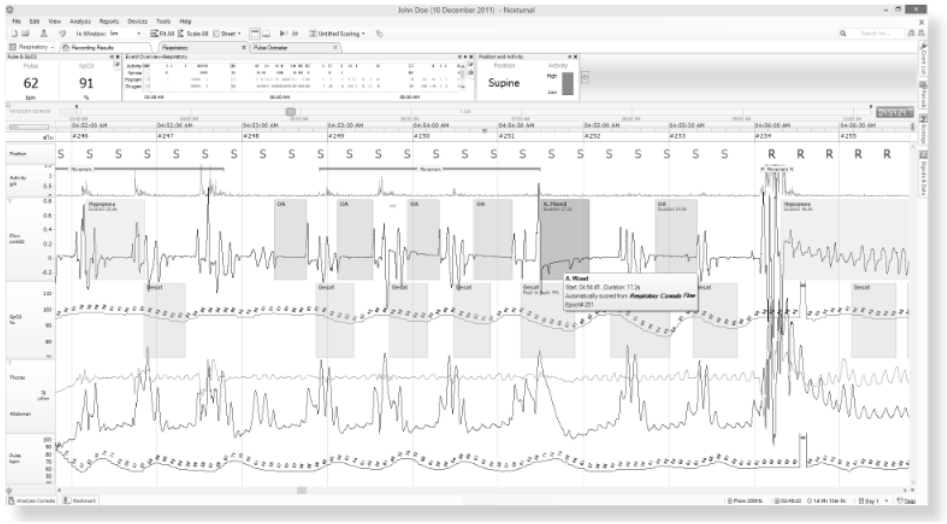

Figure 2.6: The Noxturnal Software[27]

technology, the data stream may not provide any health value without extensive processing with the provided software. In a blog post[3] on its website, it is claimed that XeThru can provide hospital-grade sleep data.

2.2.6 Summary of Platforms

Table 2.1 presents an overview of the platforms we used in our research. In general, the on-the-person platforms have approximately the same sensors available, and all but the NOX T3 has Bluetooth with an API available as an interface for data acquisition in real time. With NOX T3, it is only possible to extract the sensor data through their software, after the recording is done. With a programmable micro controller unit (MCU) only, the BITalino and Cooking Hacks (MySense HW v2) have little processing or storing capabilities on their board by default.

2.3 Health sensors

To obtain physical health recordings about a patient or user, we have to use sensors. These usually output raw data about a specific physiological process as a stream or record, and a combination or estimation of measures from more than one sensor may be used. To give the data meaning, the results from the recording are typically processed by software, then analysed by either health personnel and/or the software itself. In this section, we seek an understanding of common sensors used in health monitoring.

First, we differentiate between logical and physical sensors by using the definitions from Kristiansen et al. [43], which classify them based on

[3]https://www.xethru.com/blog/posts/xethru-delivers-hospital-grade-sleep-data, acc. 2017-10-2

Platform	Software	Internal MCU/CPU	Sensors	Communication interface	Storage
BITalino	API, OpenSignals	MCU	ECG, EEG, EMG, EDA, PO	Bluetooth 2.0	External
Cooking Hacks	API, MySignals	Connected Arduino UNO	ECG, EMG, GSR, PO, Airflow	Bluetooth, Cloud	Cloud, External
NOX T3	NOXturnal	Internal CPU	ECG, EEG, EMG, RIP, PO, Airflow	USB	Intern 1GB

Table 2.1: Overview of the health sensor platforms, key features and specification

input and implementation. A *physical sensor* converts an analog signal from the real world into a digital data stream. Implemented through software, a *logical sensor* analyzes sensor data from one or more data streams and produces a data stream as output.

In Figure 2.7 we see that output A is a result of one physical sensor. This sensor is an accelerometer, which is designed to output digital data based on acceleration input. Another example would be a digital thermometer, the input of which is the (analog) variable of temperature, and the output is temperature as degree Celsius.

The digital outputs B and C are results from sensors that process analog signals from electrodes. An electrode can measure electrical changes in the skin or body, and is useful for monitoring different organs. However, both the implementation and the interpretation of the data from electrodes may differentiate between (logical) sensors. Even though both electrocardiography (ECG) and electroencephalography (EEG) constitute methods for monitoring physiological processes, we call them sensors in this thesis.

As we pointed out, the terms sensors and methods are often used interchangeably, and we use sensors in our work. Below we show the most commonly used sensors of the health platforms named in this thesis. With this discussion we also learn about the importance and use of sensors in medicine.

- Electrocardiography (ECG) is used for measuring electrical activity

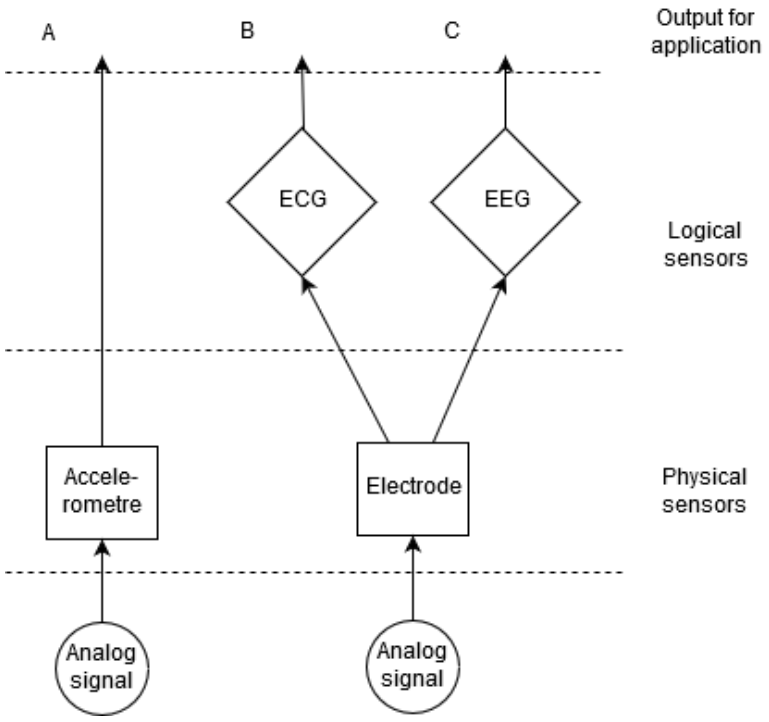

Figure 2.7: Example of logical and physical sensors

of the heart. In medicine, ECG is used in screening and diagnosis of heart conditions, such as myocardial infarction ("heart attack"). In short, a stationary, clinical ECG device is attached to the patient using electrodes, which is then able to record electrical activity of the heart, referred to as an electrocardiogram.

- Electrodermal Activity (EDA) / Galvanic Skin Response (GSR) is used for measuring electrical activity of the skin. EDA/GSR sensors are most commonly used in psychological research and therapy. They can record the electrical conductance of the skin (in practice, the moisture level), and the values measured can be used to indicate certain emotions.

- Electromyography (EMG) is used for measuring electrical activity produced by muscles. In EMG, we detect an electrical potential; that is, a voltage difference in two points of time generated by muscle cells. In medicine, the record, called an electromyogram, can be used to identify neuromuscular diseases. Both intramuscular and surface EMGs are available; however, the sensors available to us are surface EMGs.

- Electroencephalography (EEG) is used for measuring electrical activity of the brain. Placed along the scalp, electrodes measure voltage fluctuations in the brain. In medicine, the recordings can help to diagnose epilepsy and sleep disorders.

- Respiratory inductance plethysmography (RIP) is used for measuring movements of the chest. The procedure is done by placing one transducer band around the chest of the patient, and in some cases a second one around the lower stomach. RIP record analysis can help to describe various respiratory patterns and disorders.

- Airflow (nasal) is used for measuring the airflow and breathing patterns of a patient. Commonly, either thermocouple, GSR or nasal pressure sensors are placed in the nostrils to detect and measure changes caused by inhalation and exhalation.

- An accelerometer (ACC) is used for measuring acceleration. In medicine, this sensor is useful for recording positioning and movement, or nonmovement (e.g., to monitor a person's sleep behavior in a sleep study).

- A pulse oximeter (PO) is used for monitoring pulse, oxygen saturation in blood, and sometimes other physiological signals. The pulse oximeter is usually placed on a patient's finger, with a photodiode absorbing lights from LEDs. They are widely used in in sleep studies and in medicine to monitor respiratory and cardiac patterns.

2.4 Benchmarking Sensors

Benchmarking is the measuring of an object's performance, based on well-defined metrics, in which the result is often represented as a single value or a collection of values. These values should say something about the object's performance relative to other tested objects, a ground truth or best performance result. It is important that the testing is adequately documented, to make the results reproducible and comparable to other similar tests.

When we benchmark sensors, we may assume that output from one specific group of sensors from different manufacturers produce output that is similar in character and in most cases comparable. However, the format and the frequency of the output are likely to differ. As an example of the complexity of the process of benchmarking health sensors, we can inspect the experiment presented by Da Silva et al. [42]. They tested the performance of two contrasting devices from BITalino and Philips ECG, which differ in both price and technology. While the expensive Philips is a gold standard certified for clinical use, with almost a dozen electrodes, BITalino focuses on low cost and ease of use and has few electrodes. First, data acquisition methods for the devices must be established. As we described for the platforms at our disposal for this paper, the storing capabilities of sensor devices may vary. Next, because of their technical differences, preprocessing and synchronization algorithms are needed. The sampling rate may differ, or the values may not be directly comparable, need to be scaled/filtered, etc. Last, the data are analysed with metrics commonly used for signals from ECG sensors.

The example above contains the overall elements that are also found in our process of benchmarking. We can therefore identify these 4 objectives:

- Defining the benchmarking procedures.
- Determining suitable methods for data acquisition.
- Processing of data in order for the devices to be comparable.
- Analysis of data in regard to the appropriate metrics.

After completing the objectives and the benchmarking process, we have a good foundation on which to draw conclusions about the platforms tested.

2.4.1 Metrics

As we described in the last section, we should define the appropriate metrics to use in analysing the data. To better understand the term metric, we can inspect the description provided by Suri et al. [47]: "a standard of measurement stated in quantitative terms which captures the performance in relative to standard on the occurrence of event." Simply put, a metric is a quantification of an event that describes the characteristic of the measurement relative to the reference. The reference can be the ground truth or the gold standard. In our setting, an event can be a fall in oxygen saturation in a person, and the reference our expensive pulse oximeter. Then a metric should quantify the event (e.g., count the desaturation in the reference and the test oximeter).

2.4.2 Testbed and Data Acquisition

In software development, a testbed is a platform or setup to test new development in an isolated environment. Its purpose should be suited to the situation or item being benchmarked. If we are to measure analog signals converted to digital data, and in addition transport the data through WiFi or Bluetooth, we have to consider the location of the setup, minimizing signal interference or disturbance. Furthermore, a pulse oximeter that uses photodiode to measure light might be vulnerable to light, and the testbed may therefore limit light sources. The testbed is therefore defined on the basis of factors such as signal type and data acquisition method. In addition, the setup of the experiments should be adequately documented for the purpose of enhancing their reproducibility.

2.4.3 Data Quality Phenomena

The quality of data might be affected by many different phenomena. Both physical and logical sensors are likely to be affected by environmental perturbations such as light, movement and static. Hopefully, in most of the characteristic we see in the data is expected, such as a fall in saturation when a person holds his or her breath. An other example would be the thermometer showing a plausible estimate of a person or of the temperature in a room. However, some events shown by data might be

unexpected. If the thermometer suddenly indicates a fall in temperature for no reason, we consider it to be unexpected.

Whether the characteristics of data is expected or not, the reason can be known or unknown. If someone opens a window, then we know the reason for the drop in temperature. Or a person moving may explain the loss of a physiological signal. We place the events in signal data as one of the following phenomena.

- **Fundamental**: Holding one's breath causes oxygen saturation to fall, while breathing causes it to rise again. This is expected and desired characteristics we can explain.
- **Environmental**: Light and movement are artifacts, as are wireless perturbations or static. These affects is commonly unwanted; however, they may be either expected or unexpected.
- **Random**: Unwanted events and patterns we cannot explain includes sudden loss of signal and outliers in data signals.

Chapter 3

Pulse Oximetry

After the spectrometer was invented in the late 19th century, research and experiments from individuals and groups throughout the 20th century led to the modern pulse oximeter [40]. Pulse oximetry uses the principles of the Beer-Lambert law, which states that it is possible to calculate the concentration of an absorber in a solution simply by the use of light [50]. The calculation is made possible by measuring the light transmitted though a solution, using variables such as light intensity, path length and extinction coefficient of the substances at a particular wavelength. In simpler words, pulse oximetry uses the Beer-Lambert law to determine the oxygen concentration in the blood by measuring light transmitted through living tissue.

Today, a pulse oximeter is a well known physiological monitor that can record events related to hypoxemia, which is a condition of low oxygen in the body. By monitoring the oxygen saturation in blood, it is possible to discover health conditions preventing oxygen uptake in the body close to real-time. Therefore, they are used in critical care, in anesthesia, and in tracking the oxygen saturation of neonates. Oximeters is also one of the common sensors used in sleep studies. The sensor is small and minimally intrusive. It may be attached to different body parts: a finger, an ear or the forehead are most common, in addition to feet for neonates. Pulse oximeters fall mainly into two groups: reflectance and transmittance. The time frame of this research project only allows us to explore transmittance oximeters, the most common variety. On the other hand, even though we do not discuss their differences, most of the principles of pulse oximetry we explain in this chapters are also applicable to reflectance oximeters. In this context, it is worth noting that a common feature of pulse oximeters is that they provide a real-time estimation of heartbeats per minute (heart rate), and some also provide other physiological data. However, such features are also excluded from our research, as we focus on pulse oximetry as a technology.

We start this chapter by explaining pulse oximetry and the physiological processes it depends on in Section 3.1. Then we go on with an examination of the standards for quality testing oximeters in Section 3.2, including test procedures and considerations. In Section 3.2.2, we review common

metrics and analysis methods, before exploring related work on accuracy in Section 3.4. Last, in Section 3.5, we present a summary of our findings in this chapter.

3.1 Pulse Oximetry and Hemoglobin

In this section we explore the basics of pulse oximetry, and the physiological processes on which it depends. It is based the work of Wukittsch et al. [50], and Crapo et al. [12].

Pulse oximetry uses a method based on a two-wavelength, non-invasive spectral analysis of the blood, a technique that produces an estimation of the *arterial oxygen saturation (SaO_2)*. Oxygen is transported in blood from oxygen-rich environments, to peripheral tissue though the arterials. In general, most of the oxygen transportation in humans is done by the *hemoglobin (Hb)* protein found in red blood cells, each able to bind (or load) up to four *oxygen (O_2)*. An oximeter takes advantage of this property, and the oxygen saturation is an estimation of the proportion of *oxygenated hemoglobin (HbO_2)*, relative to the total amount hemoglobin. The binding process is also reversible, as oxygenated hemoglobin becomes a *deoxygenated hemoglobin (Hb)*, also known as reduced hemoglobin or just hemoglobin, after unloading the oxygen to a peripheral tissue. The (reduced) hemoglobin then travels with the blood back to the lungs through the veins for re oxidation.

As mentioned in the Introduction, it is possible to calculate the concentration of an absorber of light by analysing the light transmitted through tissue. A pulse oximeter measures the absorption of red and infrared light by the aforementioned oxygen-carrying hemoglobin protein. The absorbance of light, valued as the *extinction coefficient*, by hemoglobin and oxyhemoglobin is shown in Figure 3.1. There we see that the absorbance of red light (wavelength 650 to 750nm) by oxyhemoglobin is less than for (reduced) hemoglobin, and the reverse is true for infrared light (wavelength 900 to 1000nm). Therefore, it is possible to calculate the ratio between hemoglobin and oxygenated hemoglobin by emitting red and infrared light through human tissue and measuring the transmitted light with a photodiode. The SpO_2 value of a pulse oximeter can be expressed as the following equation:

$$SpO_2 = \frac{HbO_2}{Hb + HbO_2} \tag{3.1}$$

Pulse oximeters are calibrated against the gold standard for SaO_2 estimation, the CO-oximeter. A CO-oximeter analyse blood samples taken of the subjects with a multi wavelength spectrometer. Different from a pulse oximeter, a CO-oximeter is also able to measure the concentration of *methemoglobin (MetHb)* and *carboxyhemoglobin (COHb)*, which also supplies the CO in the name of this particular oximeter. These two, together with Hb and HbO_2, are the major absorbers of red and infrared light in the blood. As a consequence, a CO-oximeter is able to take all of the major absorbers

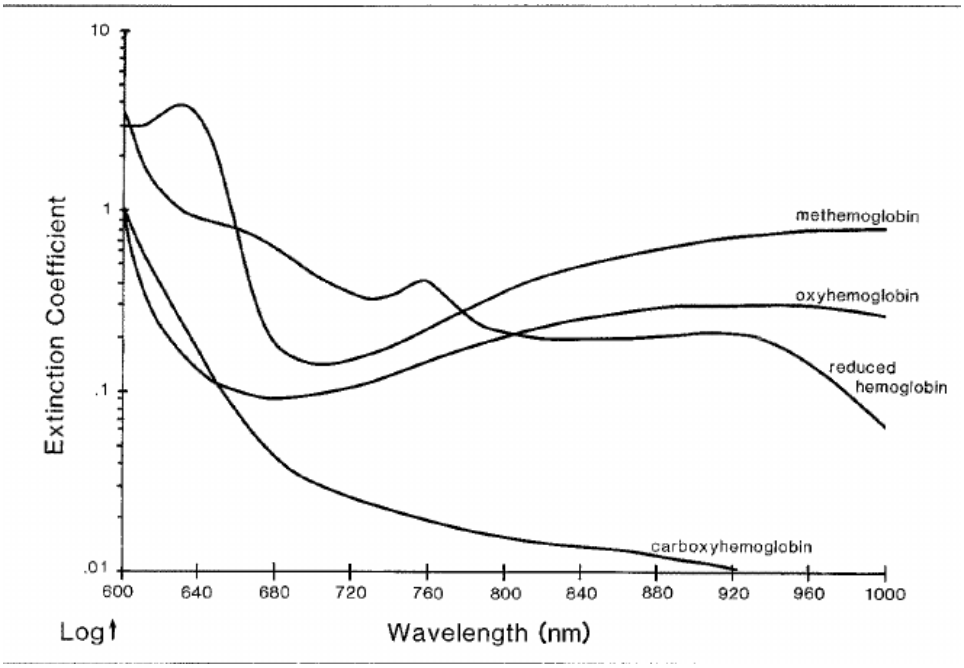

Figure 3.1: Hemoglobin Extinction Curves

in the blood into account when calculating the SaO_2 value, and therefore provide an accurate SaO_2 estimation.

Where the CO-oximeter only transmits the light through the blood sample, a pulse oximeter has to filter out other physiological components such as cartilage, bones and tissue from the equation. As the "pulse" in the name pulse oximetry might suggest, in addition to just emitting light and measuring the transmitted light with a photodiode, it uses the variable of pulsating blood in the calculation process. The nature of the pulsating blood circulation is used to filter out, or subtract, the absorbance done by non-blood artifacts, in order to calculate the SpO_2 value. By emitting the red and infrared light, and recording the minimum and maximum values within a heart beat, the pulse oximeter is able to calculate an R-value:

$$R - value = \frac{IR\frac{max}{min}}{Red\frac{max}{min}} \tag{3.2}$$

The R-value does not represent the SpO_2 value in itself, but it is empirically related to the SaO_2. When calibrating a pulse oximeter, the SaO_2 value from the CO-oximeter is compared against the R-values, and the relationship is stored in a table in the pulse oximeter's processing unit. When used afterwards, the oximeter can measure the Hb and HbO_2 values, and the internal microprocessor will estimate the oxygen saturation.

As expressed above, the value that indicates oxygen saturation measured by a pulse oximeter is called SpO2, and it is a percentage estimation of the total load of oxygen by hemoglobin proteins. The blood might still, and often do, contain more oxygen that is not carried by hemoglobin. The total amount of oxygen in blood is indicated by the *partial pressure of oxy-*

gen (PaO_2). PaO_2 describes the arterial oxygen tension, and it is measured in *millimeters of mercury (mmHg)*. It is useful to understand the relation between SaO_2 and PaO_2 when explaining the process of desaturation and (re)saturation, as explored below (this information is also important later to both the design of the oximeter tests and in explaining our results).

The hemoglobin follows the blood through the circulatory system, unloads oxygen to peripheral tissue, and as a consequence the PaO_2 falls. The oxygen-hemoglobin dissociation curve shows the relation between PaO_2 and SaO_2, and is presented in Figure 3.2. In the figure we observe oxygen saturation on the y axis of the blue graph and the total pressure of oxygen on the x axis. We start with an investigation of the graph's sigmoid shape. At high PaO_2 levels over 80 mmHg, the SaO_2 value has little effect of increased oxygen pressure on the blood. As we move leftwards in the graph and the PaO_2 falls, especially below 50 mmHg, the oxygen pressure has a more linear influence on the SaO_2 value. Let us further explore the this behavior with an example: In order to lower the SaO_2 value from 97% to 90%, a corresponding 30 mmHg or greater drop from the initial 100 PaO_2 is needed. On the other hand, a 10 mmHg drop in pressure from 60 PaO_2 results in a drop from 80 to 70% SaO_2.

From these data we can learn that most hemoglobins bind new oxygen at a slow rate when they are fully loaded. Additional oxygen is instead transported in the blood. Furthermore, the SaO_2 value falls at a slow rate from a high initial value. For example, an initial SaO_2 of 97% indicates that the PaO_2 is also high, and given a fixed rate of unloading oxygen to peripheral tissue, the fall in PaO_2 will not have a considerable affect on the SaO_2 until it falls closer to 80 mmHg. By observing this, we can identify an important property of the oxygen transportation. When the blood containing oxygen and oxyhemoglobin arrives at peripheral tissue, we assume (at least practically) that mainly the oxygen floating in the blood is being released to the tissue. Next, after reaching a certain point (usually between 60 and 80 mmHg), the oxyhemoglobin starts unloading its oxygen.

We have now discussed normal behavior in the oxyhemoglobin dissociation curve. However, additional impacts could also be considered. The red and green graphs in Figure 3.2 are models of a left or right shift of the curve, which is caused by factors such as pH or temperature. Even though the potential shift of the graph causes the relation between SaO_2 and PaO_2 to vary, this is not important to us. What we should note from the curve is its sigmoid shape, and how the SaO_2 behaves at higher PaO_2 values. In the next section we will learn more about the calibration and accuracy testing procedures of pulse oximetry.

3.2 Accuracy

A variety of studies exists on the accuracy of pulse oximeters, which is the standard evaluation indicating their quality. Newer studies on determining the accuracy of an oximeter reference the ISO 80601-2-61:2011

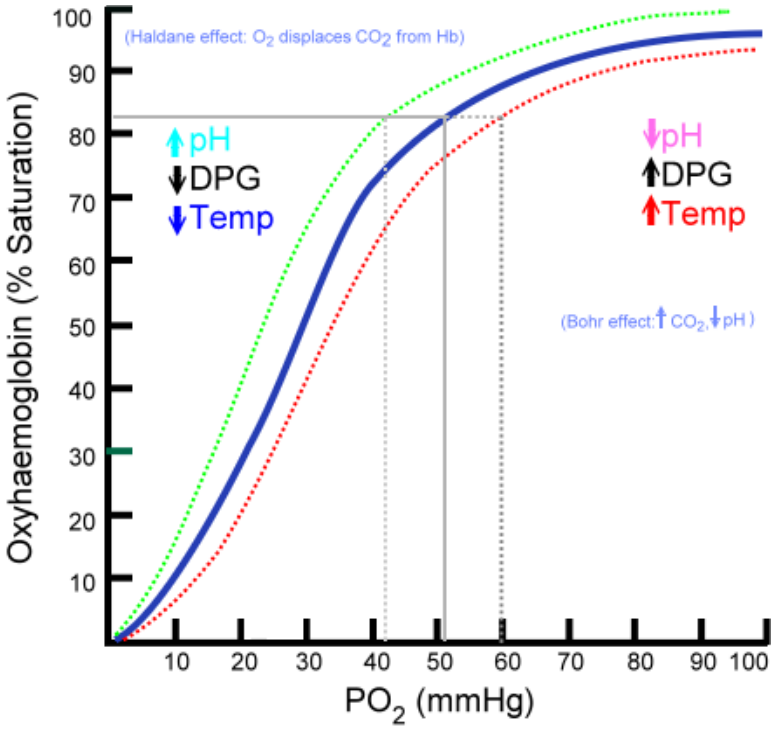

Figure 3.2: Hemoglobin's Oxygen Disassoation Curve

[22] (hereafter referred to as ISO of 2011), which is seen as an international standardization document for pulse oximeters. In December of 2017, the International Organization for Standardization published a revised version [23] (hereafter referred to as ISO of 1017, and the writers as the committee), which was further corrected and updated very recently, in February 2018. Most resources we found in the research process of this paper reference the 2011 version. Therefore, the time limit of our master's thesis does not allow us to rewrite this section according to the recently revised version. However, information in papers that is in conflict with the newest ISO standard is corrected. Also, it is worth noting that while the US Drug and Food Administration (FDA) recommendation document [17] (hereafter referred to as FDA 510k) often cited below refers to the ISO of 2011, the 2017 version of ISO 80601 also uses the FDA 510k below as a reference. We therefore also use the FDA 510k as a reference in some parts, and use the ISO 2017 as a control instance. The FDA 510k is a guidance document that is meant to assist the industry in preparing the documentation needed to demonstrate the safety of new medical equipment.

3.2.1 Testing procedures

The gold standard for both the calibration and measurement of accuracy is comparing measurements from the pulse oximeter against values from blood gas analysis done by a multi-wavelength CO-oximeter, or a radiometer [36] [24] [23]. The Food and Drug Administration (FDA) in the

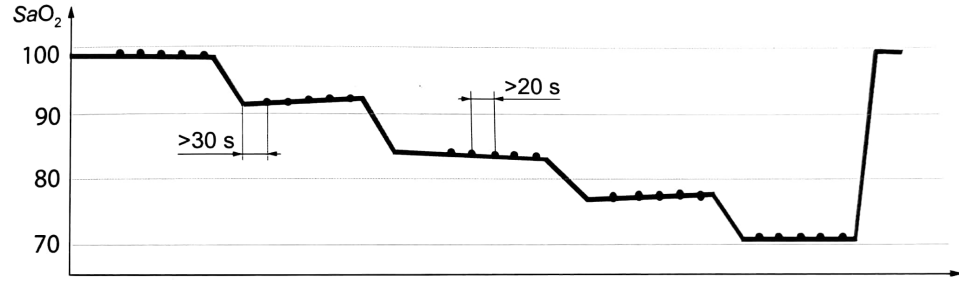

Figure 3.3: Oxygen Saturation Plateaus [23]

.

United States has premarket guidelines [17] on how to assess the accuracy of pulse oximeter. The FDA recommend an *in vivo*(i.e., clinical) testing of accuracy, which should include at least 200 blood drain samples analyses paired with the corresponding reading from the pulse oximeter. The 200 matched samples must be evenly spread out in the SaO_2 range, from 70 to 100%. They further recommend that at least 10 healthy subjects be tested who vary in age, sex and skin pigmentation; the greater of two persons or 15% of the group should be of dark skin pigmentation. The ISO 2017 specifies inclusion criteria of ages between 18 and 50, in addition to persons being classified as ASA Category 1, which also means no or minimal alcohol use [1].

The most common method to achieve a spread of test samples with saturation values between 70 and 100% is by having the test subjects breathe a gas mix containing *nitrogen (N_2)*, *carbon dioxide (Co_2)* and *oxygen (O_2)*. Changing the ratio between oxygen and nitrogen in the mix causes more stable periods, called *plateaus*, of PaO_2 values than is otherwise caused by breathing normal room air. As a result, arterial blood draws can be taken, R-value calculated, before the latter is matched against a simultaneous (or correlated in time) reading from the pulse oximeter. The relation is then stored as the resulting SpO_2 values in a table in the pulse oximeter[24].

Figure 3.3 shows a visualization of the plateaus mentioned above. The vertical line is the SaO2 values and the plateaus, and dots represent the blood draws. In the ISO of 2017, the recommendation is that the readings of a plateau should stabilize for at least 30 seconds before the first samples are matched, and additionally 20 seconds between each set of samples. With this method it is possible to achieve the range of samples as described and shown in Table 3.1. It is important to note for later discussion that the process of creating stable plateaus, as described above, requires medical considerations and possible supervision of medical personnel.

Non-invasive testing

The ISO of 2017 states that non-invasive laboratory testing is theoretically possible against other pulse oximeters, but such methods have not yet

SaO_2 plateau range %	Target number of samples
100 to 97	5
97 to 92	5
92 to 85	5
84 to 78	5
77 to 70	5
Total	25

Table 3.1: Example of target plateaus and ranges [23].

been shown to successfully demonstrate correct A_{rms}. However, the ISO of 2017 provides a suggestion for a non-invasive testing procedure. Instead of testing a pulse oximeter against a CO-oximeter, it is possible to test against a second pulse oximeter used as a reference, if the second oximeter is traceable to a CO-oximeter. By doing this, it is possible to drop the procedures that include blood draws from test subjects. Still, the rest of the testing procedure, and requirements for data analysis, are the same as with *in vivo* testing.

The paper suggests a testing procedure that includes breathing gas mixes in order to achieve oxygen saturation plateaus between 100 and 70%. Then, a total number of acceptable data pairs should be acquired to demonstrate statistically the specified SpO_2 accuracy (for instance, by following the plateau scheme similar to that of Figure 3.3, a total of 200 matched samples distributed on 10 subjects, from each 20 sample periods during different plateaus). As with use of the SaO_2 values from a CO-oximeter, SpO_2 values from the second pulse oximeter are used as the reference value from which the A_{rms} is calculated. It is important to note that the A_{rms} value would be relative to the gold standard CO-oximeter, including the error for the reference pulse oximeter.

Last, the standard proposes that other profiles for noninvasive testing are possible (e.g., a continuous data collection during gradual changes in saturation).

Functional Testers

Pulse oximeters are not intended to be recalibrated after being released to market. However, devices exits that can test oximeters without the use of *in vivo* test procedures such as described in the last section [51] [29]. These kinds of devices are known by different names, such as calibrators, simulators or functional testers. While differences exist in purpose and use, in this paper we follow the definition of the ISO of 2017 and call them functional testers.

A common feature amongst functional testers is that human tissue is

not required in order to test pulse oximeters. In short, and as a high level of generalization, these devices instead use a simulated finger or other simulated tissue to give the impression to the oximeter that a finger is placed between the LED and photodiode of the device. Then, instead of letting light pass through the tester, the simulated finger itself also contain a LED and a photodiode, which then "communicates" with the pulse oximeter. By doing this, the functional tester is able to (1) measure the light from the LED of a pulse oximeter and (2) send light into the photodiode and read the resulting SpO_2 values. They can therefore control the wavelength of the light emitted, and whether the R-curve is defined correctly. Hence, such testers are purely mechanical and electrical devices, and they do not measure oxygen, hemoglobin or other physiological factors.

Therefore, the ISO of 2017 states that no other means of verifying correct calibration of pulse oximeters exists besides methods mentioned in Section 3.2.1. Functional testers cannot test the accuracy of oximeters; rather, they ensure that the devices are acting according to the design of the manufacturers. Likewise, testers cannot determine if the design has been done correctly. Since pulse oximeters cannot be re calibrated, to correct potential errors the defective components must be replaced, or the oximeter redesigned.

Nevertheless, functional testers could prove useful in instances such as the periodic control of oximeters in use. In the work of Milner and Mathews [29], they used a tester to check over 800 oximeters currently in use in hospitals in the UK, and found that over 30% had technical issues that may lead to malfunction or wrongful SpO2 estimations. This result suggests that even with pulse oximeters not designed to be recalibrated, use may cause sensor errors. Therefore, the use of functional testers or other similar devices may be useful to control oximeters already implemented in medical environment. A possible case in point would be a doctor testing the equipment lent out to a patient.

3.2.2 Data Analysis

To investigate the state of the art in pulse oximeter quality analysis, we examine the two documents most used as references in the literature, the ISO of 2017 and the FDA 510k . In Chapter 3.4 of this thesis, we also explore related work on quality testing of pulse oximeters, which may add additional tools for analysis suited to our purpose.

The ISO of 2017 states that accuracy is the metric for pulse oximeter quality. The time limit of our research does not allow us to delve into the argumentation, but the committee states that their definition of accuracy represent a combination of both systematic and random components of error. The definition of accuracy, A_rms, which is also commonly used today by the manufacturers, is expressed as the root mean square between the tested pulse oximeter and the CO-oximeter used as reference. The formula is displayed in Equation 3.3.

$$A_{\mathrm{rms}} = \sqrt{\frac{\sum_{i=1}^{n} \left(Sp\mathrm{O}_{2i} - S_{\mathrm{R}i} \right)^2}{n}} \tag{3.3}$$

The standard also provides specific instructions for determining quality. The accuracy of pulse oximeters shall be a root-mean-square difference of less than or equal to 4% over the range of 70% to 100 % SaO2. The standard for the SpO2 values from the reference oximeters is that they shall be traceable to SaO2 values from a CO-oximeter. Furthermore, the CO-oximeter should have a SaO2 performance of 1% (1 standard deviation). The accuracy testing must be done according to the standard, using the methods we describe in Section 3.2.1. Then the paired SaO2 and SpO2 data points are pooled for all subjects, and A_{rms} is calculated using the formula in Equation 3.3. All pulse oximeters released to market should be labeled with the accuracy values, either within specific ranges or between 70 and 100% in general.

In addition to those described above, the FDA 510k guidelines add recommendations for analysis methods and graphical visualizations methods for the premarket documentation. The first is a Bland-Altman plot (which we learn more about in Section 3.2.2) and the second is an error plot, which can be understood as the distance between the reference value and the test object value (i.e., SaO_2 versus SpO_2-SaO_2).

The guidelines also include specifications about the demographic information to record under testing in laboratory conditions. Studies to determine accuracy should include number of subjects and samples taken, inclusion and exclusion criteria, specific laboratory conditions and subject motion, as well as information about the desaturation profile, the target plateaus and ranges.

Relation plot

A relation plot, or scatter plot, is a commonly used presentation of the correlation between two sets of data. Typically, data are paired in time and then plotted, with one value on x axis and the other on the y axis. When using pulse oximeters, one can plot the reference SpO_2 value on one axis and the test SpO_2 values on the other. In addition, a correlation coefficient can be calculated and plotted as a trend line.

Bland-Altman plot

In most of the related literature we mention in this paper, the authors base their discussion and analysis around results of a visualization called Bland-Altman plot. This statistical analysis method was introduced by J Martin Bland and Douglas G Altman [8], and further discussed in their paper

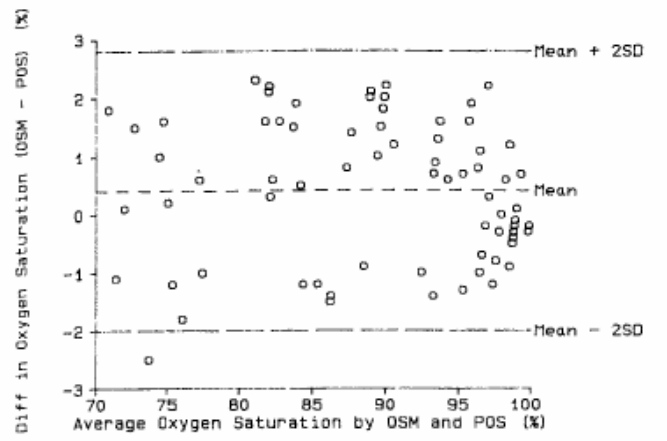

Figure 3.4: Bland-Altman plot example[8]

"Measuring agreement in method comparison studies" [7]. This section is based on these two papers.

The background for the work of Bland and Altman is their perception that the commonly used methods for measurement comparison, using correlation coefficients, often analyzed the results inappropriately and were misleading. When comparing measurement methods, the goal is to determine if the agreement between them is sufficient. If the differences are not clinically important, we can replace the old method, or use the two methods interchangeably. However, it is worth noting that defining what is clinically important is a matter of clinical judgement, and cannot be answered by statistical methods.

Therefore, Bland and Altman developed a new analysis method that tries to assess the degree of agreement, i.e, how much a new method is likely to differ from the old, famously known as the Bland-Altman plot (or analysis). The method is particularly useful when we cannot be certain if either of the measurements provides the right results. Even methods assumed to be the gold standard may not be without errors. Therefore, they argue that the new method should be plotted against the mean of the two. Therefore the Bland-Altman plot the differences between two methods against their mean.

Further, Bland and Altman provide a method to estimate the variation, or to predict where the range of the differences fall (called the *limit of agreement* in their argument). They estimate that the differences fall within a 95% limit of agreement, assuming they are distributed normally. The limit of agreement is defined as the mean of difference ($\bar{d}$), $\pm$1.96*sd of the differences s_d). The mean of the difference is also known as bias, and the standard deviation as precision.

Figure 3.4 shows a Bland-Altman plot of measurements of oxygen saturation from pulse oximeters. In the x axis we see the mean of the two different measurements ranging from 70% to 100%. Each circle is the difference between the first and the mean of the first and second

measurement. As we explain above, the mean marked with a dashed line is the mean of all the differences. The other two dashed lines at the top and bottom show the limit of agreement, or the standard deviation — 2 (1.96) — of the differences. By observing the figure, we learn three characteristics about the differences. The first is that the circles are spread out, and the differences are assumed to be distributed normally. If we observe strange patterns in the plot (e.g., by a histogram), we can check so see if the differences are abnormally and execute corrective actions. However, Bland and Altman note that non-normal distributions may still fall within the limits of agreement, though the outliers might be weighted on one side of the mean. The second characteristic is that the differences are concentrated around 0.4. The third is that the limit of agreement is -2 to 2.8%. As a result, we can therefore expect 95% of the measurements to be within the ±2.4% from the mean difference of the mean difference of 0.4%.

3.3 Inaccuracy and Limitations

The accuracy of a measurement from a pulse oximeter can be affected by different known and unknown variables, such as static perturbation from other electrical devices, device malfunction from use, and physiological characteristics in patients. The fact that pulse oximeters are medical equipment that record physiological processes also add factors we would not usually see in sensors recording environmental events. In this section we go through the both the technical and implementation challenges of pulse oximetry.

3.3.1 Averaging

Sample rate is the rate of which pulse oximeters measure and calculate SpO_2 values. In the research for this thesis, the time did not allow us to investigate the sample rate of each pulse oximeter, or how the manufacturer designed it. We can instead analyze what we now know about the calculation ofSpO_2 values, and then present possible outcomes and implications of different sampling rates.

The pulse oximeter calculates a SpO_2 value at each heartbeat. The calculated value may or may not be a correct measure of the real SaO_2 value. In order to filter out, or attempt to correct inaccurate values and outliers, the pulse oximeter is implemented with a filtering method known as averaging, also called averaging time. Averaging (A) is a mean calculation of a number of samples, and it is to filter, or *smooth out* data [14].

To illustrate the effect of averaging let us inspect the example in Figure 3.5. A person holds their breath for 10 seconds, and the oxygen saturation is recorded with a pulse oximeter at 1 Hz. The blue line is the hypothetical ground truth change of the SaO_2 value. The oximeter estimates and outputs raw values (red line) without averaging. For some unknown reason, the raw SpO_2 values (in red) is up to +-2% of the SaO_2 values. The yellow line is a mean of 3 measures, A = 3. As we can see, it smoothes out the raw

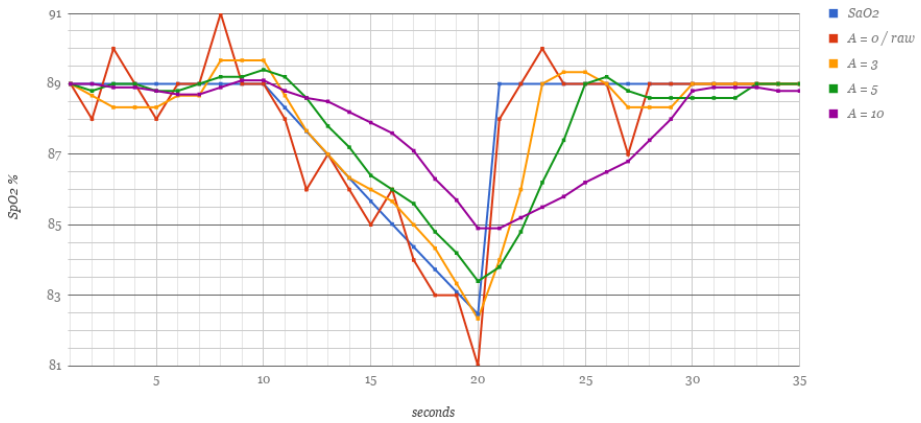

Figure 3.5: Example of averaging and its impact

measures of red, and shows a trend more similar to the blue SaO_2 graph representing the real values. The start, the rate of fall in oxygen saturation, and the lowest SpO_2 level are about the same. We may therefore say it is a better estimation of the SaO_2 value than the blue line representing the raw output. However, it takes 2 seconds longer than the blue line for saturation to back to 89%. With even higher averaging, the smoothing effect is more visible, as shown with the green and violet lines. At A = 10, the lowest saturation level is 3% higher than the lowest SaO_2 value. In Chapter 4 we further analyze this effect, and also learn about the possible consequences averaging has on detection of apneas.

3.3.2 Response Time

The fact that pulse oximetry uses the physiological characteristics of blood circulation in humans to estimate a SpO_2 value, implies that there may be accuracy parameters related to delay. Called response time in this thesis, the total delay can be explained mainly by the three factors we describe below.

The first is the time elapsed for the blood to be transported from certain locations within the body to its destination. To explain this further, Figure 3.6 shows the circulation of blood. As a general abstraction, it demonstrates that the oxygen-poor blood travels from the heart to the lungs (in blue), where it is oxygenated. It then goes back through to the heart and out though the arteries to peripheral tissues in the body (capillaries in figure). In our case, we understand the delay as the travel time of oxygen-rich blood from the lungs to the pulse oximeter placed on a finger.

The second factor in response time is the averaging described above. The displayed SpO_2 value of a pulse oximeter is a mean value calculated over a range of seconds or readings, usually between 2 and 20 seconds. A long averaging time may cause "smoothing" out variations of blood saturation levels, and will therefore cause delay before desaturations is

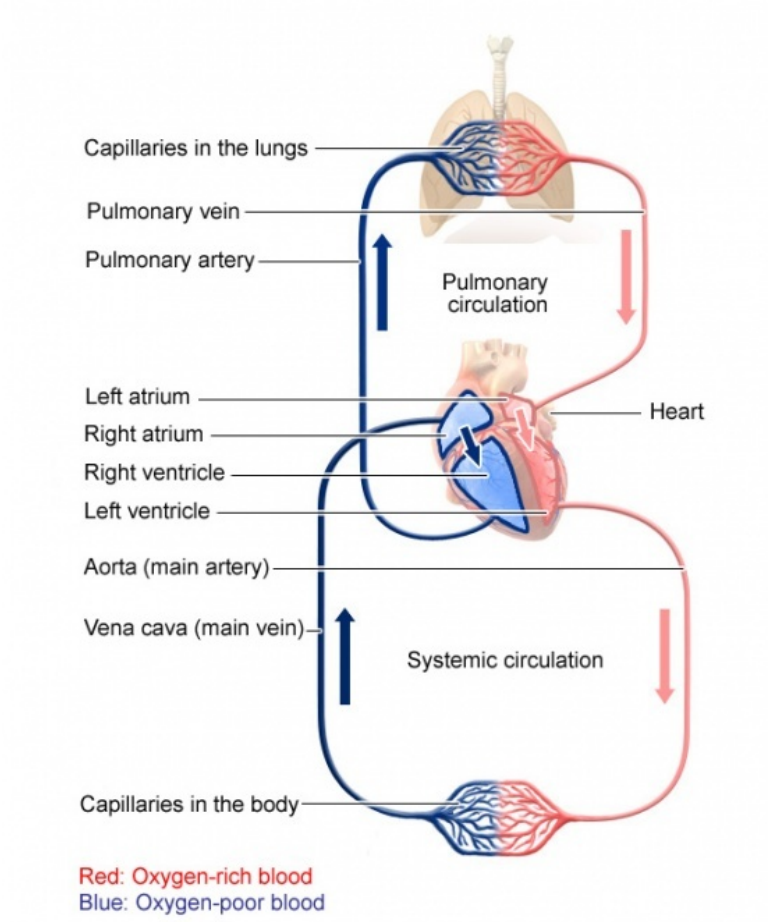

Figure 3.6: The Circulatory System[5]

notable in data.

A third factor concerns purely technical aspects such as data acquisition and microprocessor calculation time. We can assume that these factors have little importance for us as we consider milliseconds to be too short to effect our project in the scale of the two other factors.

However, since we do not know everything about the internal workings of each pulse oximeter, there may be additional design factors we do not mention in this section. In any case, studies show that the time elapsed from de- or (re)saturation, caused by breathing in reduced oxygen gas mixes or holding the breath until it is visible in the SpO_2 readings can vary from half a minute to up to a few minutes [35] [41] [10]. It is important to recognize these findings later when we analyze requirements for our testing procedures.

3.3.3 Environmental

In addition to the general design properties of pulse oximeters and the physiological nature of humans, we also need to consider possible environmental effects. Potential factors affecting accuracy are listed below.

- Movement artifact: it is well known that movement can cause disconnection or wrongly indicate desaturation of SpO_2 [3] [34].
- Ambient lights: a recent study with five sources of light reports a less than 0.5% difference between the control and the light source measurements on subjects of light skin [16].
- Nail polish. Minerals and other ingredients in nail polish may impacts the results of readings [11] [9]
- Low perfusion: the ISO of 2017 mentions low perfusion as a possible source of error.
- Hygiene and cleaning: universal precautions towards infection control [36], and that cleaning the sensors might affect accuracy [29].
- Positioning: the sensor for the pulse oximeter should be positioned according to the documentation or manual. Its design does not allow us to test more than one sensor at the same location. A subject's medical condition may disturb the baseline oxygen saturation of one of the fingers, resulting in errors in the readings.
- Building material quality: one possible unwanted effect of low quality would be for even small errors in emission wavelength of the LEDs produces error on the photodiode readings and resulting in incorrect calculations on the SpO_2 estimation [29]. Low-cost sensors may cause less accurate measures.
- Temperature: in Section 3.1 we noted that temperature changes may affect the hemoglobin extinction curve.

3.4 Relevant assessments on accuracy

Some important related work on accuracy was mentioned earlier in this paper. However, it is also useful to investigate other literature that assesses tests for the accuracy of pulse oximeters. From this we can also learn about suitable methods for testing, their procedures, and tools for analysis.

Lipnick et al. [24] tested the accuracy of six inexpensive pulse oximeters not cleared by the FDA, against a CO-oximeter.

They did so with the hypothesis that the pulse oximeters did not meet the ISO standard for accuracy. Executed by a professional laboratory, the study used FDA guidelines for (invasive) accuracy testing, which also meet the ISO of 2017 requirements. Three pulse oximeters were placed on each subject. They breathed a gas mixture to reach 10 to 12 stable plateaus in the range of 70 to 100% SaO_2. Hands were wrapped in warming band to ensure good blood circulation. A total of 536 matched samples from 22 healthy subjects were obtained. A demographic table included ethnicity and skin tone. Bias was plotted against SaO_2, and precision was the SD of the bias. A_{rms} was also calculated, and the requirement of $\geq$3% accuracy used. Further, Bland-Altman analysis with mean bias of differences, SD of differences and regression lines was included.

A table of the results was also presented, with results from 10% ranges and all paired observations, mean bias, precision, A_{rms}, and limits of agreement. Their conclusion on the accuracy of the pulse oximeters is based on the calculated A_{rms} value, and if it meets the FDA standard.

Phattaraprayoon et al.[35] measured the accuracy of two sets of pulse oximeters with non-invasive methods, one for wrist compared with one for palm, and one for ankle compared with for sole. They tested the pulse oximeters to see if the different locations were comparable. In addition, they also recorded the response time for obtaining the first sample. The study hints that they used heartbeat from a third device for synchronization; nevertheless, readings from two oximeters were paired at the same time at intervals. The time to obtain the SpO_2 samples was also obtained; however no method or definition was provided. Demographic data were also collected. The test samples were taken when the subjects had SpO_2 values at the range of 85 to 98%. For analysis, they used Student's t-tests and regression analysis. In addition they provide a Bland-Altman analysis, including bias and precision. The study included 150 subjects, and 145–147 tests for each of the paired test objects. Figures were presented with plots for relations with correlation coefficient and p value, in addition to a Bland-Altman plot. A table with data for each oximeters was also listed. Their conclusions are based on both the calculated limits of agreement and the correlation.

Macknet et al. [26] compared the first commercialized pulse CO-oximeter with standard CO-oximeters, both of which measure the total hemoglobin in blood. They included specific procedures, with preparations, presentation of the physical examination and test procedures, and criteria for process termination. For analysis they calculated bias, precision, and the Arm_s. The results are presented in scatter plots, with linear regres-

sion statistics and a Bland-Altman analysis. The authors also included a table with the accuracy by range.

Most of the literature for comparing pulse oximeters use the same procedures and analysis methods. The most common procedure is by comparing a pulse oximeter to a CO-oximeter, which also is the Gold Standard in the industry. Furthermore, in order to achieve desaturations, test subjects usually breathe gas mixes, or experience a natural change in saturation [48] [10]. Some authors also include criteria for inclusion and exclusion [49]. Of the studies we investigated written after Bland and Altman's first paper in 1986[8], use Bland-Altman analysis in addition to calculating the A_{rms} of the pulse oximeters.

In their clinical practice guidelines about patients' purchase of pulse oximeters, Pretto et al. [36] note that they are used in unsupervised environments and generally lack device specifications, regulatory approval, or clinical evaluation; therefore, they cannot be recommended at this time (2014).

3.5 Summary

In this chapter we explored pulse oximetry and learned about the important events in the process of estimating SpO_2. A summary of our findings is visualized in Figure 3.7, which represents an abstract generalization of the chapters highlights.

First, pulse oximetry depends on the circulatory system, or bloodstream. The blood travels from the heart to the lungs, back through the heart, and out to the pulse oximeter location, which in our case is on one or more fingers. Oxygen is loaded by hemoglobins in the lungs, and unloaded to peripheral tissue in the finger. Then the blood travels back to the heart to begin circulating anew. We can see the blood circulation in red at the top left of the figure. For the purpose of this thesis, it is sufficient to assume that the response time (t) of an oximeter is determined primary by the elapsed time for a particular amount of blood being transported from the lungs to the pulse oximeter. As we see in the top right part of our high level figure; HbO_2 and Hb are transported from the lungs with the blood and irradiated with light; absorption is then measured and calculated before SpO_2 is shown as output.

Within the microprocessor (or other similar processing unit) we have the 5 important events displayed in the bottom part of the figure. First, a light emittance diode (LED) sends red and infrared light through tissue, and a photodiode on the other side of the tissue measures the

remaining light that is not absorbed by HbO_2 and other absorbers. The results from the reading appear as point 1. In point 2, R values are calculated by using the absorbance properties of oxygenated hemoglobin and reduced hemoglobin. Then a stored calibration table is used to look up the R value, and an SpO_2 estimation is found in point 3. In 4, samples may be averaged intentionally to smooth out outliers and perturbations. It can be an average of n samples. Last, the estimated output in point 5 is the sum

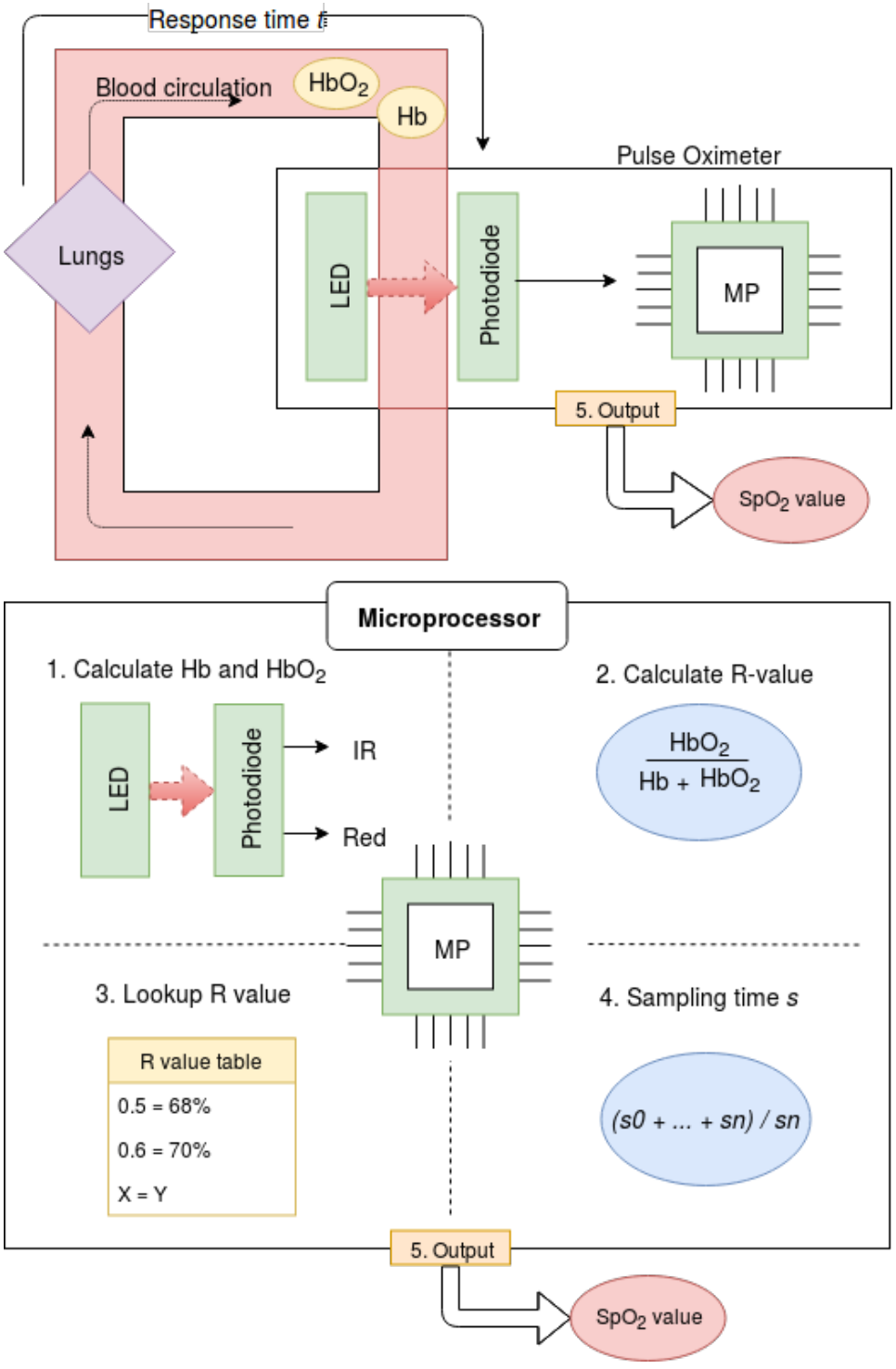

Figure 3.7: General abstraction of pulse oximetry

of all the previous steps.

Chapter 4

Obstructive Sleep Apnea

We start with identifying the characteristics of sleep apnea in Section 4.1. Taxonomy and scoring rules appear in Section 4.2. In Section 4.3 we delve into the diagnosis of sleep apnea, before a discussion in Section 4.4 of apnea events and breath-holding.

4.1 Characteristics

Sleep apnea is a sleep disorder caused by irregular breathing during sleep caused by apneas. An apnea is a partial or complete collapse of the upper airway [37], and has many causes. A normal cause is the tongue of a sleeping person moving backwards in the mouth and causing a blockage of the airways. The respiratory passage blockage often causes reduction or absence of oxygen supply. To resume normal breathing, the brain awakens the person. Sleep apneas therefore often involve periods of awakenings throughout the night for affected persons. The day after, the person usually does not remember any snoring, breathing cessations, or being in a waking state. For this reason, indications of a sleep disorder are often discovered by partners or roommates who experience snoring or gasping from an affected person.

Most people diagnosed with a sleep disorder have obstructive sleep apnea (OSA), which is defined as repeatedly having either complete or partial blockage of airways throughout the night. OSA is further defined in the next section. Central sleep apnea (CSA) is a less common form of sleep apnea. Rather than blockage of airways, the common reason is irregularities in the part of the central nervous system handling the respiratory effort that signals the body to inhale [2]. The symptoms and health implications are the same as for those suffering from OSA; however, it may be difficult to diagnose CSA before OSA is treated. A third kind of sleep apnea is mixed sleep apnea (MSA), which can be a mix of symptoms from both OSA and blockaCSA.

4.2 Taxonomy

As we outline in the introduction to this chapter, OSA is a breathing disorder caused by partial or complete blockage, of the airways during sleep. The total blockage of the airway during a nights sleep is *apnea*, while the reduction in airflow caused by partial blockage is defined as a *hypoapnea* [20]. Berry et al. have set rules to score apnea and hypoapnea [4]. In general, the former is scored by a 90% drop in signal from a respiratory sensor, and the latter is scored by a drop $\geq$30% drop, associated with a 3% drop in oxygen saturation. Also, both events should last $\geq$10 seconds.

The standard method for diagnosis of OSA is polysomnography, and the result from such a sleep study result in a polysomnogram. The data contain readings from the different sensors, and by analysing the data from different sensors, a specialist in sleep studies can both diagnose OSA and determine its severity.

AHI: A common measurement score for the severity of sleep apnea is the Apnea Hypoapnea Index (AHI) [20] [38], which is expressed as the number of apnea and hypoapnea events per hour. The severity of a potential sleep disorder diagnosis is defined by this classification. Persons with fewer than 5 event per hour are deemed to be without a sleep disorder. From there, the severity increases with the frequency, as seen below.

- None/minimal: AHI < 5 per hour
- Mild: AHI $\geq$,but < 15 per hour
- Moderate: AHI $\geq$15, but < 30 per hour
- Severe: AHI $\geq$30 per hour.

ODI: The oxygen saturation index (ODI) reflects respiratory events per hour associated with desaturation. It is defined as number of arterial oxygen saturations/hour $\geq$3% by Berry et al.[4]. Health sensor platforms for home diagnosis often use the ODI to accompany respiratory events to calculate AHI. However, because the ODI algorithms may differ between manufacturers [30], and the accuracy of the ODI depends on various factors we discuss in Section 3.3, it can not rule out mild OSA with certainty [36].

RDI The respiratory disturbance index (RDI) is defined as ODI + respiratory effort related arousals (RERA) per hour of sleep. This index is used for the classifications of sleep monitors below.

4.3 Diagnosis

The most common sleep study is polysomnography (PSG), which is the standard test for the diagnosis of sleep apnea [37]. The traditional laboratory study involves recording physiological signals using methods such as electroencephalography (EEG), electromyography (EMG), electrocardiography (ECG), airflow, and oxygen saturation among others. The process

also requires the patient to stay overnight in a laboratory for observation by medical personnel and for the data to be recorded.

In Figure 4.1, the top (A) shows a typical setup for polysomnography in a laboratory. Sensors measuring airflow are placed in the nostrils. Wired electrodes are attached to measure brain activity, eye movement and snoring. In addition, a respiratory belt around chest and belly measures respiratory effort, and a pulse oximeter finger clip estimates arterial oxygen saturation. It is also common to monitor heart activity.

The Polysomnogram marked as B in Figure 4.1 is a result of the sleep study. In the breathing event row, we see that breathing was absent for a period, which was followed by a decrease in oxygen levels in the top row. As a consequence, the patient had a period of wakefulness or easy sleep, visualized in the REM sleep stage row.

As we see in the illustration, the number of sensors attached makes the patient immobile, and the study may be experienced as uncomfortable or intrusive. In recent years, developments in mobile health devices have enabled monitoring with a different type of setup. Whereas the standard sleep study is executed in a laboratory, it is now possible to acquire mobile monitoring devices for home use. An example of a clinically certified monitor is the NOX T3 we learn about in this paper. As described in Chapter 2, this type of portable device enables physiological data to be recorded at the patient's home. The differences between regular polysomnography and the mobile NOX T3, for example, is the number of channels included to record physiological data. Furthermore, while PSG requires medical attendance, NOX T3 is a home monitoring device that can be set up by a doctor or technician and then used at home without supervision. Classification and requirements for the different types of sleep study are is presented by the Center for Medicare and Medicaid Servics [28]. In Type I, the sleep must be attended by a sleep technologist, and a full set of sensors are commonly used in PSG. Type II and Type III require a minimum of 7 and 4 channels, respectively, and are used unattended. So are Type IV monitors, which require channels that allow calculation of AHI or RDI from airflow or thoracoabdominal movement (breathing).

4.4 Events

Berry et al. [4] define desaturation associated with a respiratory event as "a drop from a baseline SpO2 preceding the event to the nadir in the SpO2 following the event". In simpler words, for measurements from a pulse oximeter, an apnea event is a continuous drop in oxygen saturation from a baseline, and the event proceeds until the saturation rises again. However, they state that identifying the baseline saturation may be difficult if the desaturation events occur back-to-back. This challenge may also be reinforced by factors we learned about in Chapter 3, as well as factors investigated above.

As we have learned, both the sigmoid shape of the oxyhemoglobin dissociation curve and averaging affect pulse oximetery's ability to measure

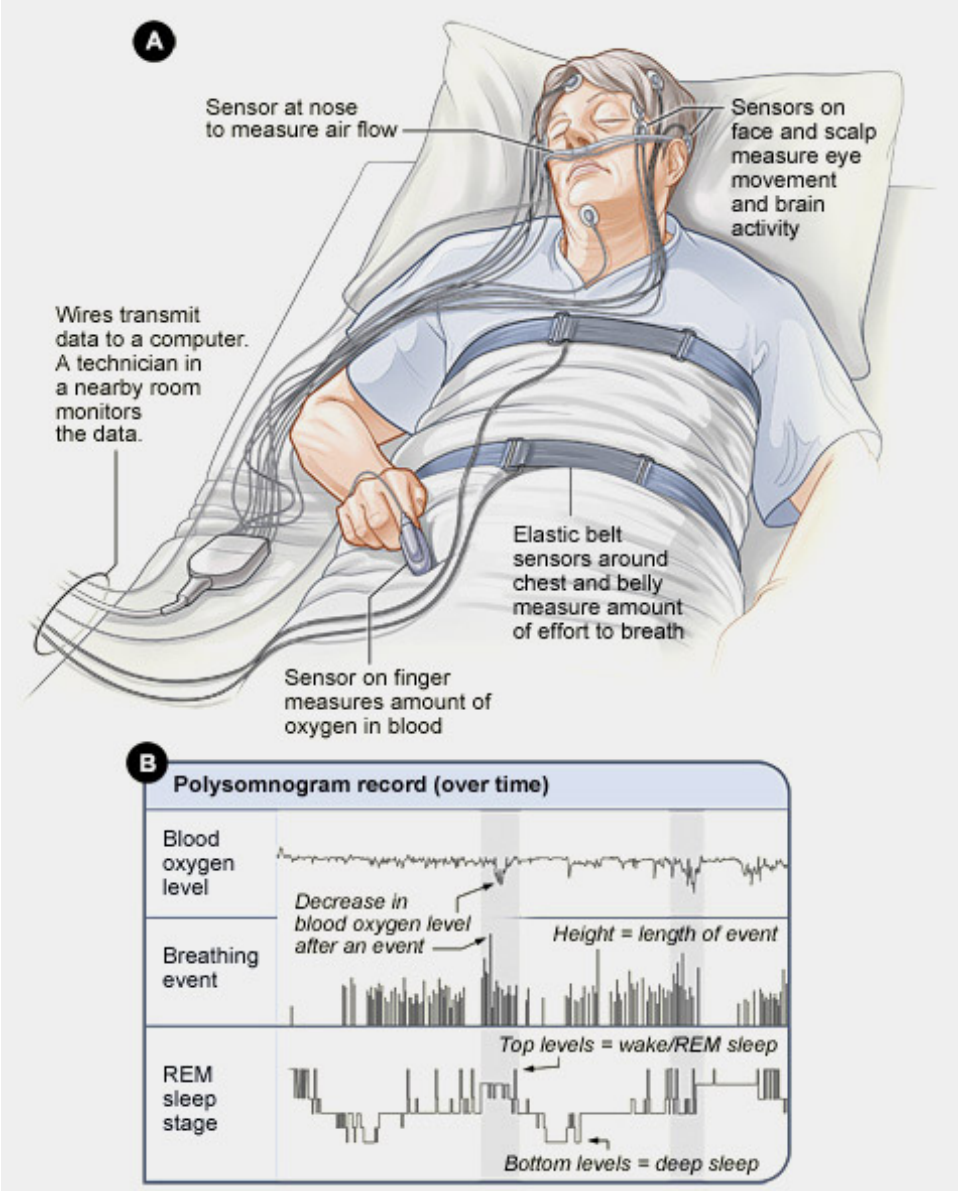

Figure 4.1: Polysomnography(A) and polysomnogram(B)[31]

desaturation events (e.g., at sea level, the normal baseline oxygen saturation of healthy persons is 97±2% [12]. For a pulse oximeter to measure the 3% desaturation from an initial level of 97%, the PaO2 must drop significantly more than from an initial level of 93%. In addition, a high averaging may smooth out potential dips in saturation. In this section we examine the desaturations and averaging in detail.

4.4.1 Averaging

In Section 3.3 we noted that averaging can cause the pulse oximeter to smooth out events. As a study from Farre and colleagues [14] points out, underestimations of oxygen desaturation are systematic, and are caused by the limitation in what they call the *dynamic response* of an oximeter. This dynamic response can be understood as the pulse oximeters ability to measure the real SaO2 values of the test subject, and is affected by averaging time and other filtering algorithms. The example in Figure 4.2 illustrates the impact of averaging.

A supposed test subject hold their breath for 10 seconds, and 4 identical pulse oximeters estimate the SpO_2 values. Their only difference is the internal averaging (T). The calculated output from the oximeters is given with T = 0 (blue), 3 (red), 5 (yellow) and 10 (green). Sampling rate is 1 Hz, shown as dots on each line graph. We assume the initial saturation to be 93%, with an mean drop rate of 0.45% per second. The calculation of each sample is done by computing the mean of the T's last measurements. That means for T = 0, the output simply drops 0.45% for each new sample, since there is no averaging. For T = 3, a mean of the 3 last SpO_2 measurements is calculated and given as output, etc.

We start by observing T = 0. The SpO2 level drops by about 4% over the time period of 10 seconds. That is a higher drop than the ≥3% requirement from AASM to score an apnea event. Next, for T = 3, we see a reduction in drop from the baseline to the nadir in about 0.40%. The drop is still above 3%, and we can still score an event, but the value is closer to the limit. At T = 10, the desaturation is no longer a desaturation event, since it is only a drop by 2.3%. In general, we see that the higher averaging, the flatter desaturation period. In addition, the rise in saturation back to baseline may be slowed down.

This example is only an illustration of the effect of the pulse oximeter's averaging when estimating a person's SpO2. In reality, oximeters may have other, similar implementations of averaging, and may also contain other algorithms (e.g., to filter outliers). Therefore, in sleep studies it is important for the averaging to be as low as possible, and for the oximeter to provide samples at a satisfying rate. For use in sleep studies, the recommended averaging is 3 seconds, and sampling rate 10Hz [36] [45]

4.4.2 Rate of Fall and Breath Hold

Apnea is defined as a partial or complete blockage of airways. In this section we use the term to refer to complete blockage. It is possible to

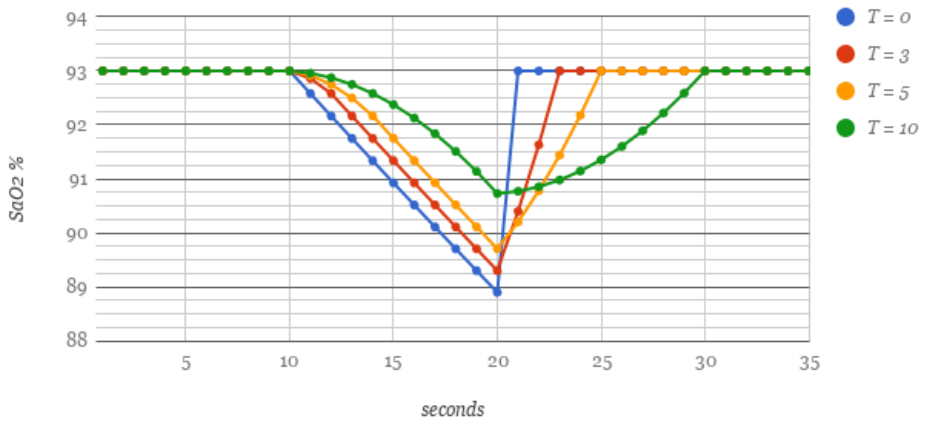

Figure 4.2: Illustration of averaging time(T)[14]

simulate apnea in a waking state, by breath being held. The result of both sleep and waking apnea is that no air, and therefore no oxygen, finds its way down to the lungs. The difference is that one is involuntary and the other takes place of the person's free will. Although the mechanisms are the same, when you start holding your breath, its saturation begins to fall. The rate of that fall is the object of closer investigation is this section.

Strohl et al. present a study to examine the relation between desaturations from breath-held apneas and those during sleep [44]. They instructed one group of healthy subjects to simulate apneas by breath holding, and compared the SaO_2 values against a second control group of patients having apneas in sleep. They simulated both obstructive and non-obstructive apneas; breathing with and without respiratory effort on a closed airway. The breath-holds lasted 10 to 25 seconds, and were initiated from functional residual capacity (FRC), which is the state of the lungs after a normal passive exhalation. The study found no significant difference between obstructive and normal breath holding. Also, the rate of fall is not affected by simulating obstructed breathing using the Mueller or Valsalva maneuvers[1]. Furthermore, the results in the study indicated no difference between apneas simulated while awake and real apneas in sleeping patients. The study does not rule out other explanations; however, it suggest that the rate of fall in saturation is much determined by the initial oxygen saturation. In Figure 4.3 we see the relation between the initial SaO_2, and the rate of fall in SaO_2 from two subjects (black and white dots) during sleep. From an initial 94 to 96% SaO_2, we can expect the rate of fall to be half of an initial saturation from 84 to 86%.

A second a newer study on breath holding from FRC is presented by Sasse and colleagues [39]. Here, the nonsmoking test subject was instructed to not perform any respiratory maneuvers while holding their breath for 35 seconds. Meanwhile, blood draws were obtained and later analysed to measure the PaO_2. The results revealed a greater drop than observed

[1]ihttps://en.wikipedia.org/wiki/M%C3%BCller%27s_maneuver

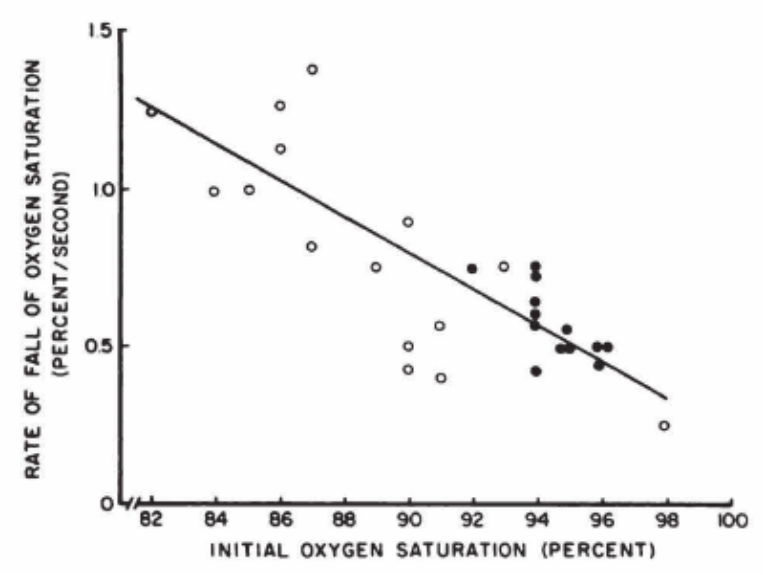

Figure 4.3: Rate of fall in saturation

in earlier studies at the time, including the one described above. Over the period of the first 35 seconds of breath holding, the PaO_2 declined an average of 50 mmHg from the initial 110 mmHg. Since the breath holding was initiated from a high initial PaO_2 value above what a pulse oximeter is able to measure, the study indicates that holding the breath long enough causes measurable changes in SpO_2 values.

In this regard, it is interesting to investigate the reason for the breakpoint of breath holding. M.J. Parkes[32] dives into possible rationales while discussing the difficulties of explaining it.

He states that the easiest way to quantify breath holding is by the duration, which may be affected by various of factors. He points out that the tolerance of discomfort is not equal for all subjects. And even within subjects, duration may be increased by distractions and successive trails, physiological factors such as the starting long volume, or an unconscious effort to breathe. All of these factors are likely to affect the breakpoint of breath holding. It is mainly the physiological signals that work against the strength of will of the person. When holding the breath, the respiratory signals from the brain do not stop; instead, the person rather closes the airway and controls the muscles. Parkes therefore sees the breakpoint as being determined by the relation between the effort of the person and the physical (negative) return from breath holding. He also cites the common misconception that if you hold your breath long enough, you fall unconscious.

4.4.3 Alternatives to Breath Hold

In this paper we identify breath holding as the best option to influence changes in arterial oxygen saturation. Within our limits of our benchmark-

ing tool being non-medical, we see other alternatives that might affect the oxygen saturation as less suitable.

The commonly used method for lowering oxygen saturation involves breathing gas mixes normally found in breathing air but using a different ratio between oxygen, nitrogen and carbon dioxide. Since we did not have such gas mixes available for this research, we evaluated an alternative method of changing the ratio of the gas mixes in air, consisting of a subject breathing in a locked container. Even though this might give us more stable and gradual changes in oxygen saturation, it introduces even more medical considerations for us to investigate. By holding breath, the oxygen saturation falls very rapidly; however, this occurs over a short period of time (usually between 10 and 30 seconds). By having test subjects breathe in an locked container, they will be exposed to a potential shortage of oxygen for an extended time. By using this option, a more thorough medical investigation is needed. We evaluate breath holding as safe, because when a person hold his or her breath, at some point the brain overrules the strength of will and starts the breathing process again. In contrast, we see breathing in a container as introducing the risk of a sort of slow suffocation or carbon dioxide accumulation in the body. The brain of a healthy person therefore acts as a "fail safe" in our research, as the person would eventually start to breathe when the urge to do so is higher than the strength of will. Therefore, we avoid any methods of altering the air's oxygen composition. It is also worth noting that we found no methods in the literature for lowering arterial oxygen saturation other than breathing gas mixtures or holding the breath.

Part II

Design and Implementation

Chapter 5

Preliminary experiments

In this chapter we report on a series of experiments conducted on the health sensor platforms used in the benchmarking process to investigate their behavior. In Section 5.1 we discuss our goals and expectations for our experiments. We start with our reference sensor, NOX T3, in Section 5.2. Next, we explore the BITalino in Section 5.3, and the CH in Section 5.4. Finally, in Section 5.5, we investigate the different methods of synchronization and their usability.

5.1 Introduction

In this thesis we test pulse oximeters that are unique in design and technology. We may therefore expect data output to be distinct in both character and format. Typically, a platform provides software for displaying the data, or as a method of data acquisition. However, in our research we are not interested in a visual representation of the data. In order to test quality, we ideally want sensor data as raw and as unprocessed as possible.

The experiments described later in this paper focus on the benchmarking of pulse oximeters. To make this process as smooth as possible, it is useful to detect possible difficulties, errors and unwanted incidents concerning a possible implementation at an early stage. As follows, we describe preliminary experiments done with the goal of determining methods of data acquisition. We also discuss the state of the data quality and possible challenges. Next, we establish the possibilities for methods of synchronization between sensor data. We can specify the goals of these experiments as follows: We can specify the goals of these experiments as the following:

1. Establish a method of data acquisition and data storage. The method should also be able to provide the best possible representation of the data (within a reasonable time frame). By "best", we mean to strive to limit perturbations within our control, such as noise and other factors, that have negative effects on the data quality.
2. Find synchronization methods for the oximeters between the platforms.

Consequently, all experiments in this chapter follow a common procedure. First, we do not need any additional expectations beyond those indicated in the documentation of the sensors. We only want to observe the behavior of data from each particular sensor, and to discuss the possible challenges introduced. However, it is important to note that our brief testing in this chapter is insufficient to indicate any quality statement; we mean only to survey the sensors and their possibilities.

5.2 NOX T3

5.2.1 Data Acquisition

As previously described in Section 2.2, the NOX T3 is a closed environment. Data are presented to us only through their software Noxturnal.

We connect the device to a computer manually with the pre-installed software. We can then observe the results in the program. However, we want the data in a file as we want to compare the data with other sensors.

By searching through our alternatives in the software, we found two for output. We could export the data as an Noxturnal file type. This option is not practical since it can only be opened and viewed by the Noxturnal software. The second alternative allows us to copy the data as comma separated values, with rows of timestamp and value pair separated by tab space. If we want data from a second sensor, these data has to be copied to a second file.

5.2.2 Data Characteristics

The readings from NOX constitute the reference values in our benchmarking. We can observe the behavior of the samples of SpO_2 data from this platform, and assume that this is as good as it gets. Below we see a dump of a recording.

```
11:17:28:868   97
11:17:29:202   97
11:17:29:535   97
11:17:29:869   97
11:17:30:202   97
11:17:30:535   98
11:17:30:869   98
11:17:31:202   98
11:17:31:535   98
11:17:31:869   98
11:17:32:202   98

...

11:36:54:944   98
11:36:55:277   98
11:36:55:610   98
11:36:55:944   98
```

```
19 11:36:56:277  98
```

From this we learn that the sample rate is set to 3 each second, at a fixed interval. Even though this interval is changing a little over a period of about 20 minutes, we can say that the sensor provides a new estimation $\approx$ 333ms. However, this interval changes between sensors (e.g, for an accelerometer, a new reading is presented every 100ms, as seen below).

```
1  11:17:28:959  0,00973402754377517
2  11:17:29:059  0,0194995858738656
3  11:17:29:159  0,039062234057039
4  11:17:29:259  0,00976555869808937
5  11:17:29:359  0,0195852199838154
6  11:17:29:459  0,0390489576270459
7  11:17:29:559  0,0194637376432305
8  11:17:29:659  0,00976555857576278
9  11:17:29:759  0,00979711847651998
10 11:17:29:859  0,00979711847651998
11 11:17:29:959  0,00976555857576278
12 11:17:30:059  5,40032268547819E-05
```

We are not able to find specifications about the averaging in the documentation. However, as stated earlier in this paper, the recommended averaging for sleep monitoring is leq3 seconds. We may therefore assume that this frequency is implemented in NOX as it is a sleep-monitoring device.

5.3 BITalino

5.3.1 Data Acquisition

For data acquisition from BITalino we use software developed by Svein Petter Gjøby, which consists of one application (app) for storing data, and one specifically adapted to fit the protocol of BITalino [18]. The Collector app collects and stores data to either a file or an external database. For our use, we store the data to a file on an Android smartphone with the apps installed. The second Android app, called Bitalino, is a wrapper application to collect data from the BITalino platform. In this app it is possible to control the functionality within the BITalino device, including channels and sampling rate.

5.3.2 Data Characteristics

As the SpO_2 values are presented in a 10-bit format. In order to compare them against other SpO_2 they are converted into 8-bit values with the formula in Equation 5.1, where *nbits* is the number of bits in the channel.

$$(0.25 * (2 * *(10 - nbits)) * spo2value) - 0.8 \tag{5.1}$$

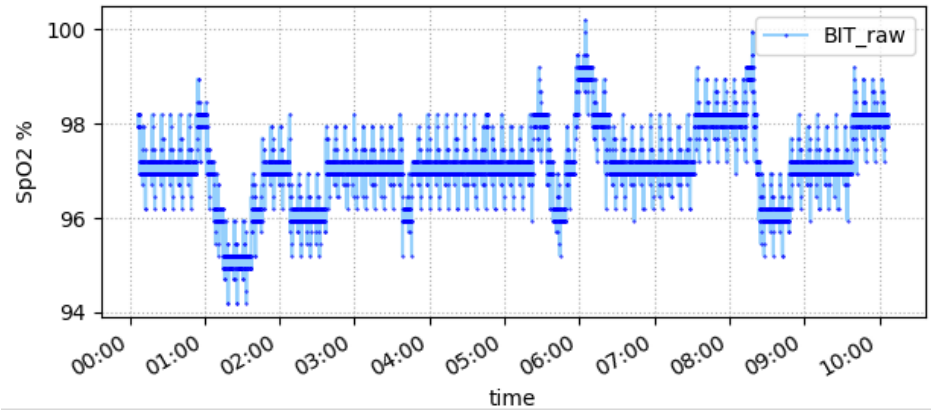

Figure 5.1: BITalino pulse oximeter plot

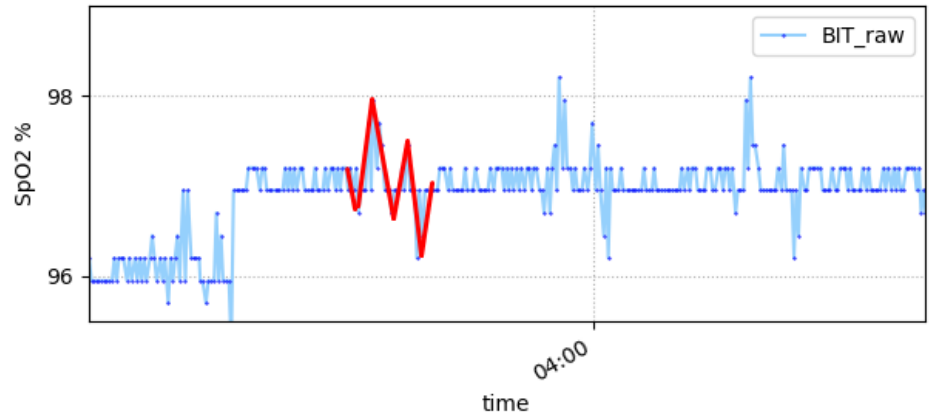

Figure 5.2: BITalino oximeter pattern cut

Noise

Figure 5.2 shows a plotted recording from a pulse oximeter from BITalino. The dark blue dots illustrate the values from the individual samples, with blue as the trend line. The first thing we notice is that there is no clear line showing stable SpO2 values. The blue samples show that two parallel lines run throughout the recording. That might indicate that each SpO2 value is represented by more than one 10 bit value. In Figure 5.1 we see a close up of the signal. If we ignore the pattern indicated in red, we see that the values are not spread more than 0.2 to 0.3%. Therefore, we could use the value to the nearest integer to get stable values. However, as we also see in the figures, a periodic pattern occurs in the data. This pattern is about 2%, from high to low. Using different channels does not eliminate this pattern, which persisted through testing of the three BITalino boards used in this thesis.

5.3.3 Summary

As we mention earlier in this thesis, the pulse oximeter from BITalino is a stand-alone device that can be used without being connected to the BITalino platform. The oximeter has both drivers and software to visualize recordings and store them on a computer. However, we are not able to complete this option. The device is not found as such when we connect it to a Windows 10 computer. The reason might be that we did not find the right drivers, an incompatibility with Windows 10, or some other unknown. However, a search through common problems on the Internet reveal that a special USB mini-cable is needed. When purchased from the BITalino web shop, the pulse oximeter comes with a special USB mini-to-channel cable. Therefore, we have to use the BITalino platforms as a part of the data acquisition. The CMS-50+ can be found in various web shops on the Internet. An investigation of the literature did not identify any common problems such as the one described above. We can therefore assume that what seems to be a disturbance pattern, as described above, is caused by the BITalino board and not by the pulse oximeter itself. We do not see it as our task to try to identify or fix the signals from the BITalino board, nor does the time frame of this research allow us to do so. As we have established a method for data acquisition, we therefore use the data as it is in later experiments.

It is worth noting that the application does not always work according to the instructions. It is not easy to start the data acquisition process, and it may stop while recording. Instead of developing a new data-acquisition method for BITalino, we chose to use it regardless of the perturbations in data and the problems with the mentioned applications. Instead we introduce guidelines in the benchmarking experiments. First, changing channels for the oximeter between recordings may eliminate bad channels. Second, the time frame available for each experiment should be long enough to make sure the app is working and providing data. Other than that, we use the data as is.

5.4 Cooking Hacks

5.4.1 Data Acquisition

In contrast to the two platforms discussed earlier in this chapter, Cooking Hacks lacks an already implemented method of data acquisition for our purpose. Therefore, we examine this platform more thoroughly.

Cooking Hacks provides documentation [19] that highlighting the possible methods of connectivity, visualization and extraction of data. For each sensor, different code snippets presented serve as examples of intended use. In general, you can either display your data at the TFT display or send it though a serial port to your computer. As we are interested in raw data to be compared against other pulse oximeters, we use the code example for displaying the data in Arduino IDE's serial monitor.

First we use the code from documentation(Section 6.2.2.4) without

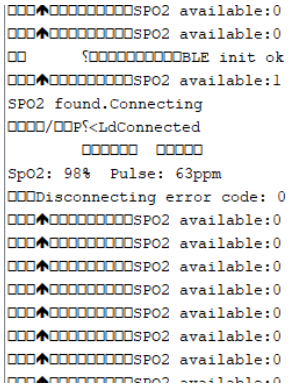

Figure 5.3: Serial output from Cooking Hacks

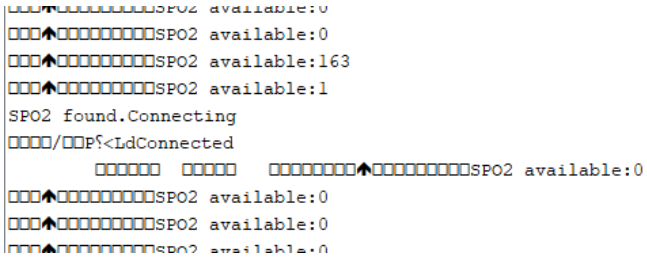

Figure 5.4: Serial output from Cooking Hacks with failed reconnections

any modification but specifying the media access control address(MAC address) to our pulse oximeter. Our first result is shown in Figure 5.3. MySignals connects to the pulse oximeter, then disconnects after data are received. Then it waits for the oximeter to reconnect, which introduces some problems. The time until a successful reconnection follows no apparent pattern, and may vary from 6 to 30 seconds, even never reconnecting. Additionally, when connected, the board may not be able to receive any data from the oximeter before disconnecting, as we see in Figure 5.4. As a consequence, we are not even close to receiving updated SpO2 estimations at a rate that is satisfying for our benchmarking.

Limiting external factors

Both of the code examples in the documentation specify the described behavior of disconnection, except for the failing reconnection. Therefore, the problems we experience may be caused by what we defined as environmental phenomena, as discussed in Section 2.4.3. Our research lab includes several computers and other devices, and it is located near other workrooms. Our first approach was to limit external factors that might affect MySignals' ability to reconnect with the pulse oximeter. The new location is checked by scanning for Bluetooth and WiFi signals, and WiFi and Bluetooth were deactivated on the laptop. However, the connection problem did not improve, even with no electrical devices sources other than the laptop used for data acquisition.

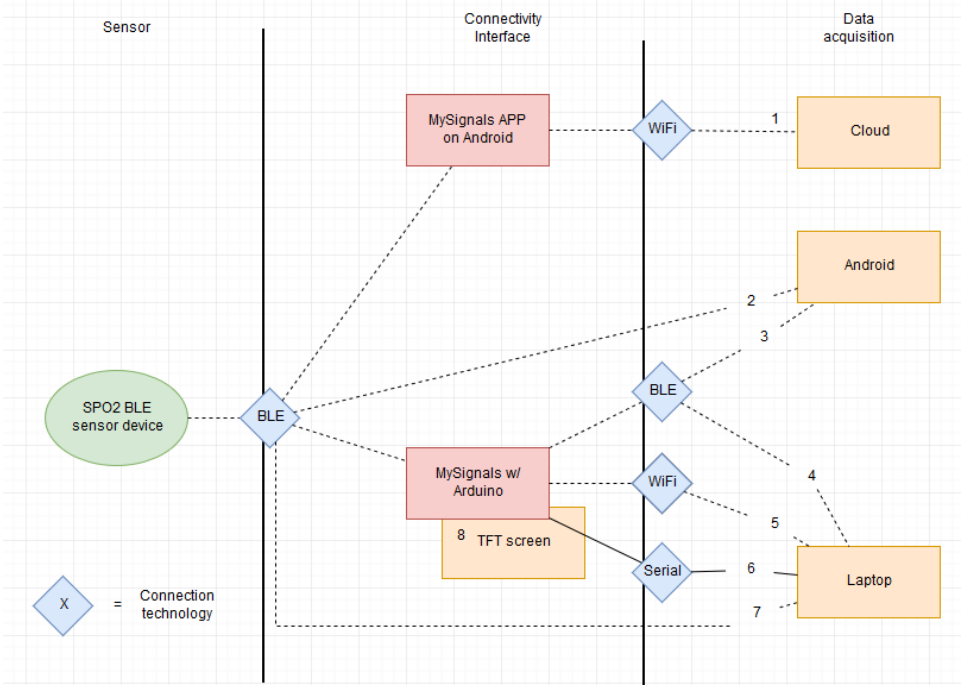

Figure 5.5: Data acquisition for Cooking Hacks' MySignals

We therefore assume that environmental phenomena are not the main problem. MySignals is connected to a laptop computer, which also provides power to the unit. In addition, a power supply can be connected directly to the Arduino unit. In this context, other connection possibilities are explored for removing the Serial/USB connection with the laptop, and the possible source of connection disturbance.

Writing New Code

An overview of how we investigated the documentation, and identified the different data acquisition method is presented in Figure 5.5. As we see in the figure, there are many different possibilities for data exchange with the pulse oximeter. We can program the Arduino to send data to a computer with Bluetooth (4), WiFi (5), and Serial (6), as described above. In addition, the documentation describes use of the MySignals APP (1, which was not possible at the time); connecting to an Android device over Bluetooth (2 and 3), and showing results on the TFT screen (1). However, none of the alternative methods gave better results or proved easier than trying to improve alternative 6. The series of experiments exploring these alternatives is presented and discussed in B.1 at the end of this paper.

The documentation for MySignals HW v2 fails to explain the reason for the code behavior, providing only a short description about some of the functions. As a result, initial testing with modifying the example code, such as simply removing disconnection lines or changing delays, halted the data stream or provoked other unwanted behaviors. However, without a

deeper understanding of the code, solving the connection issue by altering the example code is an experimental task. As it appears to us, repeated attempts to connect to the pulse oximeter should be avoided. We therefore investigate how MySignals operates the pulse oximeter and write new code in this section.

A basic description of Bluetooth Low Energy technology is presented in Section 2.1. With this information in mind, and with the help of the similar example code of the Body Temperature BLE sensor in Section 6.2.5.4 of the MySignals documentation (which is similar to our code example), we investigate the example code. The following code snippet shows important parts of the code, and is explained in the comments above the function calls.

```
1  // Scan for BLE devices
2  MySignals_BLE.scanDevice(MAC_SPO2, 1000, TX_POWER_MAX);
3
4  // Connect to oximeter
5  MySignals_BLE.connectDirect(MAC_SPO2)
6
7  // Subscribe to data
8  MySignals_BLE.attributeWrite(connection_handle_spo2, SPO2_HANDLE,
       0x01, 1)
9
10 // Wait for data
11 MySignals_BLE.waitEvent(1000)
12
13 // Get the value stored
14 spo2 = MySignals_BLE.event[13];
15
16 // Unsubscribe to the oximeter stream
17 MySignals_BLE.attributeWrite(connection_handle_spo2, SPO2_HANDLE,
18   0x00 , 1);
19
20 // Terminate connection to oximeter
21 MySignals_BLE.disconnect(connection_handle_spo2);
```

After we identified the parts needed for subscribing to the data stream from the pulse oximeter, we read and reverse engineered the source code in the installed MySignals library. Figure 5.6 presents a Unified Modeling Language (UML) diagram showing the flow (from top to bottom) of calls used to connect to the oximeter and to subscribe to its data stream. In the figure we see that commands are not sent directly to the pulse oximeter; they go through the BLE module on the MySignals board. Against the end (bottom) of the diagram, we inserted a loop. This is our desired behavior, as we unsubscribe after the loop.

Figure 5.7 shows the internal logic of how data are moved from the serial buffer to output. In short, data are placed in the serial buffer and actively gathered by the Arduino device. We come back to this stage later in this section.

In the new program we write we do not want to unsubscribe from the stream. Instead, we reset the module and reconnects in case of discovered errors such as a continuous series of out-of-bound values (such as 0). When

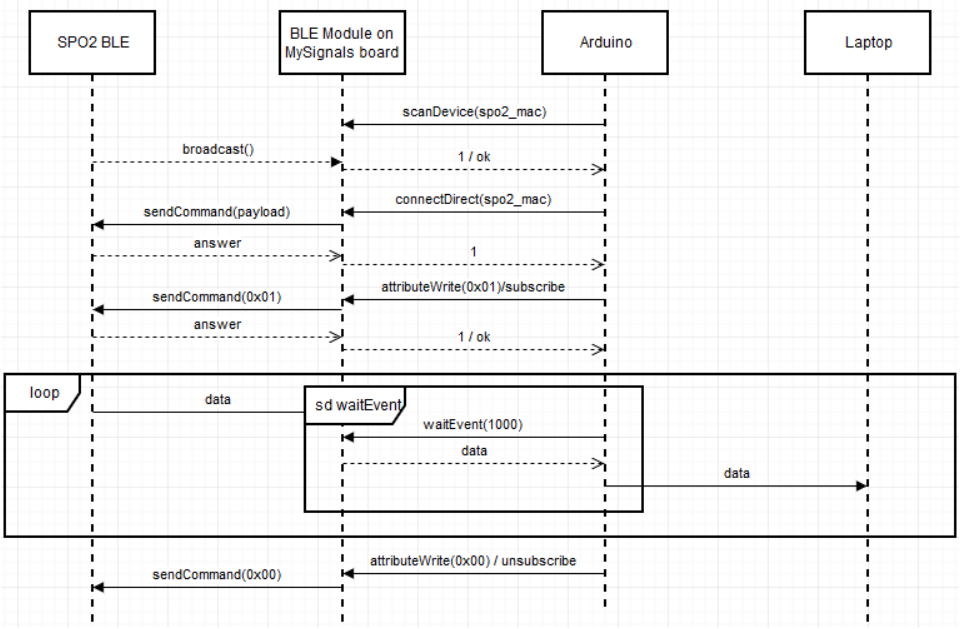

Figure 5.6: Flow model of protocol to connect and subscribe to the SPO2 device.

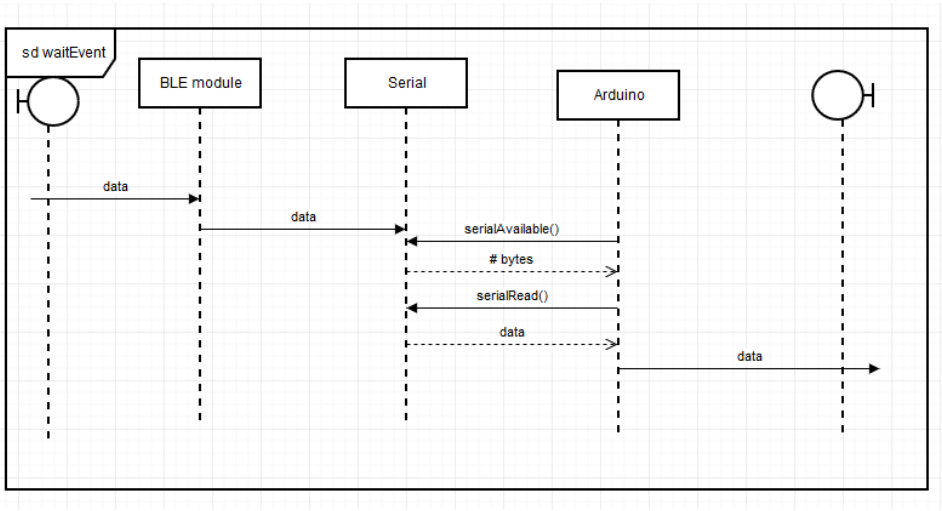

Figure 5.7: FLow model of waitEvent's internal functions

we do this, we hope to avoid a situation in which an entire benchmarking experiment becomes of no use because of an early error that could have been fixed by restarting the module. Other than that, we want all of the data received from the pulse oximeter to be at the highest possible sampling rate, with no filtering done by the Arduino processing unit. The output format is modified to present data as comma-separated values. The code is presented in Section B.2.

5.4.2 Data Characteristics

In the previous section we identified and chose a data acquisition method for Cooking Hacks' oximeter. By testing CH against NOX, we get two types of results that differ in character. Shown in Figure 5.21, the first is an example of a recording we can use as is in our benchmarking. The second type of results we discuss in this section.

Normally, healthy persons have 97±2% arterial oxygen saturation at sea level. We also know that to avoid volatile readings, it is common for pulse oximeters to have implemented averaging, and possibly other algorithms.

In the recordings from Cooking Hacks, we have sudden drops in the SpO_2 values, resulting in values around 0–50% SpO_2, as shown in Figure 5.8. Since the SpO_2 value is usually a mean value calculated over a period of time, we can assume it is unlikely that the SpO_2 levels drop and rise 50 to 90% in milliseconds. We therefore can define those figures as *invalid* or *out of bound values*. These values appear on certain conditions, as detailed in the Appendix. We cannot know the exact reason; however, we assume that it has to do with our design of the data acquisition tool, how Arduino communicates with the MySignals board, or/and how MySignals communicates with the pulse oximeter. An example appears in Figure 5.7, where we detail how the Arduino fetches data from the BLE serial buffer. The library code for fetching data from the pulse oximeter is done in two operations: first, checking available bytes in the buffer then fetching the data. A possible explanation of outliers in data might be that the Arduino fetches invalid data, caused by an empty buffer or a buffer containing garbage.

We further explore our theory by comparing data from CH with a simultaneous recording from NOX T3 (NOX). We place the NOX and CH pulse oximeters on index and ring finger, correspondingly. Then we both start and stop each of the recordings about the same time. We manually synchronize the two sets of data by comparing the SpO_2 values to best fit, and then plot the data.

The graphical plot is visualized in Figure 5.9. As mentioned in the last section, the new program we upload to the Arduino contains no value filtering. The plot therefore includes the outliers we discussed above, and the lines cover a huge portion of the graph.

When we plot a dotted representation instead, as seen in Figure 5.10, a clear pattern of values close to the NOX values in orange is visible. Most outliers are placed below 40% SpO and appear to be random. It is evidently

191769	97	65	-0.25
191951	97	65	-0.25
192134	14	0	-0.25
192316	97	65	-0.25
192498	97	65	-0.25
192680	5	4	-0.25
192863	97	65	-0.25
193045	25	0	-0.25
193227	97	64	-0.25
193410	97	64	-0.25
193592	14	0	-0.25
193774	97	64	-0.25
193957	97	64	-0.25
194140	97	64	-0.25
194322	97	64	-0.25
194504	25	0	-0.25
194686	97	64	-0.25
194869	97	64	-0.25
195051	14	0	-0.25
195233	24	77	-0.25
195416	97	64	-0.25
195598	97	64	-0.25

Figure 5.8: Cooking Hacks SpO_2 values, extract from recording

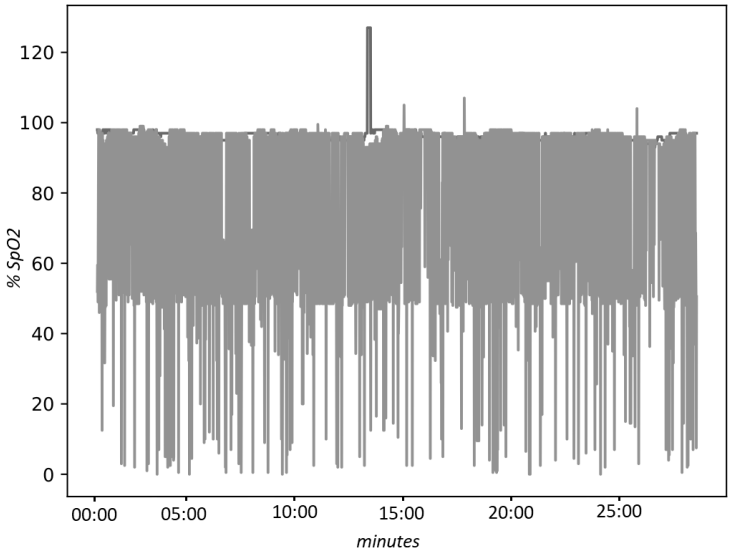

Figure 5.9: Cooking Hacks line representation of recording

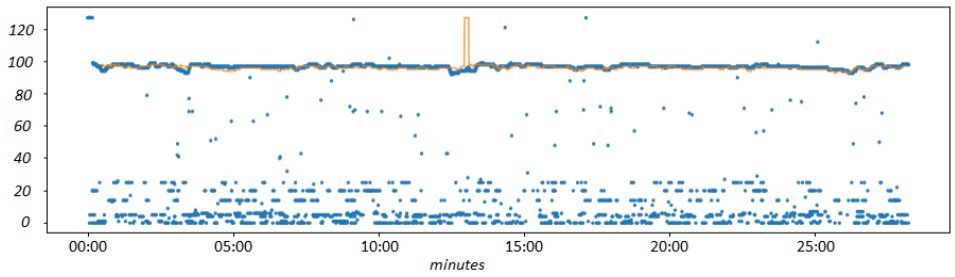

Figure 5.10: Cooking Hacks Dotted graph representation

not meaningful to compare the raw data from Cooking Hacks against data from NOX. Instead, we argue that outliers we define as invalid SpO2 values be removed with a simple filter.

In Section 4.4.2, we learned about the rate at which the oxygen saturation level falls under breath holding. It is unlikely for the oxygen saturation to drop more than 1% from one sample to the next, when the sampling rate is above 3Hz (3 per second). The filter therefore should be "drop values that fulfil the following condition":

```
1 filtered.spo2values = ch.spo2values[(ch.spo2values[i] -
      ch.spo2values[i-1]) > -2]
```

The filtered values only contains those where any one sample subtracted with the last sample is more than –2. We thereby remove values that constitute a drop in more than 1% SpO_2. We also remove all values above 100%. All other values are taken into account in the data analysis.

An alternative to simply removing values, would be to calculate new values for each we remove, e.g. by using interpolation. However, there is two reasons for why we do not think this solution is necessary. The first is the sample rate. In order to compare data samples from CH and NOX, we need to align samples in time. This means we have to either up- or downsample, and interpolate the values in the CH data. "Missing" values is therefore calculated in this process. The second reason is that the SpO_2 values only differs from the previous and subsequent values by 1. However, for most of the samples the change would be 0(as seen in the example data in Figure 5.8). We therefore argue that because of the small value difference, and the possibility of the values being garbage values as described earlier,, we see it as sufficient to filter out the values and leave the interpolation to the data synchronization process.

5.5 Synchronization

Each of the health sensor platforms uses different data acquisition methods. When we benchmark pulse oximeters, we need synchronization methods to compare data. In the chapter about pulse oximetry, we described the normal methods for testing pulse oximeters, and how the data samples are synchronized. However, the non-medical specification of our research does

Platform	NOX	Other
Cooking Hacks	Accelerometer, newline Nasal Airflow	Time, SpO_2 pattern
BITalino	Accelerometer, RIP, button	Time, SpO_2 pattern

Table 5.1: Overview of synchronization methods

not allow us to have stable periods of oxygen saturation achieved through breathing gas mixes. Instead, we have to explore alternatives.

In this section we look into synchronization methods involving other sensors, time and the SpO2 value itself. The sensors available for each platform are identified in Section 2.2.6. We use this information to define a matrix of possible synchronization methods between the platforms. The condition for a suitable sensor is that the output of the sensor from one platform must have the same characteristics as a corresponding sensor of NOX T3. To give an example, the breathing pattern from a nasal airway sensor has the same repeatable pattern as one from a RIP band; they both measure the respiration of a person. As for exclusion premise, sensors we use cannot be personal or not intended for reuse, as it would include additional expenses for equipment. ECG is an example of a sensor that we exclude as a synchronization sensor, as it uses disposable electrodes. Because of the time limit of our research, we also prefer synchronization methods that are easy to use and implement.

In Table 5.1 we see a summary of suitable synchronization sensors and other methods.

Column 1 contains the platforms, Column 2 the sensors in common with NOX, and the last column other alternatives to explore. Generally we see that we can use sensors that record acceleration and respiration. We also look into the possible consequences of the use of time and the SpO_2 value as synchronization methods. The rest of this chapter explores our alternatives and draws a conclusion on each of them.

5.5.1 Respiratory Synchronization

Since the test subject necessary breathes through the benchmarking process, it could prove useful to have a synchronization mechanism that involves respiratory patterns. For that reason, sensors that record breathing pattern are discussed in this section.

Cooking Hacks

MySignals HW comes with a Nasal Airflow sensor, Figure 5.11,for measuring breathing rates. The three thermocouple sensors placed on prongs measure both oral and nasal thermal airflow.

We start the test procedure to analyse the sensor data in relation to our use. First we copy the example code in the documentation and upload it to the MySignals board. Then we place the nasal airflow sensor according to the instructions and pictures in the documentation. Last we start the

Figure 5.11: Cooking Hacks' Nasal Airflow Sensor

sensor recording and observe the result to establish a sense of behavior of the sensor data as a subject of discussion.

The graphs in Figure 5.12 show the output from NOX RIP bands at the top and output from CH at the bottom. In this thesis we define NOX's output as the reference behavior, and the top subplot shows a normal breathing pattern expected from a person who is awake. On the other hand, the output from CH below does not show a clear breathing pattern. Watching the stream live and adjusting the placement of the sensor did not result in the desired behavior.

Our brief testing shows that the nasal airflow sensor was either too sensitive for us to identify the best placement for use, or that it is not able to provide a stable stream of data. On another note, using nasal and mouth sensors requires a high level of hygiene effort. Moreover, a closer inspection of the sensor reveals that it consists of movable/bendable parts of sensor technology that may be damaged while running benchmarking procedures on multiple subjects, potentially causing additional expenses for replacement if broken. As a result, Cooking Hack's nasal airflow sensor is impractical to use as a synchronization method in our study.

BITalino

BITalino include two RIP bands in their kit, both of which show strange behavior in the data. Presented in Figure 5.13, one recording shows the typical behavior of the signals from both RIP bands. The bands are placed at the indicated locations in the instructions, chest and belly, and there is no apparent difference between the two locations, or between bands. Values range between –50 and 50. At the start of a recording, the pattern may or may not show a breathing pattern. Generally, the data have three states: noise, dead or breathing pattern. At the start of the recording in Figure 5.13, we see 1.5 minutes of noise. Then we have 5 minutes with a more or less stable breathing pattern. Last, just before the 7-minute mark, the signals "dies," and locks up at –50.

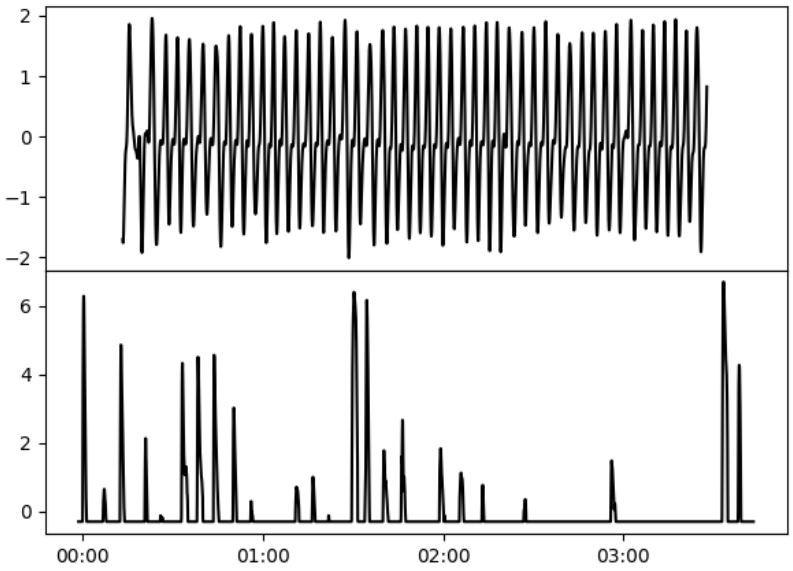

Figure 5.12: Graph of breathing pattern from NOX and CH

If we were to adjust the bands by either gently moving, stretching or twisting them, the signal may change from one of the mentioned states to another. In addition, if we are to achieve a breathing pattern, the signal is likely to eventually "lock" itself in the dead state until the band is adjusted.

We do not see it as our task to further explore or explain the behavior of BITalino RIP bands in our research. Nevertheless, the instability of their output proves them to be unfit for our use. If we were to use them, a likely consequence might be that we would not be able to synchronize some records with the NOX.

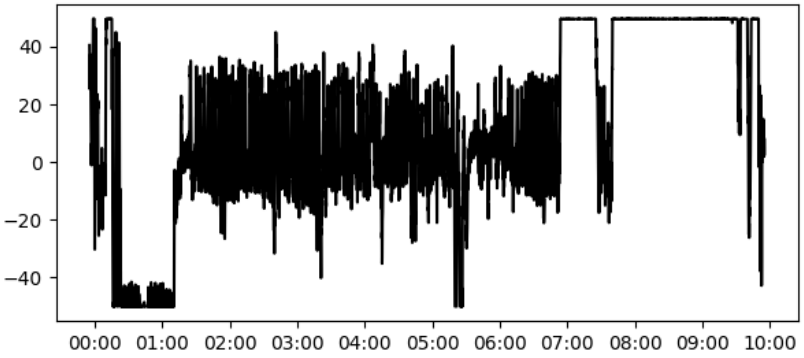

Figure 5.13: BITalino rip bands graph

Figure 5.14: Cooking Hacks body position sensor

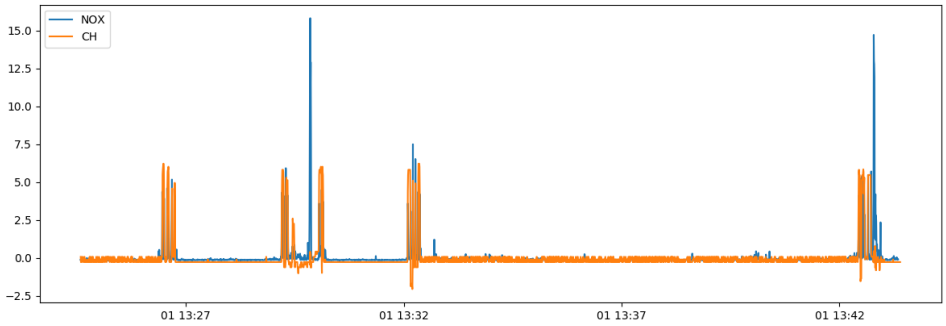

Figure 5.15: Synchronized accelerations from NOX and CH

5.5.2 Acceleration

All three platforms have an accelerometer available for possible use in synchronization. Unlike the two other platforms, NOX's accelerometer is inside the central unit. The sensor is used to indicate sleeping position when used for sleep study, as the central unit is usually attached to a t-shirt at the chest of the user. As we describe when speaking of data acquisition in this chapter, data from each sensor are separate, and even have different sampling rates. However, we assume the timestamp on the sensor data to be equally related in time.

Cooking Hacks

Cooking Hacks' body sensor is a belt that can be placed on a person to record their body position. It uses a three-axis accelerometer that provides individual values. However, we only need one axis to synchronize an event with NOX, which we specify in the code. We attach the accelerometers to the NOX central unit and begin the recording. We flip the sensors three times at the start, a few times in the middle, and three times at the end of the recording. Then we normalize the data around zero and plot it into the same graph. The accelerometer is shifted to where the start is aligned with

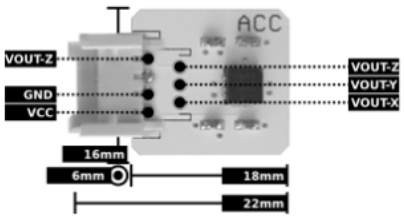

Figure 5.16: BITalino accelerometer

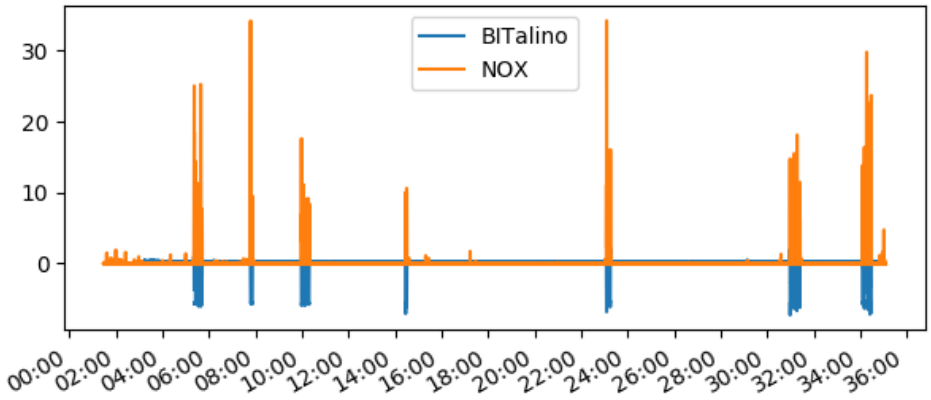

Figure 5.17: Accelerometer plot of NOX and BITalino.

NOX accelerometer data, indicated by the first three vertical pillars, as seen in Figure 5.15. The orange is the CH plot and the blue is the NOX plot.

Each sample from Cooking Hacks, contains all sensor data. Therefore, the accelerometer is related in time to each data sample from its pulse oximeter.

BITalino

In Figure 5.16 we see the accelerometer that can be connected to the BITalino board with a cable. It has a three-axis sensing feature.

We use same approach to test the acceleration sensor as we describe above for Cooking Hacks. In Figure 5.17 we also see the same type of pillars for BITalino (blue) and NOX (orange), although the values from BITalino are also located below zero. As for Cooking Hacks, each data sample from the platforms consists of data from all specified sensors.

5.5.3 Timestamp

In sensor data records, each sample is often paired with a timestamp. The timestamp might be from an internal clock, and it might represent actual or elapsed time. The platforms we use in our research each use different timestamp methods. The NOX uses its internal clock to apply timestamps on each sample, and BITalino has its timestamps from the Android with the

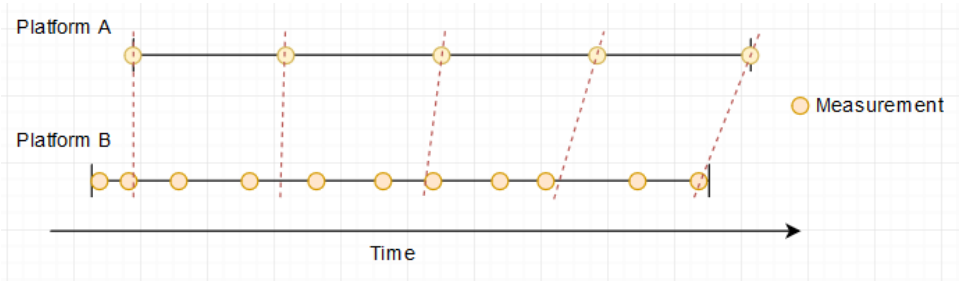

Figure 5.18: Example of the synchronization and time skew problem.

installed BITalino APP. The last platform, Cooking Hacks, uses the Arduino mills function, which represent milliseconds elapsed from the starting the device.

Using timestamp as synchronization introduces a challenge, as the clock might not run at the same speed. This might also be a problem in general when calculating accuracy, since we then compare two data streams against each other. The problem is illustrated in Figure 5.18. We have two platforms, A and B. The clock calculating the timestamp in platform B is faster than the clock for platform A. The stippled lines illustrate the correlation between the readings. Then we see that platform B is further and further behind as the time goes forward. If we were simply to compare the two data sets without testing the speed, we would not get the right results.

If we the clock of one platform to be faster than the other, it is possible to account for this by applying the difference to each sample. As an example, platform B is 1% faster than platform A, so we can add 1% to each data sample. the We continue to use NOX as the gold standard, and the time stamp of the two other platforms is compared against it. The accelerometers explored in the last section do clearly indicate acceleration events. Therefore, we investigate our results from them.

Cooking Hacks

In order to investigate the time skew, we take a closer look at the start and end of a plot from CH and NOX. Displayed in Figure 5.19, a zoom into the start shows four synchronized acceleration events. In blue, the NOX acceleration consists of either one or two vertical spikes. In orange, each CH event is represented by one vertical spike in the graph. At the same time, Figure 5.20 shows the end of the same recording. If we then look at the last spike from CH, we see that the difference is less than a second over this period of about 15 minutes. Our benchmarking experiments later in this paper are likely to be less than 20 minutes.

BITalino

Conducting a similar experiment with BITalino and NOX showed no visible time skew over a period of 30 minutes. This result might be

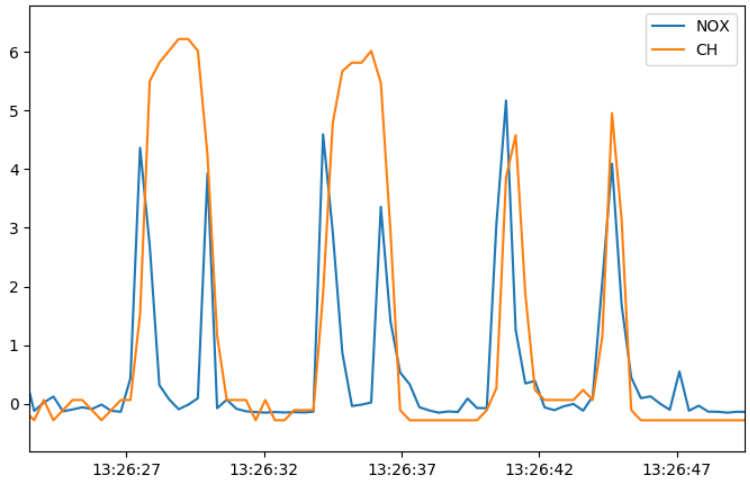

Figure 5.19: Accelerometer data from NOX and CH, at beginning.

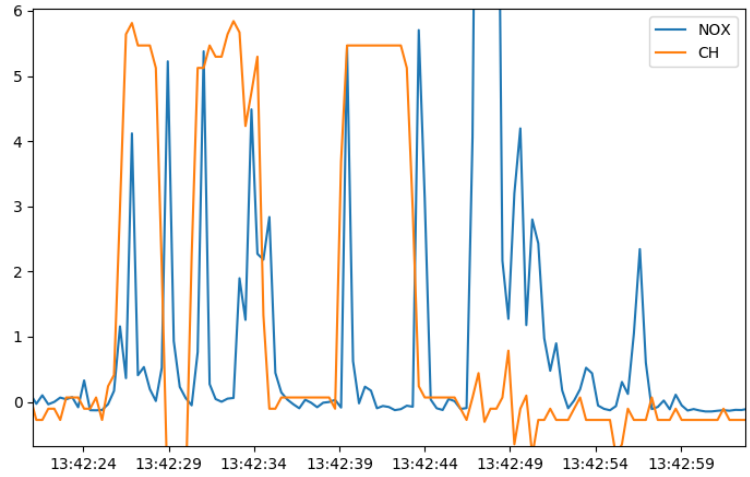

Figure 5.20: Accelerometer data from NOX and CH, at end.

explained by similar quality of the time hardware in both the NOX T3 central unit and the Android smartphone.

5.5.4 SpO2 Synchronization

It might be possible to synchronize the data using only the oxygen maturation. We know that by holding breath, SpO2 values would eventually start to fall. Repeated desaturation can make it possible to synchronize two sets of data using only the oxygen saturation.

Two 12-minute recordings from NOX and CH are presented in Figure 5.21. Here we see that it is possible to synchronize the desaturations by comparing them visually.

5.5.5 Sample Synchronization

When we find a suitable synchronization method for data, we are able to find start and end points in the data, and events are therefore synchronized. However, in order to compare data, the difference between two samples must be calculated. Each data sample can either be the raw measurement from a sensor, or an estimated sample. Since sampling rates from different platforms are not equal, one or both records are objects of resampling.

The time limit of our research does not allow us to investigate the best practice for resampling. However, we want to leave the data from NOX mostly unchanged. NOX has a fixed sample frequency set at new sample every 333ms, with 334ms every third. To achieve a fixed interval between samples, we downsample NOX to 334ms, as mean. With that as a basis, we have to downsample the data from the other two platforms to fit NOX because of their higher sample rate.

5.5.6 Summary

The synchronization methods we explore above show different degrees of success. Using the respiratory pattern as synchronization did not provide stable data for either CH or BITalino. The best method is found to be synchronization with an accelerometer. By simply plotting data into a graph it is easy to identify both the start and end of recordings.

The time skew experiment showed that that for CH there is a time skew of about <1 second in 20 minutes. In our short setting of benchmarking, we claim that <0.1% time skew is too short to affect the quality of our results. Nevertheless, even though the time skew is barely notable in the short time of our benchmarking experiments, the differences would have to be taken care of when recording a night's sleep over 7–9 hours.

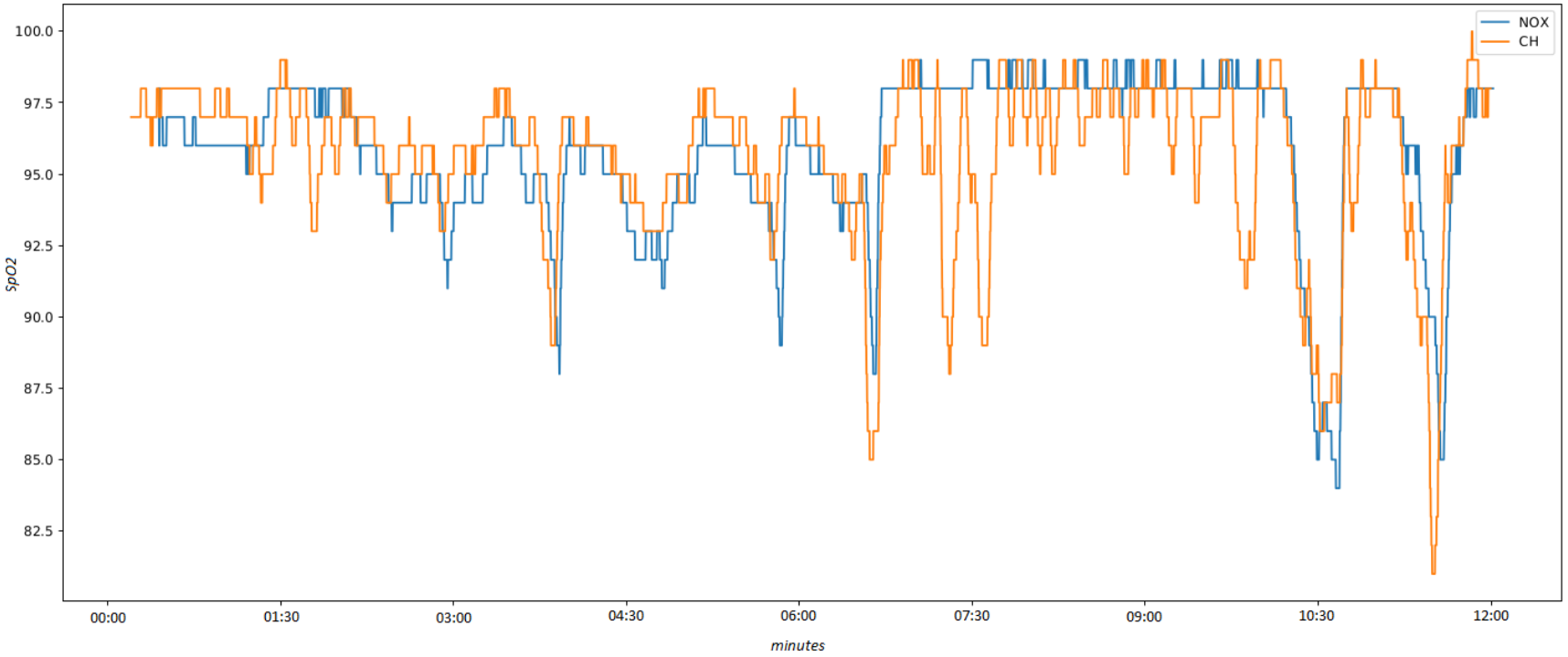

Figure 5.21: Comparison of CH and NOX pulse SpO_2 values

Chapter 6

Requirement Analysis

In Part I of the thesis we learned about health sensor technologies, pulse oximetry and obstructive sleep apnea. We identified important definitions, guidelines and procedures for our research. In order to better understand the technical challenges and what kind of data to expect, we did preliminary experiments, as described in the last chapter. In this chapter we analyse and present the requirements for the design of the experiments.

Section 6.1 goes through important requirement limitations, and the scope of our benchmarking tool. In Section 6.2 we discuss desaturation events. Data quality and defining the metrics for our analysis are the topics in Section 6.3. We define the requirements for the benchmarking protocol in Section 6.4, before we conclude this chapter with a summary of our requirements in Section 6.5.

6.1 Limitations and Scope

In this paper we develop a benchmarking environment and tool for pulse oximeters. Even though we want our tool to provide conclusive results, and to follow industry standards as much as possible, the nature of our non-invasive, non-medical testing method introduces some limitations. This section also covers the framework and scope of the experiments.

6.1.1 Non-invasiveness

As described in the last section, our protocol for accuracy testing is non-invasive. According to the ISO of 2017, implementing alternative, noninvasive methods for testing accuracy is possible, if not recommended, if the reference pulse oximeter is directly comparable to a CO-oximeter.

The content of benchmarking methods we define must dispense with a need for health personnel. Earlier in this paper we demonstrated that the common procedure to obtain comparable samples from a certain range of oxygen saturations includes the test subject breathing in gas mixes with a certain ratio between nitrogen and oxygen. Even though these sorts of gas mixes usually contain only gas already in the air that we normally breath,

our protocol cannot rely on having test subjects breathe gas mixtures that we offer. Likewise, the gold standard for testing accuracy is by matching the SpO_2 values with blood draws analysed with a CO oximeter, requiring medical assistance. Therefore, the methods we use in our benchmarking protocol are both non-invasive, and non-medical, which means that it should obviate the need for any medical guidance.

Furthermore, we are not in possession of medical equipment to use in the benchmarking process. In addition to the oximeters and platforms, the tools at our disposal are common devices such as smartphone and computer for data acquisition and analysis. Nevertheless, it is important that we achieve results that are comparable with the international standards mentioned.

6.1.2 Scope

To enhance the usability of our method in benchmarking pulse oximeters, the protocol should be generic and not depend on a specific technology. The oximeters mentioned in this paper are meant to represent a variety of pulse oximeters from a subset of the market. Our research aims to develop a benchmarking tool for testing inexpensive pulse oximeters without medical certification. They are ambulatory devices intended for home monitoring, or devices in a relevant development stage of home health monitoring.

Although we test oximeters from different sensor platforms, we do not benchmark or test the platforms themselves, such as their ease of use or grade of adequate documentation. Nonetheless, we describe possible behavior of the platforms or other circumstances that may affect our results in a report of the benchmarking experiment.

6.1.3 Test population and ethics

The test subjects we recruit should be healthy persons varying in age, sex and skin pigmentation according to the ISO of 2017. As physiological variations between the subjects may affect the results, we need to document their demographic details. At the same time, environmental circumstances such as location, temperature, and setup are also subjects of documentation.

Furthermore, it is important to define inclusion and exclusion criteria for the test population, according to the ISO of 2017, in addition to other potential illnesses that may cause a physical risk for the subjects. Therefore, each subject fills in a predefined health declaration before the experiment.

As specified in the last section and above, the experiments are designed to be carried out without the need for supervision by health personnel. The test population is specified so that subjects can be drawn easily from the general population, excluding those who might introduce medical risks. We mean to include specific instructions and precautions in the protocol to avoid medical issues. However, it is important to note that our research originates from a technological point of view. Even though we discuss

medical implications, and test procedures are explained in detail, we cannot know and do not mean to fully explore all of the health implications that may be introduced by using our benchmarking protocol. Therefore we emphasize that any participation in our work is strictly voluntary, and the accomplishment is founded on each person's own strength of will.

6.2 Desaturation Events

Sampling rate, response time, technical differences or external differences may cause data streams to be noisy and not easy comparable. To be consistent in our identification of events of desaturation and apnea in data, we also need to define the characteristics of desaturation events. In Section 4.2 we learn about the definitions of sleep apneas. With this in mind, we observe a cut from a recording done in preliminary experiments (Chapter 5), as displayed in Figure 6.1.

There we can see two plots from NOX (blue line with red sample dots), and CH (orange line with green sample dots) over a period of about 3 minutes. Light green lines show events from NOX, and purple lines events from CH. The terms used in the characteristics have the following meanings.

- *Baseline saturation*, the saturation level from where the desaturation event starts, and/or level after a post-event re-saturation.
- *Start* of a desaturation, the start of a period of desaturation from baseline saturation.
- *Nadir*, the bottom of the lowest value in a desaturation, succeeded by a (re)saturation.
- *End*, the point when saturation levels increase after the nadir.
- *Length* of the desaturation, from *start* to *stop*.

In Figure 6.1 we can see points of start, the nadir and the end of the desaturation event. According to the AASM manual, this desaturation event can be scored as an apnea, following the criteria of $\geq$3% desaturation, in a respiratory event of $\geq$10 seconds.

However, in the figure we see that the starting point of the apnea event is not at the indicated baseline saturation. The reason for this is that both sensors had a previous drop in oxygen saturation, but the event was either too short or the desaturations too low to be defined as an apnea.

To simplify the data examination process, we do not pay that much attention to baseline oxygen saturation in our analysis. We are still be able to identify desaturations. The length of the desaturations we define as the start to the end of the nadir. It is worth noting that our definitions may depart from those in the literature. Notwithstanding, desaturations alone are not enough to score apnea or hypoapnea in sleep studies.

To underline the importance of clear instructions on how to score a desaturation event and (equally important) to be able to tell them apart, we can examine Figure 6.2. Even though the figure is a cut of the same recording as the one we investigate above, it is not trivial in this

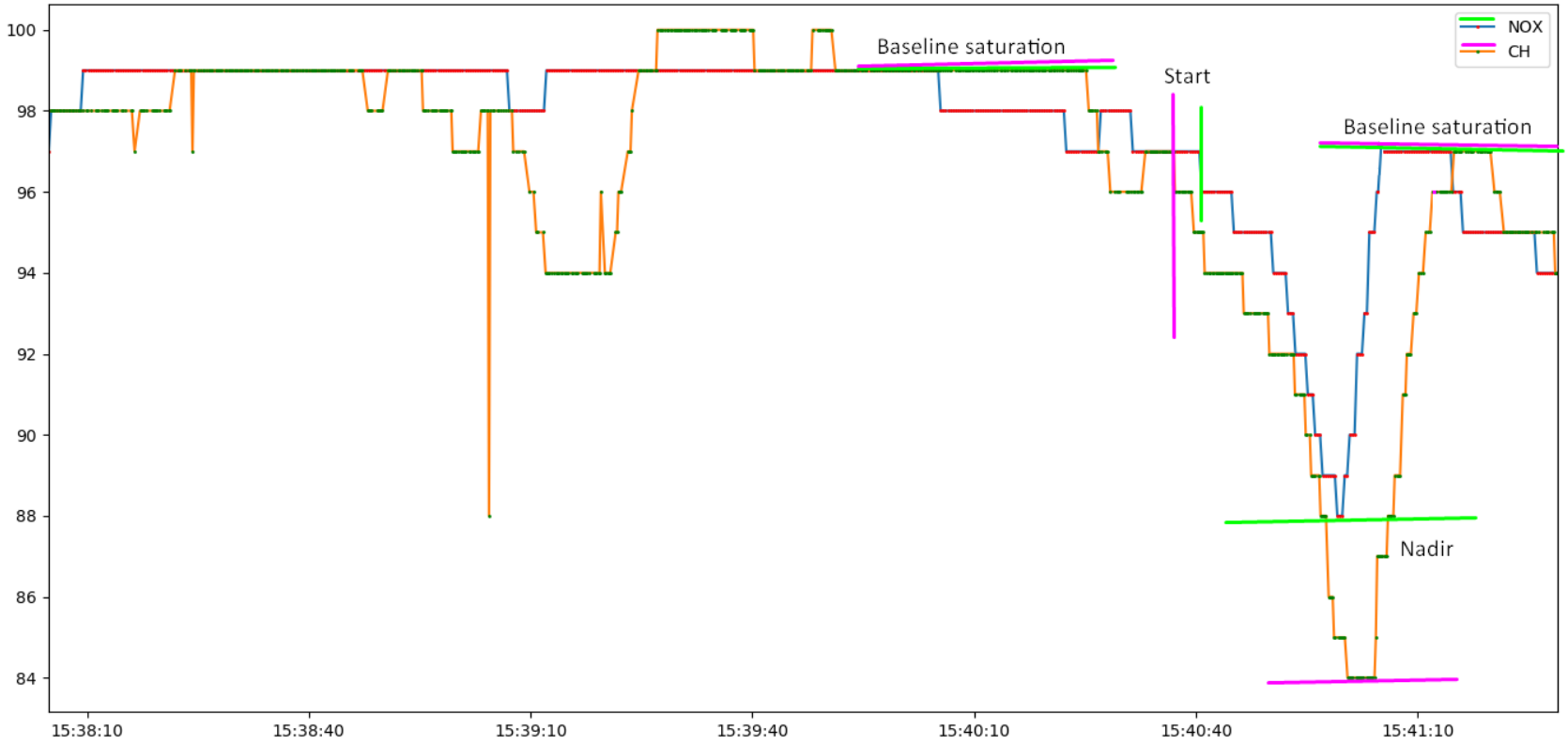

Figure 6.1: Apnea event definitions

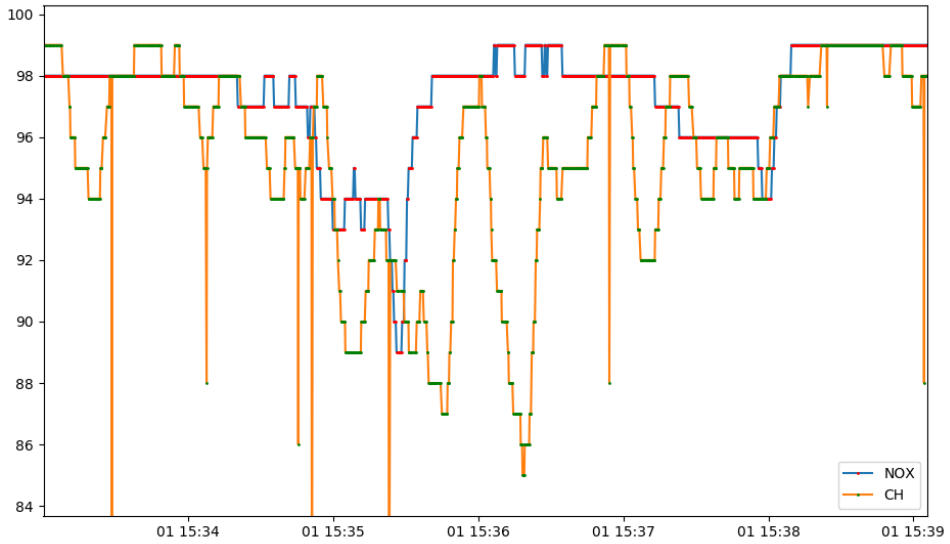

Figure 6.2: A second section of the same recording as Figure 6.1

case to pinpoint the starts and the following nadirs associated with each desaturation event. In the recording from which both figures are taken, the test subject did not follow a breathing script. The test person was lying in a supine position, simulating apneas at his own convenience. This might also be a case close to a real world setting, as frequent desaturations may simulate a person with a severe obstructive sleep apnea diagnosis. However, for the sake of our research, results such as those in Figure 6.2 can be very complicated to analyze.

The discussion above of indeterminable desaturation events leads to the conclusion that the testing procedure we design has to include specific instructions for the test subject to follow. The instructions must both facilitate the analysis process and enhance variety in SpO_2 values. This is the topic of the next section.

6.3 Data Quality and Metrics

As we discovered in earlier chapters, the related work on accuracy testing uses the practices and procedures recommended by guidelines from the FDA and the ISO of 2017. Most studies have used results from gold standard CO-oximeters to match against the tested pulse oximeter, and almost all of them use the method of breathing in a gas mixture to gain stable saturation plateaus for sample taking and matching.

Most pulse oximeters are calibrated according to defined standards; additionally, they can be certified for clinical use. Clinical certification is not specifically discussed in this thesis; however, we can assume that a certification at least includes the requirements of the ISO of 2011, as it is recommended by the FDA and often mentioned in the literature. Furthermore, we can also assume that accuracy testing is done by a third-

party laboratory and/or a control protocol is used.

On the contrary, the pulse oximeters within the scope of this thesis have not been tested by a third party. Therefore, as users of the oximeter, we cannot know with certainty the correctness of the recorded data. Earlier we referenced the CESAR project, which aims to develop an automatic sleep apnea diagnosis tool as an alternative to polysomnography, the gold standard of sleep diagnosis. CESAR depends on inexpensive health sensors aimed for the private market, and is not bound to specific or medical grades ones. It is conceivable that inaccurate data from an inexpensive oximeter might obscure the success and usability of such a project, since the tool may trigger incorrect decisions about a person's possible sleep disorder.

In addition to be able to measure the accuracy of pulse oximeters, it is also important to design the benchmarking tool with the goal of determining their usability in the detection of apneas. We want to determine a oximeter's rate of success for detecting desaturations or, in contrast, the rate at which it falsely indicates them.

As an example of the difference between accuracy and the ability to detect desaturations, we can inspect the following example. A doctor investigates the record of a night's sleep of one of her patients. Let us assume for simplicity that a pulse oximeter only was used in the monitoring, even though the use of additional channels is more common in home-monitoring devices. In preparation, the accuracy of the inexpensive pulse oximeter was determined. The result was within the predefined 4% root mean error, and the device was approved for use in the sleep study. Now, when examining the records, the doctor finds a only a few desaturations $\geq$3% per hour, too few to even indicate a mild sleep disorder. However, the reality of the situation was different. Throughout the night, this particular patient actually did have apneas frequently, with desaturations up to 4%. Still, they extended for only 10–12 seconds. Even though the averaging of the pulse oximeter was set at 3 seconds, design decisions from the manufacturer caused this oximeter to underestimate desaturations to avoid erroneous readings.

For this reason, our benchmarking tool should implement metrics to track and quantify behavior such as the one described above. The goal for our benchmarking tool is not to revise or disprove manufacturers' labeled accuracy values, and without including standardized methods our method is likely to be insufficient for doing so. Instead, we test the oximeters in relation to obstructive sleep apnea, using the international standards as guidelines. Accordingly, the metrics we define assess both (1) accuracy according to industry standard and (2) the oximeter's performance in detecting desaturations. Then, our research and benchmarking tool combine to act as a control instance, or a third party controller, for the implementation of inexpensive pulse oximeters in sleep apnea monitoring.

T

6.3.1 Accuracy

The standardized method for determining the quality of a pulse oximeter is to calculate the A_{rms} between the tested oximeter and the CO-oximeter used as the reference. We argue that this quality measure, known as accuracy, is also the most appropriate quality metric to be used in our benchmarking tool. We therefore calculate A_{rms} for each experiment and also as a total. In addition to determining the accuracy of the pulse oximeter, we also perform a Bland-Altman analysis. As stated earlier in this paper, this analysis method is common when comparing two methods of measurement. By using it we can gain valuable data to evaluate any over and underestimations. Here we calculate the limits of agreement, bias and precision. In addition, we implement a relation plot.

Earlier in this paper we point out that the accuracy of pulse oximeters tends to decrease towards lower oxygen saturation levels. Research shows that a pulse oximeter is most accurate near 100% SpO_2, and remains unchanged down to 70% at best. Because of the characteristics of our methods, we can assume that our quality measures will remain the upper bound, regardless of the metrics we choose. Therefore, even though we do not implement testing procedures recommended by international standards, we are at least able to determine the upper bound of quality.

This thesis has not investigated how much data are necessary to prove equally significant to the 200 data samples required in the international standards. The validity of our accuracy testing depends on the variations of SpO_2 values for all subjects. In advance of the experiments, we cannot know how the total spread in SpO_2 might be. Therefore, we leave the question of sufficient data quantity to be discussed when evaluating the benchmarking protocol.

6.3.2 Classify Desaturations

In addition to having a metric for accuracy, we see it as necessary to implement a classification system for desaturations. We need this in the analysis process of the results, to better understand the behavior of the pulse oximeters in measuring desaturations.

To quantify a pulse oximeter's ability to measure changes in saturations correctly, we use a binary classification system, including true positives, false positives and false negatives. We exclude the true negatives from our classifications system, as explained at the end of this section.

Our classification is strictly based on events we can identify in the data, by comparing synchronized recordings from the reference pulse oximeter and the test pulse oximeter. For each desaturation in the reference data, we classify either a true positive or a false negative. If the test object also detected a desaturation within the same time frame as the reference, it is labeled as true positive; otherwise, it is deemed a false negative. In contrast, desaturation indicated in the test object data, where there is no desaturation in the reference data, is labeled as a false positive.

Our classification system can be seen in Table 6.1. The first column

	Reference	Test object
True positives	1	1
False negative	1	0
False positive	0	1

Table 6.1: Desaturation Classification System

contains our classes, the second is desaturations identified by the reference pulse oximeter, and the last is desaturations found in the test object data.

Until now, we have purposely only used the term desaturation. A desaturation is defined as SpO_2 values of $\geq$3%, with no requirement of duration. On the other had, we also want to identify apneas in the data. We use the AASM definition that requires a $\geq$3% drop in oxygen saturation over a respiratory event period of $\geq$10 seconds. The reason we want to differentiate between the two of them is because we also want to analyze and discuss any patterns regarding the duration and depths of the drops in desaturation. Note that all desaturations are included in the calculation, not merely the ones accompanying simulated apneas.

The classification system described above is based on events identified in data. In this case, it is not easy to quantify a true negative. Let us say there are no desaturation events in the data from both oximeters. And let us quantify the true negative and give it the number 1; in 1 of 1 cases there are true negatives. Then let us assume 1 desaturation in the reference data and 2 desaturations in the test object data. Now we see the challenge: the reference data are no longer negative. Nevertheless, we can argue that the test object has at least one less true negative than the reference. Or to see it from another perspective, the test object has 1 more *false positive* than the reference.

Without dividing the data into fixed phases or time frames to quantify the true negatives, we cannot implement it into our binary classification system. Also, we argue that counting the false positives covers the absence of true negatives.

6.3.3 Procedures

The ISO of 2017 defines "procedures for non-invasive laboratory testing on healthy adult volunteers." It states that a pulse oximeter's values can be compared against a second one, if the latter is directly traceable to a CO-oximeter. Further, the data pairs acquired need to be sufficient to demonstrate the specified SpO_2 accuracy statistically. It also states the number of samples to be matched and the diversity of the study population.

The standard procedure for determining accuracy includes matching samples (data points) paired relative to time. This process is achieved by having test subjects breathe a specific gas mix, resulting in saturation plateaus. The stable periods of saturation make it convenient to pair data points, since the SpO_2 value is unlikely to be very volatile. However, as

defined earlier in this chapter, our methods exclude any use of medical equipment or personnel. Therefore, we have to consider our options.

The first choice is to record SpO_2 values from subjects, without simulating any apneas or provoking any desaturations. This option may be sufficient to measure the stability of the pulse oximeters, and to determine the rate of false positive desaturation. It is also possible to study additional literature and conduct statistical experiments to explain the behavior and the effect on accuracy towards lower SpO_2 values. However, since we want to determine a pulse oximeter's ability to detect desaturation correctly, a better experimental procedure would be to simulate apnea. Then, we would both be able to investigate the behavior of desaturations, and to measure the accuracy towards lower saturations levels.

Results from preliminary experiments show that comparing one raw data stream against a second may not be easy or best practice (at least without guidance, and subjects acting freely). In the chapter about OSA, we learned that desaturations is scored based on specific terms. After establishing a baseline saturation, an apnea is scored if the desaturation exceeds 10 seconds, and the drop from baseline saturation to the nadir is $\geq$3%. As noted in the literature, and experienced in the preliminary experiments, identifying a baseline saturation may not be trivial if the desaturations appear back-to-back without a break.

Even though we include the awake apnea simulations in our procedures, we have to assure that data from the test pulse and reference pulse oximeters are comparable. This can be assured by including synchronization techniques when possible, yet, the simulated apneas should be clearly distinguishable in time. We have to consider the length of the apneas, the possible design of the pulse oximeter, and the physiological processes on which oxygen saturation depends. The next section offers a deeper investigation into desaturation events.

6.4 Benchmarking protocol

The previous section illustrated the challenge of data sampling and clarified why we need specific instructions for testing pulse oximeters. As a start, we define all of the procedures we develop that are relevant to testing as being part of a benchmarking protocol. This protocol consists of elements that together take into account all of the requirements described in this chapter.

In general, we can divide the benchmarking protocol into two parts: preliminary tasks and benchmarking procedure. The goal for the former is to ensure that the latter is properly completed. Included in the first part are all of the preparations, considerations, necessary documents, and forms resulting from the requirements previously mentioned in this chapter. However, other necessary elements (such as recruiting for test population), or technical aspects (such as deciding the appropriate data acquisition method from each oximeter) are not part of the benchmarking protocol. Our protocol's purpose is as a guide for the completion of benchmarking

experiments.

The second part comprises the instructions for the benchmarking procedure. The ISO of 2017, and the guidelines from the FDA, recommend that 200 samples be paired to determine the accuracy of a pulse oximeter statistically. With a proposed test population of 10 subjects, 20 paired SpO_2 and SaO_2 samples are needed from each subject. In addition, the samples should be evenly spread in the range from 70 to 100%. Our non-invasive, non-medical method does not allow us to pair samples in the standardized manner. Instead we have to establish our own agreement of the amount of data required. When doing this, we need to consider the time frame of our research and the feasibility, in addition to how much data is sufficient to determine the quality of a pulse oximeter in the setting of OSA. Unfortunately, the time limit of our research does not allow us to delve into the theory of statistics. As a consequence, we want to collect as much data as possible, although, within the limits of the following two specific items. First, the actual benchmarking should be short enough to not be exhausting for the test subject, or lead to any medical risks. Second, we may assume that the ISO instructions have considered the total subjects to ensure diversity in the test populations, and it is therefore sufficient to recruit 10 test subjects.

Breathing Script

The breathing script has to take into consideration the physiological, ethical and technological limitations we identified earlier in this thesis. Then, within its limits, it should assess the two main tasks for the testing. In order to get the spread of SpO_2 values between 100% and down as low as to 80%, we have to enhance the possibility of lowering SpO_2 values. As we know from the theory of pulse oximetry, the partial pressure of oxygen in a person decreases over time as oxygen is unloaded from the hemoglobin to peripheral tissues. It is therefore essential that the breathing script include periods where the test subject halts the oxygen supply. In Section 4.4.2, we also learned that any effort to breathe or change the pressure in the lungs does not affect the rate of fall in saturation. Therefore, we see breath holding after a normal breath out as the best practice for this purpose. As we also discover in Section 4.4.2, the rate of fall might be determined by the initial saturation. The breathing script should therefore encourage breathing that potentially lowers the baseline saturation.

Ensure readability: The script must facilitate easy interpretation of the experimental results, as explained in Section 6.2. In addition to testing their overall accuracy, we also test the oximeters' ability to detect changes in the oxygen saturation when simulating OSA events.

6.5 Summary

Based on the discussions in this chapter, we identify the following general requirements for our benchmarking tool:

- We develop a tool for inexpensive, non-invasive, non-medical benchmarking of health sensors.
- There is no need for extra tools beyond the pulse oximeter sensor, e.g. - easy of use, structure
- Test population should be within the healthy general population
- The location of the experiments should be flexible
- We design and implement a benchmarking protocol, with breathing script.
- The benchmarking tool should be able to benchmark other sensors not covered in this essay with ease.

In addition to our general requirements, the requirements for our benchmarking protocol are as follows:

- It must ensure good benchmarking flow, through pre-experiments, guidance and practice.
- Accomplish goals for benchmarking script:
  - Aim for lower initial baseline oxygen saturation
  - Simulate 8 apneas, to meet ISO 80601-2017 accuracy guidelines.
  - To avoid ethical or medical issues, encourage lower oxygen saturations without pushing the subject.

Analysis requirements

- We should define a benchmark for pulse oximeters where one oximeter is the test object, one oximeter is directly traceable to a CO-oximeter as the reference, and measured the quality according to the accuracy standards in the ISO of 2017.

Chapter 7

Design

In Chapter 6, we analysed the requirements for our benchmarking tool. We also identified the limitations of developing non-invasive, non-medical quality testing of pulse oximeters. First, Section 7.1 contains the considerations for our design, including simulation of apnea, the test population, and the environment. In Section 7.2, we unveil the design a benchmarking protocol that includes proper premises, dependencies and experiment procedures.

7.1 Considerations

Our benchmarking depends on physiological processes, and in order to assure successful test results we now discuss considerations pertaining our benchmarking protocol. First we discuss the process and challenge of simulating apnea. Then we discuss and define the test population. We also summarize material presented earlier in this thesis on baseline oxygen saturation, before defining a benchmarking environment.

7.1.1 Simulating Apnea

As we found in Section 4, to score a hypoapnea a drop of $\geq$30% in sensor data must occur from nasal pressure or alternative hypoapnea sensor, with a duration $\geq$10 seconds and associated with a drop of 3% from the baseline oxygen saturation. In this context, we ideally want to test the pulse oximeter's ability to record oxygen desaturation when we simulate hypoapnea for 10 seconds or more. However, apart from its perhaps being (1) practically difficult for the test subject to properly simulate hypoapnea and (2) hard to give satisfying feedback and instruction in the process, we consider complete blockage of the airways to produce the best results. The potentially high initial baseline oxygen saturation may prove that simulating hypoapnea (also possibly incorrectly) might not be lowering the PaO_2 enough for a pulse oximeter to measure any drop in SpO_2. Also, as stated earlier, we want to test one particular pulse oximeter's ability to measure a drop in SpO_2, compared to a second reference one. Consequently, the most important goal for a breathing script is to enhance

the possibility of a fall in SpO_2 values. The benchmarking protocol therefore instructs test subjects to hold their breath from FRC with total blockage of airways.

7.1.2 Test Population

In our requirement analysis we specified that we want to recruit people from the general population, and according to ISO of 2017, the test subject must be healthy. According to the ISO standard study, the protocol should include rules for population inclusion/exclusion and for experiment termination. The former criteria are used for recruiting the test population. In addition, we designed a health declaration document (see Appendix C.3)

Inclusion criteria: The ISO of 2017 includes female and male subjects between 18-50, with ASA category 1.
Exclusion criteria: As in the standard, we exclude pregnant woman and smokers. Also, since we do not use medical personnel in our study, any persons self-reporting ongoing heart, lung, or brain problems/conditions are excluded. The test person must therefore sign a health declaration document before starting the benchmarking.
Termination criteria:: The test subjects can stop the experiment at any time.

7.1.3 Baseline Oxygen Saturation

The main purpose for this thesis is to investigate the possibility to benchmark pulse oximeters with a non-invasive method and without medical supervision. A possible conclusion of this research may be that benchmarking with our procedures is only possible with certain test subjects having particular physiological qualities. In this context, we should record and acquire additional data associated with oxygen saturation behavior. An example would be to monitor and record the oxygen saturation of a certain test subject during sleep. These experiments are not defined in the design, but are rather introduced when needed.

7.1.4 Environment

The benchmarking environment should be defined by the testbed and workbench. However, we allow our environment to be generic. The only requirement is that the location include furniture or place for the person to lie down. Next, external disturbances should be minimized to allow the individual subject to relax and stay focused.

Benchmarking configurations

As identified earlier, the sample rate should be at least 1Hz. We understand the sample rate as the time at which the pulse oximeter provides an updated SpO_2 calculation value as output. Further, from the preliminary tests we learned that pulse oximeter users are not given the option to adjust

averaging time. However, when possible, averaging should be the mean of 3 samples, or 3 seconds.

7.2 Benchmarking Protocol

In the process of benchmarking it is important that test objects have the same premises and environment. In the requirement analysis, we argued that we need a benchmarking protocol to assure both the correctness of the benchmarking output, and as a guide to follow for both setup and completion. We include both instructions for the prearrangements, such as documentation needs and forms to be filled out, and specific instructions for the participants in the benchmarking process. The benchmarking protocol is included in the Appendix. Each part is described below. We start with defining names for the roles and objects in the design, as follows.

- Test object: The pulse oximeter to be benchmarked
- Test standard: The pulse oximeter traceable to a CO-oximeter, used as ground truth
- Test population: All of the individuals recruited as test subjects
- Test subject: A person within the test population
- Test manager: Observer or test researcher

In Section 7.2.2, we first include the necessary preparations for the protocol. This is a routine for us to follow to avoid medical risks and ethical or technical challenges. Section 7.2.3 contains a guidance manual for the researcher to follow under the benchmarking process, while Section refch:design:script supplies the breathing guidance for the test subjects. The complete benchmarking protocol documents are included in Appendix C.

7.2.1 Project Description

Although we designed one benchmarking protocol, we need to implement two separate test documents, one each for the manager and the subject. The benchmarking description should be as informative as possible.

7.2.2 Prearrangements

First, to avoid any medical risk factors, we include a health declaration document for the test subjects to fill out. This document is not based on a study of the medical implications of holding the breath; rather, it checks for any illnesses or medical problems that may pose medical risks. Answering *yes* on any of the questions *exclude* the test subject from the experiments. The document is displayed in full in Appendix C.3

Next, we include the following instructions for the benchmarking manager. We do this to avoid any uncertainties, questions or breaks during the benchmarking process.

- Investigate the health declaration document before preceeding with the experiment.
- Carry out the technical preparations.
- Make sure that the test subject understand all parts of the benchmarking process, including the objective of the experiment.
- Remove any nail polish on affected fingers.

7.2.3 Benchmarking guidance

As we know, desaturations after simulated apneas are not shown instantly, with the response time being about 35 seconds to 1 minute. Therefore, it is not easy to guide the test person to achieve the desired baseline saturation. What we can do is observe the SpO_2 values after each apnea, and then give feedback to the test subject. Inform the test subject when 30 seconds are left to apnea. When it is time to hold the breath, make sure to specify that it should be held from breath out. Then notify the subject when 10 seconds have passed. Ideally, we want a spread in SpO_2 values between 100 and 70% from all experiments., Based both on experiments from the literature and from our earlier preliminary experiments, it is realistic that most subjects will be able to hold their breath for just above 10 seconds, thus lowering the saturation 10% at most.

However, as we know the rate of fall in oxygen saturation produces two positive effects: the initial baseline oxygen saturation at the start of the breath holding and its length. The test subject should therefore try to lower their SpO_2 values before each apnea. That means that calm, slow, and/or short breaths are encouraged. Also, the test subject should suppress the need for deep breaths, to avoid high levels of oxygen before an apnea. The second positive effect is the duration of the breath held. Even though all subjects should be able to hold their breath for 10 or more seconds, we expect that the total duration of breath holding to vary greatly from subject to subject. However, as the test population consists of healthy subjects only, we should also encourage the subjects to hold their breath for as long as possible, extending the 10 seconds. The discomfort that comes with the breath holding is not dangerous.

In our research we found no indication of the test subjects' positioning affecting the pulse oximeter's data quality. However, the subject should remain as still as possible under the benchmarking period, especially when holding his/her breath. To prevent the need for repositioning, subjects should therefore make themselves as comfortable and relaxed as possible. With that precaution, we hope to minimize possible movement artifacts in the data stream.

Our setup does not involve other sensors to monitor other physiological processes, such as the duration of the breath hold. The test manager has to visually observe and register the length of each apnea in an event document. Any movement, external disturbance or other unexpected event should also be noted there as well.

7.2.4 Breathing Script

Based on the requirements in Section 6.4, we describe our breathing script design in this section. An important requirement for the breathing script is that it should not be exhausting or cause any medical risks. Therefore, we limit the duration of the breathing script to 20 minutes or less. With a short time frame for the experiment, we ensure that the test subject stays focused.

Our investigation of the literature revealed (1) that subjects cannot hold their breath easily for more than 35 seconds and (2) that by repetition the breath holding time until the breakpoint of the breath hold can be improved. Therefore, to enhance the possibility of a spread in SpO_2 values, as many apneas as possible are implemented in the breathing script. As noted in the requirements, however, it is important that the apneas be visually distinguishable in the data. Furthermore, the breaks between the apnea simulations should be long enough for us to observe the results and give feedback before the countdown to the next apnea starts.

With these two limitations in mind, we propose the following breathing script: Begin each experiment with 3 minutes calm breathing to stabilize the breathing. Then every two minutes simulate an apnea from FRC (normal breath out). The test subject should be able to hold their breath for at least 10 seconds, which we define as a minimal duration. Every apnea, including the last, is followed by two minutes of relaxed breathing. In total, this process results in 8 apneas per test subject.

Definitions of events

- Breath-held - The test subject holds his/her breath for 10 seconds or more.
- Relaxed breathing - Subject takes long/slow "shallow" breaths.
- Withheld complete inhalation - Subject tries to withheld the urge for deep inhalation during relaxed breathing.
- Breath hold after exhalation - After regular exhalation, the test subject holds back the inhalation, i.e., performs a complete blockage of the airways, for at least 10 seconds.

Expected behavior

At each breath held period, the PaO_2 level will start to fall, and continue to fall as long as the test subject does not inhale. The change in SpO_2 values in the apnea period, measured by the pulse oximeter, depends on the initial baseline saturation and the duration of the breath held. At high initial levels, we might not be able to recognize any desaturation patterns, especially in situations where breath holds are just above 10 seconds. To our knowledge, the rise back to an old or new saturation baseline starts shortly after the test subject starts the inhalation. We do not see the changes of SpO_2 values in the data feed immediately after breath holding, as they depend on the averaging and response times.

By repeating the same procedure 8 times, the test subjects should also be able to learn more about their limits, and to push themself gently without pressure from the test manager. With that approach we avoid any ethical implications. We expect the test subject to improve in the duration of the breath holds and to achieve lower initial baseline oxygen saturation as the experiments proceed.

With this procedure we can achieve the recording of SpO_2 values containing up to 80 simulated apneas from 10 subjects, plus any other desaturation events. The grade of success of having SpO_2 values distributed between 70 and 100%, cannot be controlled. However, we evaluate our original design of the benchmarking protocol continuously, and implement improvements if needed.

7.2.5 Processing

As discussed, we do not see it as our objective to process signals from the test objects. We may apply filters to remove outliers that are unlikely to be real readings.

Chapter 8

Implementation of Tools

Filtering

In our research, we use a number of Python scripts for analysis and for visualizing our results and data. A reference to these scripts can be found in Appendix A.

Documents of Importance

We implement three different documents, beginning with the benchmarking instructions (C.2), which contain considerations and a guide for the researcher to follow. Next are the test subject instructions (C.2), along with a description of the purpose of the experiment. We also have a short health declaration document (C.3) to fulfil our exclusion criteria for nonhealthy test subjects. Last, we have an apnea event document (C.4) in order to note the apnea durations and other events.

Part III

Evaluation

Chapter 9

Evaluation

In the first part of this thesis, we covered the background information of health sensor platforms, pulse oximeters, and sleep apnea. Requirements for our benchmarking protocol were discussed, and we designed one that frames our non-invasive experiments to derive a method of benchmarking pulse oximeters in relation to sleep apnea.

Our methodology is based partly on theory from medicine. Our research does not mean to provide a complete picture of physiological processes; rather, we build our benchmarking protocol components on a basic medical understanding of relevant topics. Since our assumptions are mostly made by a high-level understanding of physiological processes, some of them may prove to be inaccurate, defective or, in worst case, wrong. For that reason, we divide the experiments into two parts. While the preliminary experiments in Chapter 5 focus on what data to expect as output for the given input on each platform, we test the output by using our benchmarking protocol in a Phase I of the experiments. Here we perform an analysis from the perspective of our expectations, and also discuss possible improvements to the benchmarking protocol. By doing this we hope to address any uncertainties, and also to gain the benefits of an advanced review of the benchmarking protocol before we do a larger-scale testing.

Section 9.1 contains phase one of the experiments, with a discussion of the first experiments, and suggestions for protocol improvements. An overview of the test population is presented in Section 9.2, before the synchronization is explained in Section 9.3. Section 9.4 contains an overview of the results, and calculations for their accuracy, and the Bland-Altman analysis.

9.1 Experiments Phase I

The goal for this part is to obtain an early review of our expectations towards the outcome of the experiments by using our benchmarking protocol and breathing script, and to evaluate the need for improvements and before we go deeper into the full set of experiments.

First, we test the three pulse oximeters from Nox Medical (NOX),

BITalino and Cooking Hacks (CH) previously reviewed in this paper, using two test subjects and our benchmarking protocol. As we see in the results section below, the first two benchmarking experiments are highly different in character. We therefore introduce side experiments, as mentioned in design(Section 7.1.3). These particular experiments are not part of the benchmarking of the pulse oximeters, however. The data used in the discussion of the results come from the first two experiments. Therefore, we include monitoring of the oxygen saturation during sleep of the first two test subjects, and then we examine the records to see if our assumptions of lower baseline SpO_2 levels during sleep at night are correct and useful for our benchmarking. Last, we conduct an experiment with a second and separately developed breathing script that aims to simulate breathing patterns during hypo- and central apnea.

The results of the two benchmarking, two sleep studies and one other breathing script are discussed in Section 9.1.1. The improvements on the benchmarking protocol is presented in Section 9.1.2. Last, we discuss the results from BITalino in Section 9.1.3.

9.1.1 Results and Discussion

In contrast to our analysis in the next part, in this first part we analyse each test individually, and we identify them by experiment number. It is useful to observe test results from each individual experiment to identify their different characteristics.

First, the data from BITalino and MySignals have element of noise in them. This was also discovered in the preliminary experiments, and we argued that we could therefore apply simple filters on the data sheets. For MySignals we apply the filter we detailed in Section 5.4.2. BITalino is discussed in its own section.

Experiment 1

Test subject 1 is self-reported as being a healthy nonsmoker in his mid-thirties. Before the experiment started, he had received neither breath-holding training nor familiarity with the script. The test was also started without initial training, knowledge or expectations about the oxygen saturation behavior. However, the subject did read the instructions document.

Figure 9.1 shows the output of the three pulse oximeters from NOX, CH and BITalino. The first thing we note in this graph, in general, is the absence of a desaturation period over 3% for NOX and CH. The output from BITalino is a matter of its own investigation in Section 9.1.3. When we omit BITalino from the plot in Figure 9.2 and use NOX as the ground truth, we see that the test subject has a high SpO_2 value, ranging from 98 to 96%. Our earlier models predict that with a high baseline oxygen saturation, holding the breath for 10-20 seconds may not be long enough for the PaO_2 to fall below a level measurable by a pulse oximeter. Also, from a high initial PaO_2 level, the rate of fall may be too low. CH has a desaturation

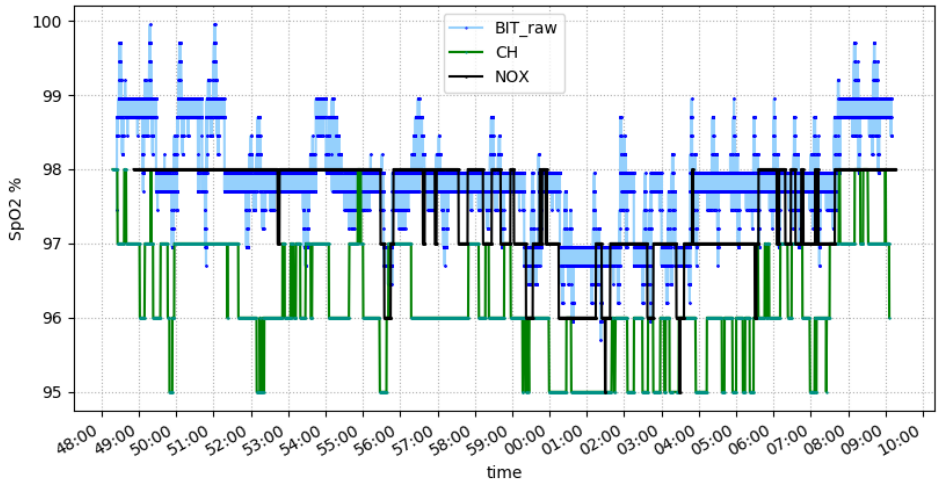

Figure 9.1: Output from NOX, CH and BITalino in Experiment 1.

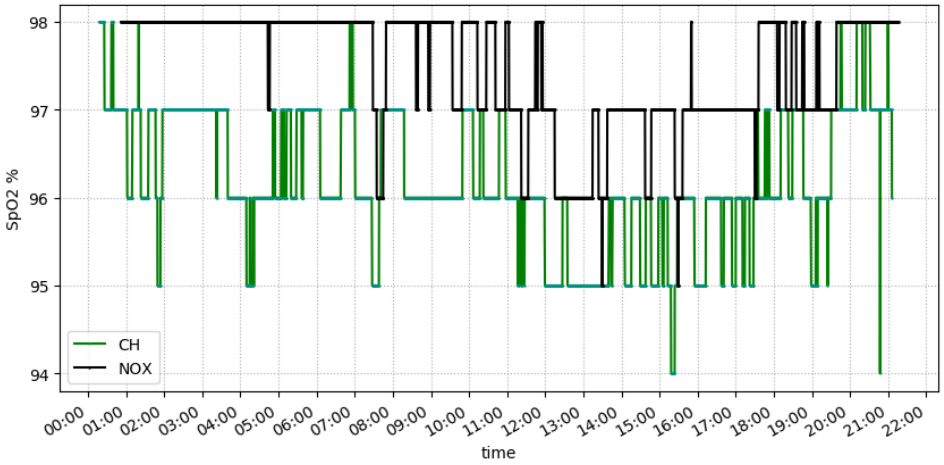

Figure 9.2: Graph of Experiment 1, with CH and NOX only

at the 7-minute mark, with a short (<10 seconds) top period of 98% SpO_2 before the 3% drop. However, according to the AASM Manual, the top might be adjusted to a mean baseline of 97% instead, and therefore no desaturation drops of $\geq$3% occur in this experiment. A third observation is that it is not possible to synchronize the data and graphs accurately by using the SpO_2 values, given no or few desaturation periods.

Follow-up questions about the subject's physical condition were asked for the purpose of gaining possible explanations for the test results. He reported a good physical condition, with a higher oxygen uptake than the general population. We also registered a low pulse throughout the test, with a mean of 54. In questions about the benchmarking process, the subject reported that he did know little about his progress in lowering the SpO_2 levels. These levels were reported by the experiment manager only after the third simulated apnea, as the subject was not able to see the

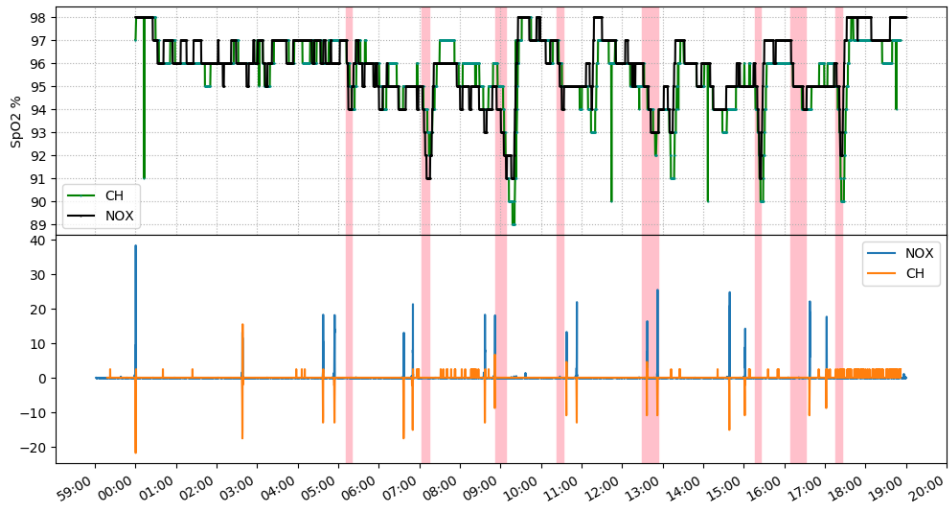

Figure 9.3: Results from Experiment 2, from NOX and CH.

display on the NOX wrist unit of the pulse oximeter. Furthermore, the health implications were not explicitly addressed in the information note, which added an element of uncertainty about the discomfort experienced.

Experiment 2

Subject 2 from the second experiment was a healthy non-smoker in his early 30s. In contrast to the first test subject, he was familiar with the breathing script and had prior experience in holding his breath, as described in the benchmarking protocol.

In Experiment 1, we learned that if the desaturation periods were few or absent, could not synchronize on the SpO2 values only. Therefore, we then used the accelerometer to synchronize and mark the apneas. In addition, we used the Noxturnal internal event detection tool as a visual indication of desaturation, and marked the desaturation periods *(pink)* in the figure with the results. In Figure 9.3 the top subplot shows SpO2 values from NOX and CH, including desaturation periods. On the bottom subplot, the aligned data from the two accelerometers are indicated by vertical lines, where the last 14 consist of start/stop pairs of simulated apneas. In the first experiment, the test subject held his breath for about 10-13 seconds, while Subject 2 was able to do so for about 10-24 seconds. As we know persons can extend their breaking point (included in background) with repeated actions. These data also suggest that Subject 2's previous training may have helped him hold his breath for longer from the start. In post-test questioning, he also reported being very conscious of his limits and of how to control his breathing.

As mentioned, the pink overlays in Figure 9.3 represent the Noxturnal's definition of the desaturation periods. The duration is the standard start until the end of an apnea where desaturations is $\geq$3%. If we then compare the pink desaturation with the apnea starts indicated by the accelerometers

	Duration		Desaturation	
Apnea	NOX	CH	NOX	CH
1	×	×	×	×
2	8s	12s	3%	3%
a	×	30s	×	3%
3	11s	11s	4%	3%
4	15s	25s	4%	7%
b	10s	32s	3%	3%
5	×	×	×	×
c	23s	18s	3%	4%
6	×	7s	×	3%
d	×	26s	×	4%
7	9s	12s	5%	6%
e	24s	×	3%	×
8	9s	9s	4%	6%
mean	13.5	17.5	3.5	4

Table 9.1: Overview of the desaturations in Experiment 2 from NOX and CH

in the lower subplot, we see that the desaturation periods are also occurring without a corresponding simulated apnea. This may be caused by the test subject's controlled shallow breathing. The desaturation events of both CH and NOX are presented in Table 9.1. In the first column we number the simulated apneas. The corresponding row contains possible desaturations that may appear after the apnea. The character a-e represents desaturation not resulting from apnea. If there is no desaturation, × is inserted instead of data. The desaturations are counted, both in total and after simulated apnea. The statistics are presented in Table 9.2. In total, 11 periods of matched desaturations with a SpO_2 drop of $\geq$3% are identified, with NOX present in 8 and CH in 10 of them. Under column desaturations, we see that 7 of the 10 desaturations for CH have a corresponding desaturation in the NOX data. One of the 8 desaturations from the NOX data is not present in the CH data, a false negative. The remaining three desaturations from CH are false positives, with no corresponding desaturations in NOX data.

In the column of AASM apneas, we count only desaturations we associated with apneas according to the AASM manual for scoring hypoapnea. This means that we count the apneas occurring about 30 seconds after at least a 10-second breath hold. In this case, NOX recorded 5 desaturations, and CH also recorded 5 at the same time. However, CH also recorded a 7-second desaturation of 3% after apnea number 6, whereas NOX did not. Instead of counting them as false positives, we instead argue that they are neither a true nor a false positive. Since we know an apnea occurred, it is likely that a desaturation was present in the blood. We therefore do not count false positives following simulated apneas.

When we observe the graph in Figure 9.3 in general, we can see a trend of CH having in SpO_2 value than estimates NOX in the desaturation

	Desaturation	AASM apneas
NOX T3	8	5
CH	10	6
True Positive	7	5
False Positive	3	×
False Negative	1	0

Table 9.2: Apnea events counted in Experiment 2

periods. In Table 9.1, we see that the mean of the events is longer for CH, and the fall in saturation is higher.

This disparity can be explained either by an overly careful estimation of the SpO2 value done by the NOX, or an overestimation of oxygen saturation changes by the CH oximeter. In Table 9.2, we see how Noxturnal classified events. In these cases, it is interesting that CH reported longer periods of desaturation. In Section 3.3.1, we explored the effect of averaging. However, as seen in the graph, the rates (steepness) of desaturation and (re)saturations appear to be similar for both NOX and CH. Also, a third explanation for the duration and desaturation in events might be placement. According to our requirements, the oximeter is placed on alternate fingers between experiments. This factor might be important for further analysis as data from additional experiments become available.

Experiment 3 and 5

In our benchmarking protocol, we regard it as important to lower the baseline oxygen saturation because of how it affects the SpO_2 values under breath-holding. We do this mainly for the practical reason that while breath-holding an oximeter fails to measure fall in PaO_2 until it reaches approximately 100 mm Hg, depending on factors such as pH and temperature. Based on data from work on apnea simulation, we further assume that the baseline oxygen saturation during sleep is normally lower than in the waking state [44]. For this reason, we saw it as valuable to record the SpO_2 levels during sleep of the first two test subjects. In doing so, we could both verify our assumptions, and possibly explain why different test subjects are more or less successful in holding their breath, or in achieving different baseline oxygen saturation.

The sleep recordings of the subjects suggest that our assumption of oxygen saturation being lower when sleeping than when awake state may be correct. However, the oxygen saturation did not fall after a period of time in which the subjects would presumably have fallen asleep. Rather, the data indicate that a test subject already has lower initial oxygen saturation at night before going to sleep. Figures 9.4 and 9.5 show sleep for Subjects 1 and 2. The top graph in both figures shows the first 30 minutes after turning on the device, and the bottom graph the entire recording. For Subject 1, 7 hours of sleep were recorded, while Subject 2, had only $\approx$3 hours, due to battery depletion.

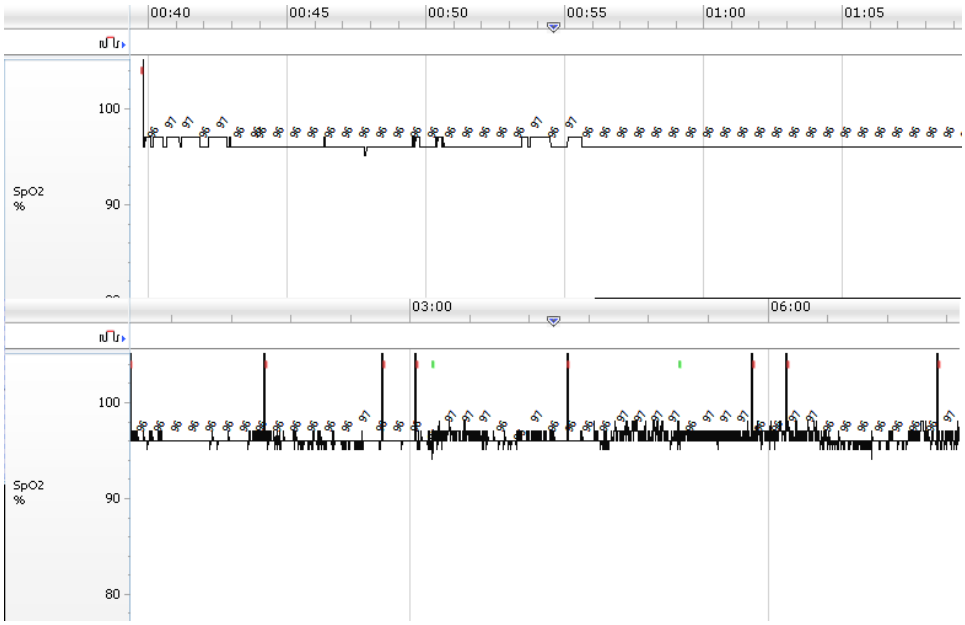

Figure 9.4: Experiment 3 result from Noxturnal

The data from both recordings indicate that the initial SpO_2 level in the first half hour guides the SpO_2 values for the rest of the recording. For Subject 1, the mean SpO_2 is 96%, with a mean of 97% in the first 30 minutes. Similarly, the mean of the first half hour is 95% for Subject 2, and the mean of all three hours is 94%. In neither of the two recordings does the SpO_2 fall as the time goes by.

In Section 3.1, we learned that the hemoglobin extinction curve moves depending on pH and temperature, thus affecting the SpO_2 readings at evening and night. In addition, other unknown physiological factors are likely to be causing a lower oxygen uptake at night. As a result, we cannot assume that a slower or calmer breathing alone, for example, is causing lower oxygen saturation during nocturnal sleep.

Experiment 4

As a second additional experiment, we wanted to test our defined breathing against a breathing script designed to simulate sleep apneas and record the effect on RIP bands [25]. In advance of the experiment, the difference between the breathing scripts can be seen as the cause and the effect of sleep apnea events. While our benchmarking measures the effect of apneas on the oxygen saturation of a person, the RIP breathing script is designed to simulate the breathing pattern of a person having apneas.

In our benchmarking research we test accuracy and the oximeters ability to record desaturation periods. A sleeping person is not voluntarily holding his or her breath, so the blockage of airways may continue for longer periods than the breaking point of a person being awake. For that reason, we found it more useful to have test subjects try to lower their

Figure 9.5: Experiment 5 result from Noxturnal

baseline oxygen saturation, and to hold their breath completely for as long as possible. Our assumption was, therefore, that by having Subject 2 test this script, the results would show little to no desaturation periods.

The RIP breathing script contains simulations of apnea, hypoapnea and deep breathing both lying on the back and on the side. For a person diagnosed with obstructive sleep apnea, all of these events can cause a drop in oxygen saturation. In total, there might be 8 periods of desaturation, located after the 3, 4, 5, 12, 13, and 14 minute marks.

Figure 9.6 shows a graph of the results of NOX, CH and BITalino in Experiment 4. The oxygen estimation from NOX T3 is 97 or above for the duration, and it measured no desaturation above 2%. On the other side, both CH and BITalino measured occasional drops in SpO_2 levels (e.g., after the 3-minute mark we can see that both have drops of 3% or more). However, examining the black plot for NOX only, no desaturations >3% SpO_2 occurred. Since the breathing script is tested on one subject only, we cannot determine that simulating both apnea and hypoapnea does not cause desaturations. However, Experiment 1 also indicated that holding the breath for just above 10 seconds might not be enough to lower oxygen saturation. Therefore, prolonging the duration of the events might cause different results.

9.1.2 Protocol Improvements

The experiments in this section show the breathing script's importance in the benchmarking experiment's degree of success. Furthermore, comparing the two subjects' tests indicates that their physical limitations and understanding of the test procedures are crucial. To address this gained knowledge, we need to include the following three specifications in our benchmarking protocol.

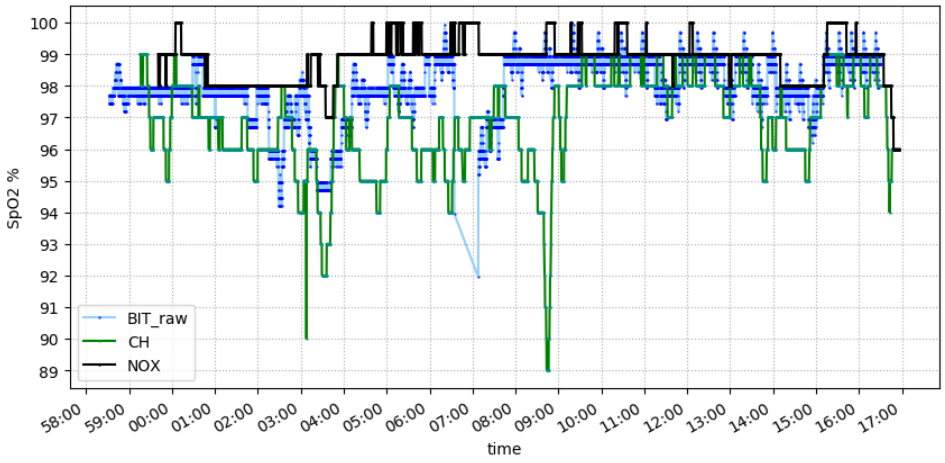

Figure 9.6: Experiment 4: respiratory breathing

Progress feedback

The first test subject said he was unaware of the progress of the desaturations. A method for subjects to see their SpO_2 values in real time is to be implemented. As the benchmarking is based partly on their strength of will and ability to control the breathing process, visual feedback could help them lower both the baseline SpO_2 and the length of the apneas. However, it is important to emphasize that the data are delayed by the response time, which may confuse and diminish the subject's ability to concentrate on the tasks.

Training

The points of distinctions in the results from the test subjects in Experiments 1 and 2 suggest that pre-experiment training in breath holding is more vital in the success of the benchmarking than first anticipated. As each experiment's discussion explains, the first test subject had no prior training (even in the lab). The second had experience with breath holding on FRC (Section 4.4.2), and knew his physiological limitations and how to control the breathing between the simulation of apneas. It is possible that the differences in the data sets in Experiment 1 and 2 could be caused and explained by physiological factors unknown to us. However, when we also take into account the feedback mentioned after the experiments, we deem it necessary to implement a training period before we start to benchmark the oximeters. Additionally, a period during which test subjects train in both breath-holding and practice calm breathing may be indicated before the benchmarking can arrive at the expected level of success. The test manager may also be able to provide better guidance if they known the subject's limitations and possibilities.

In the benchmarking protocol, we therefore include the following: After the preliminaries, a period of testing physiological reactions to the elements of the breathing script is to be carried through. In this period,

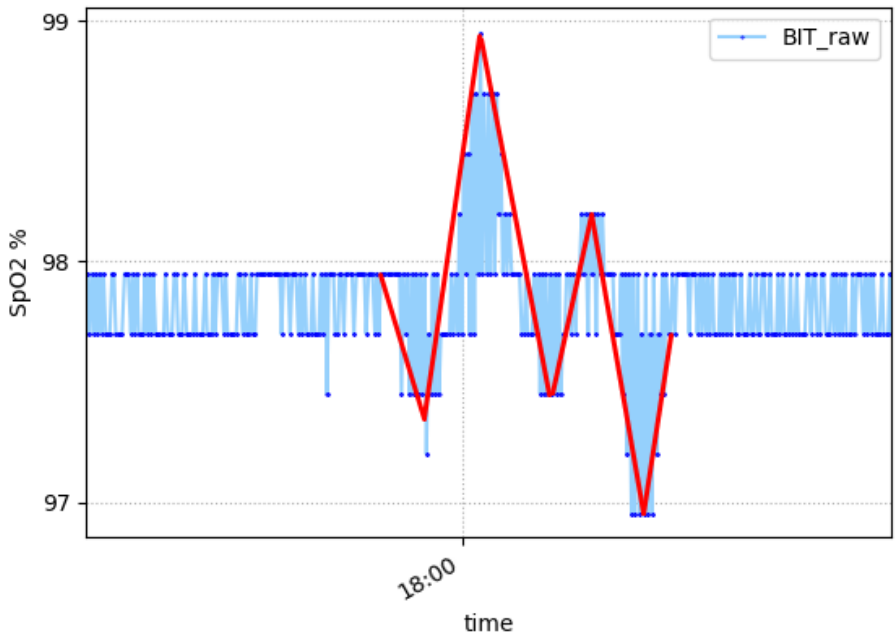

Figure 9.7: Cut from Experiment 1, NOX and BITalino

the test subject trains in both breath-holding and lowering baseline oxygen saturation, and the possible drops in SpO_2 values are observed. The training may proceed until the subject feels comfortable with a duration of breath holding that results in a fall in SpO_2 values. The test manager should note the hold time, baseline saturation levels, and oxygen drop to better guide the test subject while benchmarking.

Extended Information

The documentation for test subjects should include specific information about the risks of holding one's breath. As described earlier, we have no reason to believe that the experiments cause major health for the included test population. In most cases the participant might feel dizzy or disoriented. The document must emphasize that the experiment is strictly voluntary and can be terminated at any time by the test subject.

9.1.3 BITalino

Before our experiments in this part of the thesis, we received a new BITalino board is identical to those used in the preliminary experiments, although it came from an other production batch.

The result from Experiments 1 and 2 show that noise is still present in the data. As shown in Figure 9.7, a closer investigation of the signal reveals the same pattern (in red) discovered in the preliminary experiments. It might be possible to identify the trend through filtering, as seen in Figure 9.8. However, we specified earlier that we do not see it as our task to

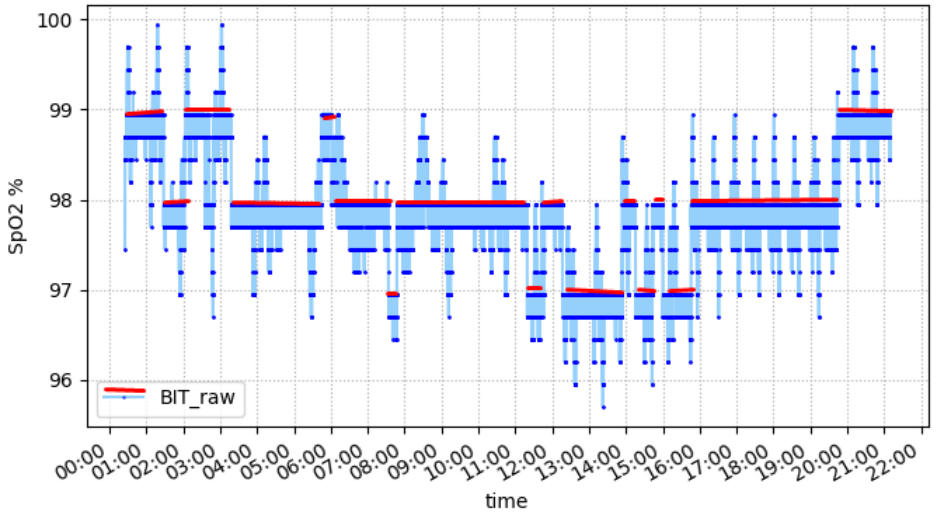

Figure 9.8: Experiment 1, trend line (red)

investigate possible noise filtering algorithms and apply them to the raw data. Therefore, it would also be difficult to determine if we found the "right" filter and values when the repeated noise pattern is higher than a rapid change in the saturation. In Figure 9.8 we can observe the pattern between 11 and 16 minutes (we inserted trend lines manually to highlight the possible change in saturation). As we see, the noise (in vertical blue lines) has a higher range than the assumed saturation change between 98 and 99. Therefore, we cannot know with certainly the difference between a noise pattern and the start of a desaturation.

In our work with BITalino, we depended on earlier work for data acquisition. From the preliminary experiments, we became aware of problems with the connection between the mobile device and the BITalino board. In addition to the problems of starting the data acquisition process, we also experienced the Collector app halting during a benchmarking experiment. At this stage, the time limit of our research does not allow us to investigate the reason for the app not working, nor to develop other methods for data acquisition. Therefore, we made a decision on the using the BITalino pulse oximeter before the rest of the benchmarking experiments.

It is possible to calculate the accuracy of the data from BITalino and to perform a Bland-Altman analysis on it. However, we do not see it as valuable, as the state of the data indicate that the device is not working as intended. Applying our apnea detection analysis would also be a tedious task, if even possible. The pulse oximeter is a stand-alone device without the need of the BITalino board; therefore, one can assume that the data noise is a result of the latter's malfunction. We also experienced problems with collecting data, making the experiments last longer than necessary. Therefore, we omit BITalino from future experiments.

9.2 Test population

Ten healthy subjects were recruited for the benchmarking experiment: 5 women and 5 men, ranging between 26 to 54 years, with a mean of 35. A level 2 was considered as medium skin colour and 8 light skin. We were not able to recruit test subjects characterized by darkly pigmented skin. However, we found this to be an optional configuration, as dark pigmentation is likely to have a maximum effect of 10% on the quality below 80% [15] . All subjects were able to complete the experiment and simulate the apneas as specified.

Demographics	
Number of subjects	10
Age in years	35+-19
Sex	
Male	5
Female	5
Skin tone	
Light	8
Medium	2

9.3 Synchronization and samples

Earlier we specify that we use an accelerometer for synchronization. However, plots show that this synchronization should be considered as guidance only, as the internal implementation of the pulse oximeter might not be the same. The delivery time for data acquisition or the time it take to store data may therefore differ. In data where the rate of the desaturation is steep, synchronization of data is utterly important. Slightly shifting the data forward or backwards produces highly different accuracy estimates. Therefore, we take advantage of our already implemented calculation of accuracy, the A_{rms}. After synchronizing the data with the accelerometer, we shift the CH data set 200ms forwards or backwards, until A_{rms} reaches its lowest point.

We can observe the effect by using data from Subject 3 as an example. First, we synchronize the plot in Figure 9.9 with the accelerometer, obtaining an Arms at ≈2.9%. However, by looking at the data plot, we see that the graphs may be aligned better in SpO2 values. Therefore, we shift the CH data approximately 6 seconds to the left, obtaining an A_{rms} ≈2.1%. Now the graphs are much more aligned, as we see in Figure 9.10. The differences in accuracy between the two results is 0.8% , a 32% difference. We also chose Subject 3 as example because of the steepness of the desaturations. Seconds of wrongful synchronization may result in percentage difference to best fit synchronization. However, a flatter graph such as shown in Figure 9.2 will not be affected as much by shifting the CH data.

In addition to shifting the alignment between the two data sets, we

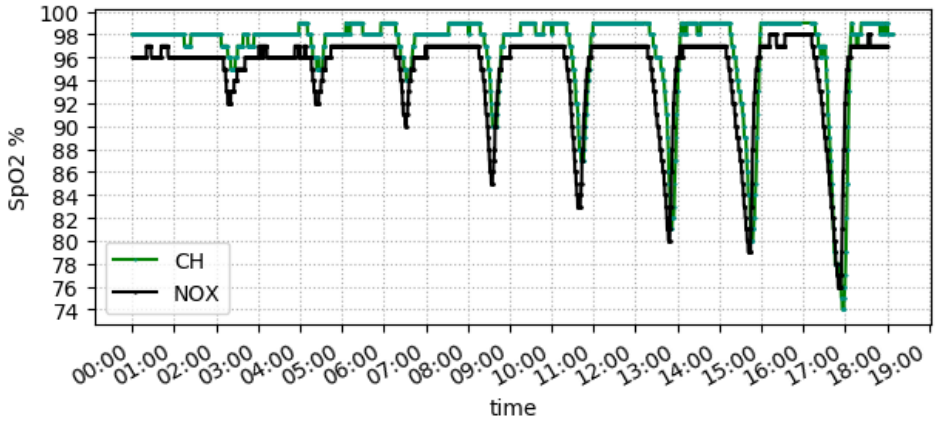

Figure 9.9: Subject 3 plot with NOX and CH

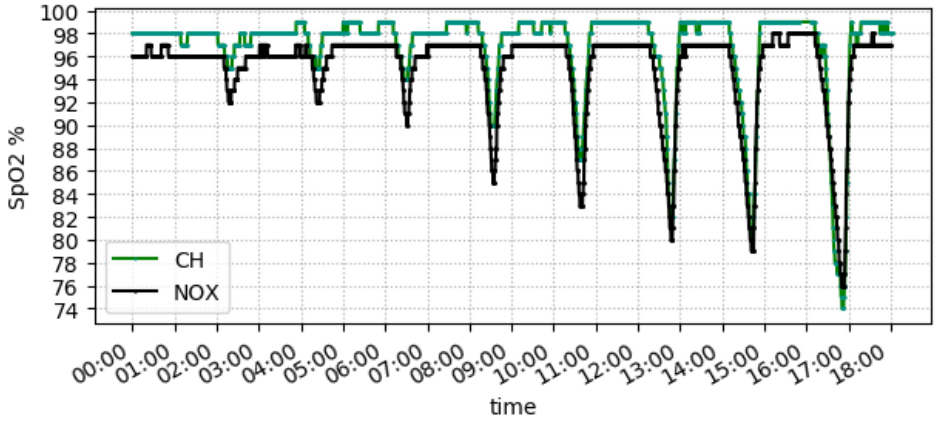

Figure 9.10: Subject 3 plot with NOX and CH, shifted

use the 19 minutes in our recording that gives us most stable values. For instance, pulse oximeters dropping out at the start or end of the recording allows us to shift the graph to the side. Also, the test subject tended to move a little to adjust their position at the start of the recording. As a consequence, apneas in data might be up to a minute earlier than expected according to time labels in the plots showing results. For relation plots and Bland-Altman analysis, 3,250 samples were paired for each subject. In total 3,254 data samples were paired, with the distribution displayed in Figure 9.11.

9.4 Accuracy

We have specified two tasks for the analysis of our results. The first is determining the accuracy of the pulse oximeter, and then performing an Bland-Altman analysis of each individual subject, and all subjects combined into one data sheet. First we present the results in Section 9.4.1,

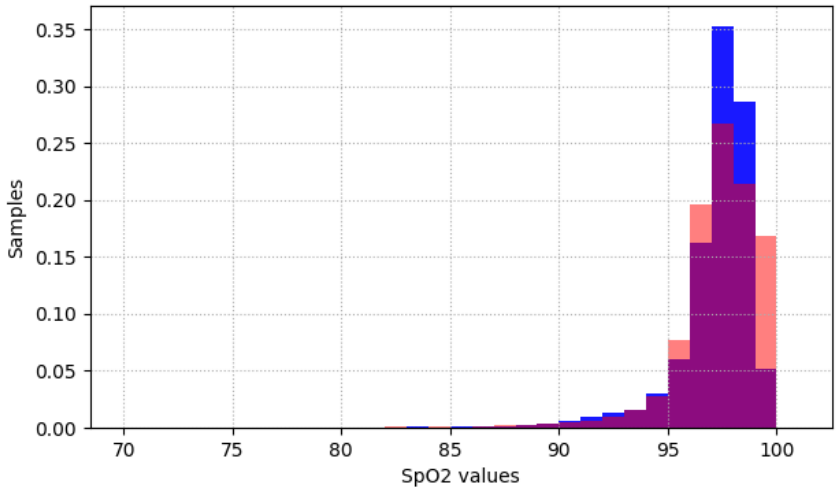

Figure 9.11: Histogram of all results

Subject no.	Accuracy	Precision	Mean Bias	Upper LoA	Lower LoA
1	1.46	1.16	-1.33	-0.17	-2.49
2	0.87	1.66	-0.22	1.44	-1.88
3	2.09	2.10	1.80	3.90	-0.30
4	1.51	2.78	0.51	3.29	-2.27
5	1.10	1.99	-0.44	1.55	-2.43
6	1.31	2.05	-0.43	1.63	-2.48
7	1.86	1.05	1.78	2.83	0.73
8	0.74	1.47	0.01	1.48	-1.46
9	0.88	1.67	-0.19	1.48	-1.86
10	1.08	2.11	0.04	2.15	-2.07
Mean	1.29 ($\pm$0.8)	1.78 ($\pm$1)	0.18 ($\pm$1.62)	1.96 ($\pm$2.13)	-1.65 ($\pm$2.38)
All	1.34	2.61	0.14	2.75	-2.47

Table 9.3: Accuracy results for each subject

before discussing them in Section 9.4.2.

9.4.1 Results

Table 9.3 show all of the results from the experiments. For each row we find calculated values for accuracy (A_{rms}), precision (2 standard deviation of difference), and mean bias(mean of the difference) for each individual subject. We also provide the upper and lower limits of agreement for all experiments. In last chapter we discussed individual experiments. However, for the rest of this paper we refer experiments that are a part of our test pool by naming the subject id. Each row in the figure labeled with Subject is therefore a experiment counted into the test pool.

In the second column we see the calculated accuracy for each individual subject. The pulse oximeter from Cooking Hacks has a labeled accuracy of

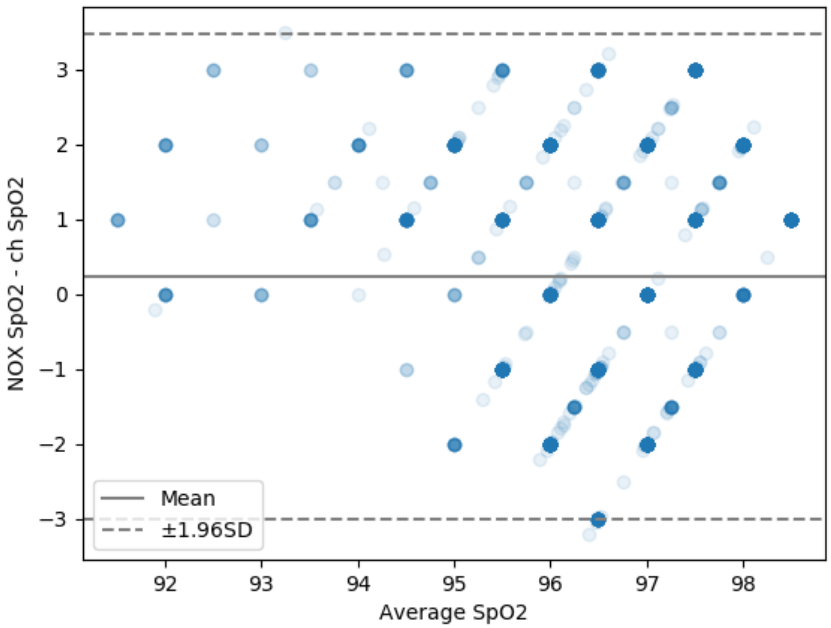

Figure 9.12: Bland-Altman plot of Subject 1 and 7.

2% between 100 and 80% and 3% between 80 and 70%. However, since little of our data falls below 80%, and only for one subject, it is not necessary to inspect or calculate these two ranges separately.

In the table we see that all subjects but Subject 3 have a calculated accuracy under 2%, and even Subject 3 is just above 2%, with 0.09%. The mean accuracy of Cooking Hacks against NOX is 1.29% for all subjects, within a range of $\pm$0.8%. A calculation of all samples combined shows a total 1.34%.

In the third column, the mean precision for the subjects is 1.788%, within a range of ±1%, while all test subjects combined is 2.61%. Where the combined accuracy varied from the mean accuracy by <0.05%, the difference between the calculated combined precision and the mean precision is 0.83% points. This can be explained as follows. While Bland-Altman plot for each subject is concentrated around their individual mean, a plot where all data are combined would result in a spread around the mean difference. To illustrate this effect, we can observe a Bland-Altman plot of Subject 1 and Subject 7 combined, with the mean bias being -1.33 and 1.78, respectively. The result, shown in Figure 9.12, is a mean bias of 0.23, which is also the mean of adding the two individual precisions. However, the precision of the combined data is 3.24, in contrast to the mean of adding their precision, which is 1.12. Related work does not include calculations of the mean; neither is the use of Bland-Altman analysis unique for our work of comparing data from pulse oximeters. However, in standard accuracy testing procedures it is common to compare just above 200 paired data

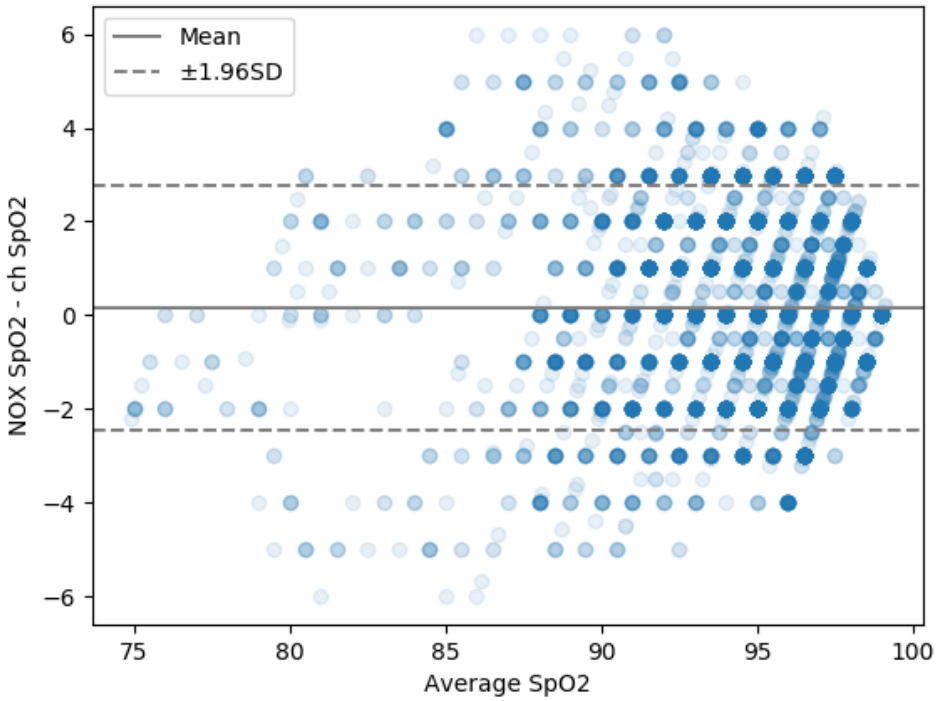

Figure 9.13: Bland-Altman plot of all results

samples equally spread between 70 and 100%. In comparison, our Bland-Altman plot contains 32,480 paired data samples, with most of them falling between 100 and 90% (as seen in Figure 9.11). This also means that the top 2–3% also has most weight in the calculation of the precision and the limits of agreement. In the figure, the values between 95 and 98% are spread around the mean bias, while the lower 3% is mostly between the mean bias and the upper limits of agreement. For this reason, we see it necessary to highlight both the combined results and the mean of the adding them. It is also interesting to analyse the difference between accuracy and the precision. Accuracy is the mean difference between each individual paired sample.

In the Bland-Altman plot of all results (Figure 9.13) we see a mean bias of 0.14%, and 2.75 and -2.47 as the upper and lower limits of agreement respectively. This means that 95% of the readings are expected to be within $\pm$ 2.61% of 0.14. The standard deviation, $\approx$ 1.30, is close to the accuracy of 1.34%.

9.4.2 Accuracy v. Bland-Altman

Accuracy, the A_{rms}, is the standard metric in the industry to determine the quality of a pulse oximeter. However, this metric is usually calculated using the gold standard method of blood draws on stable saturation plateaus. Our calculation is from two sets of continuous data, where about 2/3 of the data values are either 97 or 98% SpO_2 for both NOX and CH.

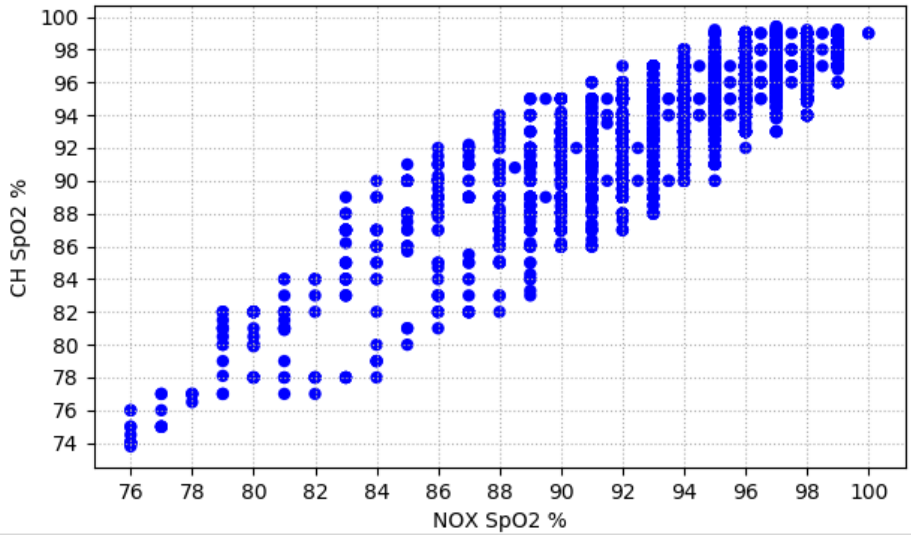

Figure 9.14: Relation plot of all results

Therefore, it is useful to compare the results of the accuracy calculation to Bland-Altman analysis. Whereas Arms is calculated by comparing the test oximeter against a reference, Bland-Altman compares the test oximeter against a mean of the test and reference oximeters. In Table 9.3, the lowest accuracy estimation comes from the test subject that achieved the highest fall in saturation values, Subject 3. This might be explained by degrading accuracy against lower levels. However, an inspection of the results provides us with an alternative explanation. First we see in the graph in Figure 9.10 that the SpO_2 values at the baseline saturations are already off by 2% before each desaturation. This seems to be a consequence of a systematic overestimation of the saturation in this subject. The relational plot in Figure 9.10 confirms this, where each dot is a paired data sample, with NOX in the x axis and CH on the y axis. The trend line in black shows most CH values being about 2% above the NOX values. Figure 9.16 show the Bland-Altman plot of Subject 3, with a calculated mean bias of 1.8% and precision at 2.1%. We also see some values as low as –5%, and as high as 6%. This is the highest spread in any of the results of any subject. However, the relational plot in Figure 9.15 shows a almost linear trend between the two data sets.

Then we can investigate the results from Subject 4, which has a higher accuracy but a lower precision. The plot of NOX and CH in Figure 9.17 shows a graph in which CH moves more between values than we saw from Subject 3 in Figure 9.10. Also, while the relational plot in Figure 9.15 shows an almost 1.0 ratio trend line between CH and NOX, the same plot for Subject 4 shows a 0.5 trend line. Then we can observe the Bland-Altman plot for Subject 4 in Figure 9.19, and the same plot in Figure 9.16 for Subject 3. In these graphs we see that the precision for Subject 3 is 0.68% higher than for Subject 4. Even so, the accuracy of Subject 4 is 0.58 higher than for Subject 3.

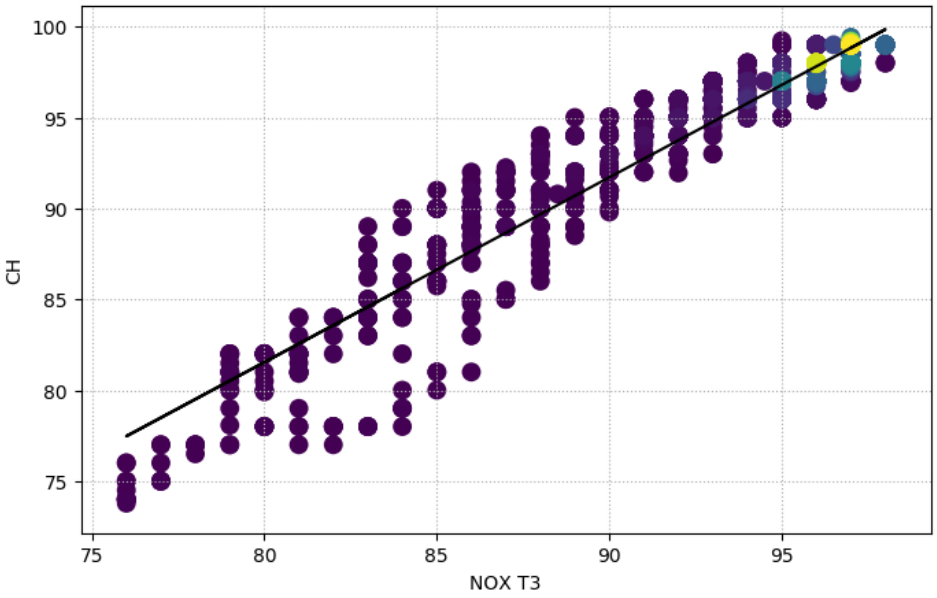

Figure 9.15: Subject 3 relation plot

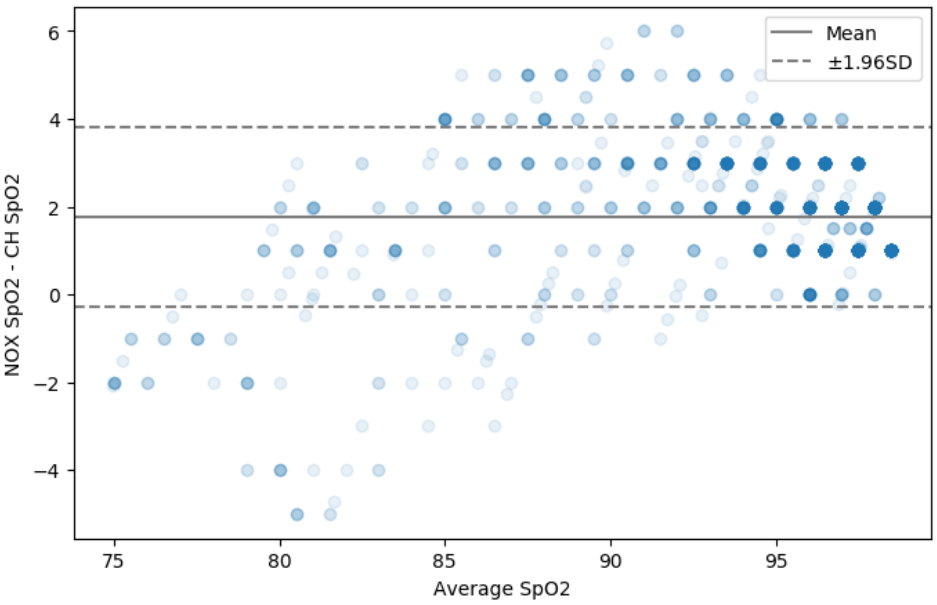

Figure 9.16: Subject 3 Bland-Altman plot

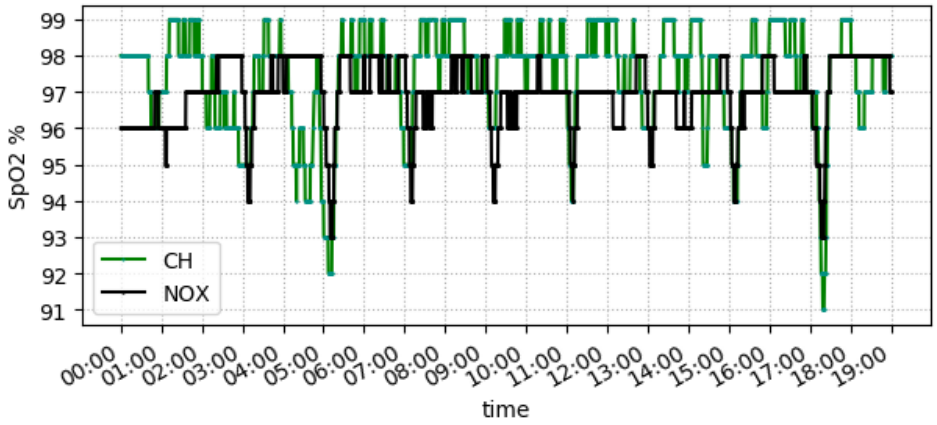

Figure 9.17: Subject 4 plot

Our example above indicates that the majority of the data samples, being the baseline saturation, have a huge impact on the calculated A_{rms}. Inspecting the mean bias and comparing the accuracy, we see that the four subjects falling within a range $\pm 0.22\%$ from zero — 2, 8, 9 and 10 — also have the highest accuracy estimations. And on the other end of the scale, the four with the highest mean bias — 1, 3, 4 and 7 — have the lowest accuracy estimations.

A second factor affecting accuracy when comparing data streams is characteristics of the desaturation. In Section 9.3 we discuss synchronization of the test oximeter record and the reference record, and we argue that because of possible design and implementation differences of the pulse oximeter we shift CH data until arriving at a best fit calculated by accuracy. Even so, the shape of the desaturations may not be equal. As mentioned, implementation of the oximeter might result in a slightly different behavior in the desaturation, or a rise in saturation. Figure 9.20 shows a zoom into the desaturation curve at the 11-minute mark. Here, both the duration and the total drop in desaturation are about the same for CH and NOX. However, while NOX's desaturation is shaped more as a line from the start to the end of the desaturation, CH desaturation is more a slope with a slow desaturation rate at start and a higher rate of fall about midway in the desaturation. Consequently, such results will worsen accuracy, even though the total duration and fall in saturation in the test oximeter data are close to the reference data.

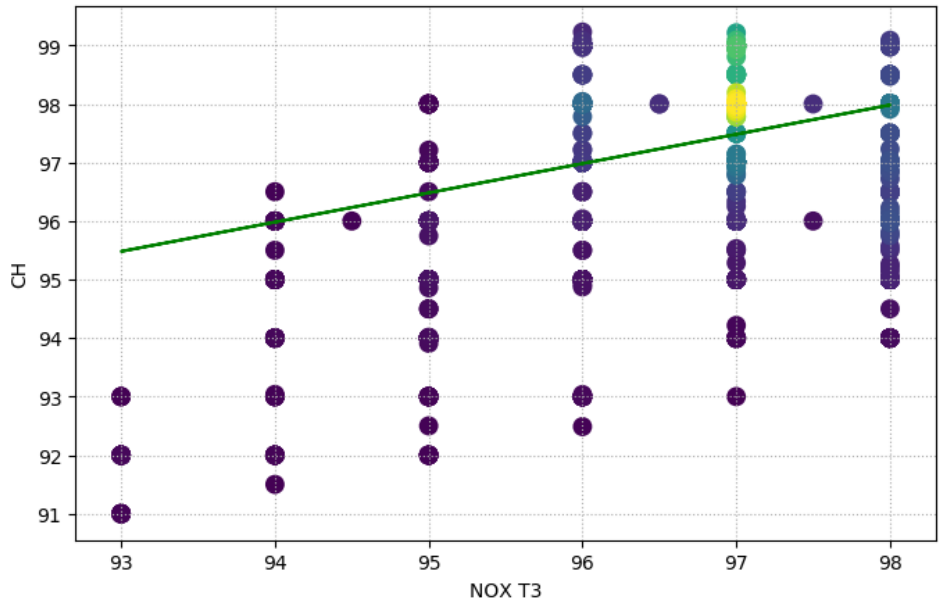

Figure 9.18: Subject 4 relation plot

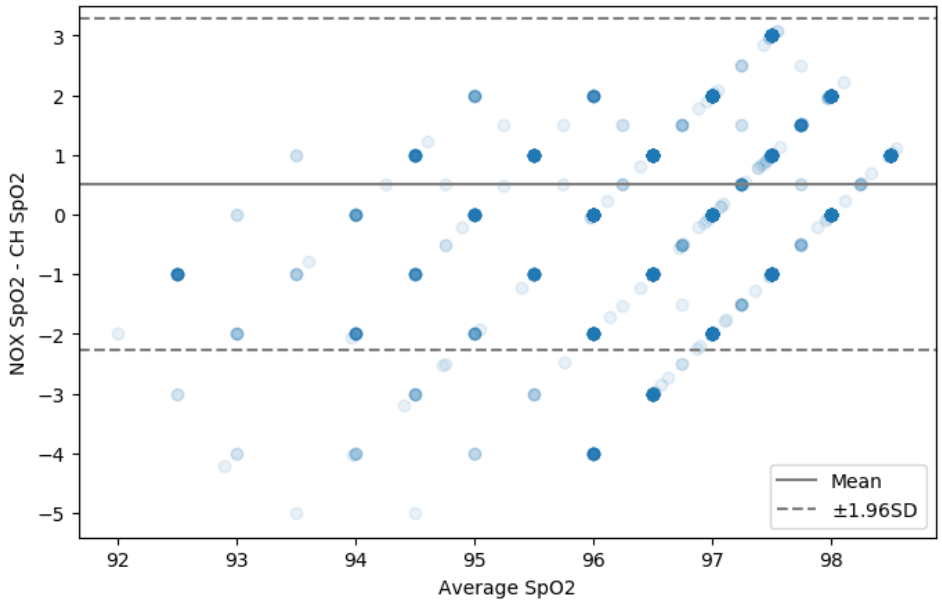

Figure 9.19: Subject 4 Bland-Altman plot

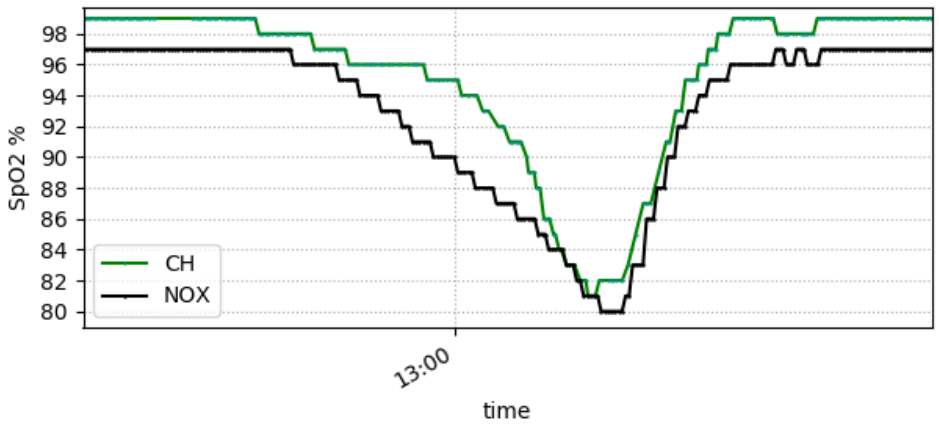

Figure 9.20: Subject 3 desaturation curve

Chapter 10

Apnea Detection

In addition to establishing the accuracy of the pulse oximeter, we also want to investigate the rate at which CH estimates desaturations where NOX also indicated them. First we present an overview of the desaturations in Section 10.1. In Section 10.2 we go into details about the desaturations and the simulation of apneas, before investigating the possible explanation of the results in Section 10.3.

10.1 Results

First we identified the desaturations $\geq$3% in NOX and CH data. Then, for each desaturation in NOX, we checked for a corresponding $\geq$3% desaturation in CH data within the same time frame. Each is classified as either true positive (TP), or false negative (FN). In total, the TP and FNs are equal to the total NOX desaturations. And conversely, if there is no corresponding desaturation from NOX where CH indicates a desaturation, a false positive (FP) is scored. Table 10.1 shows the results for each subject and in total. It also show total results without Subject 10, which is discussed below. For CH, some of the desaturation periods were split into two by a 1% rise in saturation, before continuing with desaturation. All of these were in the same desaturation period recorded by the NOX. The parentheses in the TP column are the total desaturations matching the total desaturation from CH. The other value is matching desaturation from NOX, as noted above.

Of 71 desaturations recorded by the NOX T3 oximeter, a total of 63 desaturations was also recorded by the MySignals BLE oximeter in the same time frame. CH therefore failed to record 8/71, or $\approx$11.3% of the desaturations NOX recorded. If we generalize this number, our data indicate that about one in ten desaturations is likely to be missed by the CH. As we know from the Apnea Hypoapnea Index (AHI), mild sleep apnea is scored for 5 or more apneas per our. For a person having a mean of 5 desaturation per hour over a night sleep, the results from CH will still give indications about a possible sleep disorder even without one in ten desaturations. If we look at the data from Subject 1 to 9, excluding subject ten, we see that the CH only failed to record 3 of the total 58 desaturations,

Subject no.	NOX desat.	CH desat.	TP	FP	FN
1	0	0	0	0	0
2	8	10	7	3	1
3	8	10	8(10)	0	0
4	8	11	6(7)	4	2
5	8	9	8	1	0
6	9	9	9	0	0
7	7	8	7	1	0
8	3	4	3	1	0
9	7	9	7	2	0
10	13	9	8(9)	0	5
Total	71	79	63	12	8
w/o Subject 10	58	70	58	12	3

Table 10.1:

	AASM Apneas	Desaturations
Simulated apneas	80	-
Mean duration	22sec	-
NOX total	61	71
CH total	60	79
True Positive	58 $\approx$ 95.1%	63 (55) $\approx$ 88.7%
False Positive	2 $\approx$ 3.4%	12 (12) $\approx$ 15.2%
False Negative	3 $\approx$ 4.9%	8 (3) $\approx$ 11.3%

Table 10.2: Overview of the total desaturations

which corresponds to about 5%. The reason for which it is tempting to exclude Subject 10 is discussed in Section 10.3.

10.2 Breath holding

In this section we discuss results of breath holding in general. All the subjects were able to hold their breath for longer than 10 seconds in all the simulated apneas. Neither of the test subject had training in breath hold, except Subject 3. Subject 3 was currently practising free diving, and was therefore able to both hold a controlled breathing between apneas, and hold the breath for longer than any subject.

A total account for all simulated apneas and desaturation is presented in Table 10.2. Column two and three contains desaturations event, the total number of desaturations, and the those that follow a simulated apnea(respiration reduction according to the AASM manual). First we observe the AASM Apnea column. We see that all subject was able to simulate each 8 apneas for at least 10 seconds, resulting in a total of 80 apneas. The mean duration was for 22 seconds, with a minimum of 11 to the maximum of 54 seconds. From these 80 simulated apneas, we were able to identify corresponding 61 desaturations for NOX, and 60 from

Subject no.	Apnea du	Dedu NOX	Dedu CH	De NOX	De CH	Finger
1	11s	0.0s	0.0s	0.0%	0.0%	N=m, C=i
2	17s	13.5s	17.5s	3.5%	4.0%	N=r, C=i
3	34.5s	25.0s	21.0s	12.5%	10.0%	N=i, C=m
4	22s	12.5s	18.0s	4.0%	4.0%	N=m, C=i
5	21s	16.5s	20.5s	7.6%	6.0%	N=m, C=i
6	15.5s	18.0s	15.0s	8.0%	9.0%	N=m , C=i
7	19s	21.0s	14.5s	4.5%	4.5%	N=i, C=m
8	28s	22.0s	18.0s	3.5%	5.5%	N=m, C=i
9	24.5s	29.0s	25.5s	7.5%	6.5%	N=i , C=m
10	29.5s	29.0s	19.0s	7.0%	4.0%	N=i , C=m
Total mean	22s	18.5s	16.5s	6.0%	5.5%	

De = Desaturation, du = duration, C=CH, N=NOX, i = index, m = middle, r = ring, s = seconds

Table 10.3: Results of simulated apneas, average

CH. However, CH failed to record 3 desaturations that NOX did record, resulting in 3 false negatives. 95.1% was therefore classified as true positive. CH did record 2 desaturations after simulated apneas where NOX did not record a desaturation. In these two cases, CH recorded 3% desaturation over 7 and 13 seconds. Nevertheless, 3.4% of positives were false positives.

In the Desaturation column, we take all desaturations into account. Then, NOX recorded 71 desaturations where CH recorded 79. 63, or 95.8% of the desaturation were true positives. A total of 8 were false negatives, 11.3%. In addition, the pulse oximeter from CH did record 12 more desaturations where none was indicated by NOX, 15.2%. The numbers in parenthesis in the third column are results without the results from Subject 10. As we see, most of the false negatives, 5 out of the 8, is a result of this recording. This is why it is interesting to observe the results without this recording, and we also discuss this later in this chapter.

In Table 10.3 we see the account for each individual subject, and the total mean. The second column contains the mean apnea duration. In column three and four we show the mean desaturation duration of NOX and CH. Then, in column five and six we show the mean desaturation percentage measured by the oximeters. The placement on fingers is presented in the last column. Most notable in this figure is the relation between the duration of the simulated apneas and the desaturation. As we earlier predicted, longer duration of breath hold result in higher drop in SpO_2. In general, NOX did record both longer desaturation period and a higher drop in desaturation than CH. There is no clear connection between the finger location, and the results. However, in the cases where NOX was placed on the index finger, the oximeter measured an equal or higher drop in SpO_2 values than CH did.

10.3 Classification Failures

In order to identify an explanation for the false negatives and false positives, we investigate the data from three subjects that affected these results the most. The data from subject 2, 4 and 10 is together responsible for 7 out of 12 false positives, and 5 out of 8 false negatives. The first thing we can note, is that the results from these three Subjects is within the top 4 of number of NOX and CH desaturations, with 8 and 10 for Subject 2,8 and 11 for Subject 3, and 13 and 9 for Subject 10. Each subject simulated apneas every 2nd minute, and we therefore ahead of the experiments expect a plot showing desaturations such as in Figure 9.10. There we can clearly see desaturations every 2nd minute, right after minute mark 2, 4, 6, etc. Most experiments have the same characteristics, except for Subject 2, 4 and 10.

The Subject 2 plot is displayed in Figure 10.1, Figure 9.17 contains the plot for Subject 4, and Figure 10.2 Subject 10. The plots have the same characteristics, they do all have multiple rise and fall in saturation values within a two minute time frame. In order to explain the behavior, we discuss the reason for non-periodic desaturations. Based on earlier discussion in our research, we have four possible theories.

First is that the desaturations are movement artifacts. However, all subjects were instructed, and did also in fact lay still over the time frame of the experiments. We can therefore with a high certainty exclude movement artifacts as a possible explanation of the behavior. Second, we can not with certainty know that the placement of both pulse oximeter was optimal. However, the calculated accuracy and precision of the experiments did not show to be less accurate than the rest of the experiments. We may therefore assume that placement was likely to be correct, or equally wrong. Third, it is possible for physiological factors in patients to cause arterial blood saturation, even with a steady breath. On the contrary, all subjects reported no earlier or present conditions with blood, heart or lungs. The last and most likely explanation is that the subjects did not breathe in a steady pattern, and therefore did not have a steady oxygen supply for the hemoglobins to load. We instructed them to try to breathe slowly and shallowly to lower their baseline oxygen saturation. Therefore, it may be possible that some subjects did breathe so little that the oxygen saturation fell even when they were breathing. As we know, hypoapnea is a reduction in respiration. Therefore, the recording of a night's sleep from someone diagnosed with obstructive sleep apnea is likely to contain a reduction in respiration, in addition to any apneas. The results from Subjects 2, 4 and 10 might be more representative of obstructive sleep apnea patients than the other results. Testing with additional breathing sensors would have helped us verifying our assumption above, though, the time limit of our research do not allow us to complete further experiments.

With this in mind, we discuss the consequences of possible wrongful classifications. In all three plots we discuss in this section, especially for Subjects 2 and Subject 10, we see that the desaturation of both NOX and CH is similar in character. Hence, it is likely that when analysing the data from CH, a specialist will be able to identify desaturation patterns,

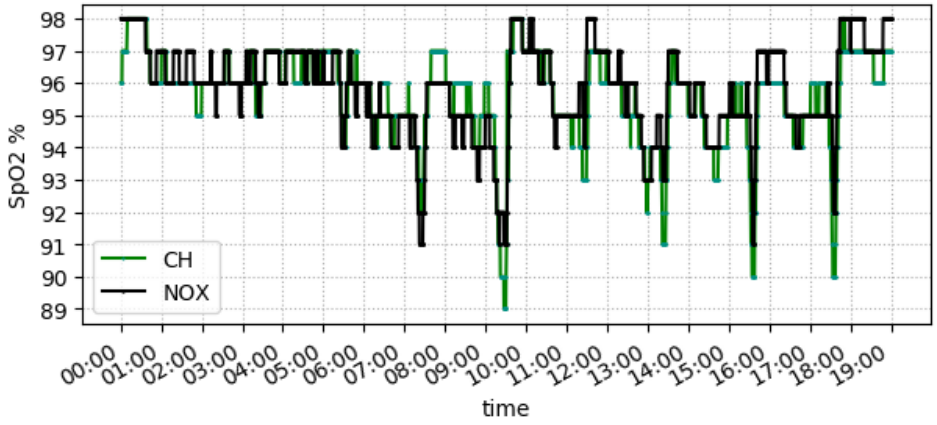

Figure 10.1: Subject 2 plot

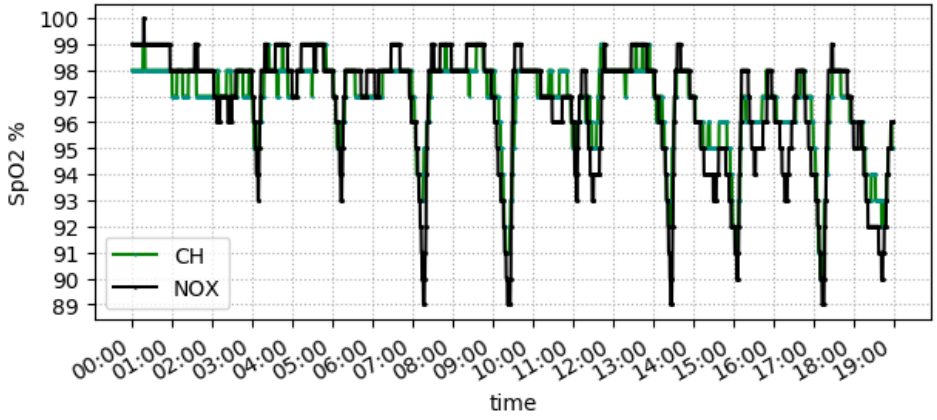

Figure 10.2: Subject 10 plot

even though our research finds multiple false negatives. That is also true for the other test subjects. Where NOX data show desaturation events, CH also has similar characteristics. On the contrary, if CH data showed few desaturation events while NOX had multiple ones, its pulse oximeter would not be suitable for sleep studies. The same conclusion is true if CH data show multiple desaturation events where NOX's show none or few. With this in mind, we can see in Table 10.1 that the number of desaturations for CH in Subjects 1–9 is the same or up to 2 more. The mean desaturations are 7.1 for NOX and 7.9 for CH, a roughly 11% difference. We inspect the results 'possible effect on the Oxygen Saturation Index by calculating the desaturation event per hour. For NOX, that calculates to a mean of 25 desaturation per hour, and 22.5 for CH.

In this thesis we classify desaturation and apnea events based on the AASM definition of hypoapnea. We therefore only count desaturation $\geq$3%. In Table 6.1 we see that for Subject 10, the mean drop of NOX' desaturations was 7%, against CH's 4%. The mean of all subjects desaturation differs only 0.5%.

Chapter 11

Discussion

In this thesis we research a benchmarking protocol to determine both the accuracy of pulse oximeters and the ability to identify desaturation events associated with sleep apneas. In this chapter, we discuss various parts of the protocol. First we address our choice of test population (Section 11.1). Next, we inspect our breathing script (Section 11.2), and last (Section 11.3) we discuss the value of our benchmarking protocols in determining the quality of pulse oximeters.

11.1 Test Population

Ideally, we wanted oxygen saturation values from all subjects in the range from 100% and close to 70%. However, in the preliminary experiments we discovered that most subjects might not even lower their saturation below 90%. In total, 5 subjects were able to lower their oxygen saturation below 90% as measured by the reference oximeter; however, only 2 below 85% and 1 below 80%.

In our work, we argued that the test population should be easy to recruit. We did this because drawing subjects from the general population would also lower the threshold for using our work to benchmark pulse oximeters. Even though we only have one person with diving experience in our test population, we assume that the achievement of reaching levels below 80% is a result of the subject's training in breath holding and respiratory control. Therefore, we can also assume that with a test population consisting of free divers only, we would have more spread in values than in our experiments. On the other hand, more training in breath holding for each subject might also improve the results. We instruct our test subjects to practice breath holding from FRC for several minutes ahead of the benchmarking. This brief period is considered more as getting to know one's limits than as training in breath holding. In Section 4.4.2, we briefly inspect what decides the breaking point of breath holding. However, more research into how to improve it, and specifically training the subject in how to do so, is considered to be an option for improving results from the test population.

11.2 Breathing Script

With the breathing script, the subject successfully achieved desaturation as measured by both the reference pulse oximeter and the test object. Even though results depend on the test subjects' training, as discussed in the last section, we now look into possible improvements to the breathing script.

The first two subjects were instructed to breathe as little as possible to lower the oxygen saturation in advance of holding their breath. However, this may have lowered their ability to hold the breath. Therefore, for the last 8 subjects, we did stress breathing as little as possible. Instead, we instructed them to breath slowly and calmly. As we see in Table 10.1, while Subjects 1 and 2 held their breath for 11 and 17 seconds on average, the rest of the test population did so for at least 19 seconds. As the results suggest, the duration of the breath holding may be more important for total desaturation than lowering the oxygen saturation in advance.

We instructed the test subjects to hold their breath for at least 10 seconds, based on the AASM classification guide's statement that the respiratory event of an apnea is $\geq$10 seconds. Even though the minimum was 10 seconds, subjects were asked to hold their breath as long as they were able. From Table 10.1 we learn that 7 in 10 subjects did so for more than 20 seconds on average, or 7 in 8 subjects with the modified benchmarking protocol. It is possible they would have pushed themselves harder if the duration of the breath holds had been 20 seconds, for example — therefore improving the fall in arterial oxygen saturation.

Our breathing script was developed to induce desaturations by having the subject simulate apneas only. A key point in our research is that the pulse oximeters are part of a health sensor platform. Therefore, a breathing script that also can be used mundanely to test other respiratory sensors, such as RIP bands, would be useful. The breathing script then would include hypoapnea and other events associated with a sleep apnea diagnosis. In this setting, the duration of the simulated events should be long enough to cause the appropriate fall in arterial oxygen saturation.

11.3 Determining Quality

The industry standard for accuracy testing uses breath gas mixes to achieve different plateaus of arterial oxygen saturation. In our work, we rely on breath holding to lower the oxygen saturations. The value frequencies in Figure 9.11 show that the distributions of values for all subjects are concentrated from 98% to 96%. Therefore, even though we get values below 90%, and below 80%, calculations of accuracy and the Bland-Altman analysis are most affected by the top 3% on the oxygen saturation scale. It is likely that statistical analysis methods exist to even out this challenge; however, the time limit of our research did not allow us to investigate them. Furthermore, the ISO of 2017 states that methods not using direct pairing with a CO-oximeter should apply the accuracy of the reference pulse oximeter to the result. For NOX and for most oximeters, this is 2%.

Then the accuracy for CH ends up being 1.34% (+2%). In addition, the literature shows that accuracy degrades towards lower levels, especially below 80% and 70%. We argue that our accuracy calculation is at best the upper bound of the accuracy of the pulse oximeter.

Instead of depending entirely on accuracy metrics and the Bland-Altman analysis, we also introduced an apnea analysis method to investigate the pulse oximeter's value in sleep studies. To investigate the relationship between our metrics, Table 11.1 combines the accuracy and precision results with those of the apnea analysis. In addition, we sort the table by descending precision and descending accuracy. In the top part of the table, we see that Subjects 10 and 4 have the most false negatives and the lowest precision as well. On the bottom part, where accuracy is sorted, no apparent pattern is discernible in the distribution of false negatives. For false positives, neither sorting condition demonstrates a relation between the two metrics. This indicates that calculating precision rather than accuracy is a better metric for our method of testing pulse oximeters. We do not see it as a part of our research to analyse this observation further. However, we can conclude that with the data we have from our experiments, accuracy calculated from the results from our benchmarking method cannot indicate the grade of success in detecting desaturation events. We cannot conclude either that Bland-Altman analysis might decide the quality. Instead, by using breath holding to gain desaturations, it is more likely that comparing desaturations from an oximeter against a reference one will give correct estimations of pulse oximeters' usability in sleep studies.

Subject no.	Accuracy	Precision	TP	FP	FN
7	1.86	1.05	7	1	0
1	1.46	1.16	0	0	0
8	0.74	1.47	3	1	0
2	0.87	1.66	7	3	1
9	0.88	1.67	7	2	0
5	1.10	1.99	8	1	0
6	1.31	2.05	9	0	0
3	2.09	2.10	8	0	0
10	1.08	2.11	8	0	5
4	1.51	2.78	6	4	2
8	0.74	1.47	3	1	0
9	0.88	1.67	7	2	0
2	0.87	1.66	7	3	1
10	1.08	2.11	8	0	5
5	1.10	1.99	8	1	0
6	1.31	2.05	9	0	0
1	1.46	1.16	0	0	0
4	1.51	2.78	6	4	2
7	1.86	1.05	7	1	0
3	2.09	2.10	8	0	0

Table 11.1: Results sorted by precision(top) and accuracy(bottom)

Part IV

Conclusion

Chapter 12

Contributions Summary

Section 12.1 summarizes of the evaluation of the quality of CH pulse oximeter. After describing our research with the BITalino pulse oximeter (Section 12.2), we summarize of our evaluation of using a non-invasive benchmarking method when determining the quality pulse oximeters(Section 12.3)

12.1 Cooking Hacks

We determined the accuracy of the Bluetooth low-energy pulse oximeter from Cooking Hacks based on a total of 32,458 paired data points from 10 subjects. We also performed a Bland-Altman analysis of each individual subject, and tallied the combined results. We find an accuracy of 1.34% compared with our reference pulse oximeter between 70 and 100%, where >90% of the values is between 95 and 100%. The mean precision is 1.71% between subjects and 2.61% for all data combined. The mean bias is 0.14 and the mean upper and lower limits of agreement are 1.96% and -1.65% (2.75% and -2.47% of the data combined).

More important, for reasons discussed in Chapter 11, we analysed CH's ability to record desaturations compared with NOX. We established that the CH pulse oximeter was able to identify 88.8% of the total 71 desaturation events recorded by NOX. Where NOX identified desaturation, CH failed to in 8 (11.3%) of the cases. In addition to the 63 true positive cases, CH recorded an additional 12 more desaturations, resulting in 15.2% false positives.

We also investigated the desaturations resulting from apnea simulations separately. Of the 61 apnea events measured by NOX, CH recorded 58 desaturation events, a total of 95.1%. From the desaturations identified in the NOX recordings, 3 were false negatives (4.9%). In addition, CH measured 2 desaturations after apneas where NOX did not measure $\geq$3% desaturation, which may or may not imitate real arterial oxygen saturation. However, as we use NOX as the reference monitor, we classify 3.4% of the CH events as false positives.

Based on the data from our experiments, we see the CH pulse oximeter as being suited for detecting desaturations in sleep studies. However, we stress that we have only tested our methods on simulated apneas.

Even though the literature indicates that the behavior of desaturation in simulated and real apneas is similar, our methods can be validated by applying the data from a sleeping person with sleep apnea disorder.

12.2 BITalino

We discussed the characteristics of the BITalino data and its possible explanations. Because of problems with data acquisition, and the unlikelihood of our being able to interpret and analyse the data, we omitted this oximeter in the benchmarking process, as discussed in Section 9.1.

12.3 Non-invasive Benchmarking

Our main objective in our research is to evaluate the use of non-invasive testing procedures for determining the quality of pulse oximeters, without the need for medical attendance. Based on background information about pulse oximetry and oxygen transport in the circulatory system, we designed our benchmarking protocol around breath holding. The results show that in recruiting test subjects with no prior training in breath holding from the general population, it is not trivial to have them hold their breath until they reach the desired saturation. Since 90% of the values are between 95 and 100%, whereas we want an even spread between 70% and 100%, we argue that our accuracy estimation is rather a confirmation or validation of the labeled accuracy of the pulse oximeter, instead of actually determining its accuracy. If we were to determine with certainty the accuracy of pulse oximeters, the test population should consist of subjects trained in breath holding. However, we see the use of gas mixtures with different oxygen levels to be a much better method of determining accuracy.

Bland-Altman analysis is often implemented for comparing two methods of measurement. Even though the reference or gold standard method is acknowledged as the best, it may not be free of errors. Therefore, Bland-Altman analysis compares the new method of measurement against the mean of the new and the gold standard method. The gold standard for determining the accuracy of pulse oximeters is to compare them against CO-oximeters. In our work, we use a second pulse oximeter as a reference. Where the reference accuracy should be added to the calculated accuracy of the test oximeter (according to ISO of 2017), Bland-Altman analysis uses only the mean as reference. Therefore, we see Bland-Altman analysis as being more useful for indicating the quality of a pulse oximeter than its metric accuracy.

Instead, we argue that using our benchmarking protocol is better suited to determining the quality of a pulse oximeter's ability to identify desaturation in a sleep study. This specific action is easy to control and does not depend on a spread of oxygen saturation levels. Instead, it is sufficient to investigate the different consequences of the classifications and to look at the behavior of the data to determine the pulse oximeter's ability to record desaturations. In conclusion, we argue that the non-invasive,

breath-holding method for testing the quality of pulse oximeters is most useful when it is related to a specific task, such as detecting sleep apnea.

Chapter 13

Open Problems

Through our research, we found problems and tasks we could not solve within the time frame of our project; most of them concerning perturbations issues and outliers in data. Early on, we specified that our task was not to perform filtering on the data acquired from the pulse oximeters. However, we did remove outliers in the data from Cooking Hacks' oximeter. When doing so, we stated that they may have resulted from our code. More time might have solved this matter or found the reason for it.

In our benchmarking process, we concluded that continuing testing the BITalino pulse oximeter only added extra tasks and time to each experiment. This was based on issues with the Collector app and the perturbations in data. First, since we had earlier work available for data acquisition for the BITalino, we had less time available for the review of the platform than we had for CH. Therefore, we relied on already developed data acquisition software. In addition to the issues with starting the data acquisition, we discovered in the experiments that the software might halt the collection of data. Therefore, we see two open problems with BITalino. First, a closer investigation of the software might reveal explanations for both the start and halts we experiences. Second, a closer investigation into the board might reveal why we experienced the disturbance pattern in data. Furthermore, we did not see it as our task to apply filters to the data from BITalino. However, because of the repeated pattern in the data, the disturbance might be removed by adding filters.

Chapter 14

Future Work

Our work shows that it is possible to determine a pulse oximeter's ability to detect desaturations associated with sleep apnea. However, the desaturations in our data are superficial, and the arterial oxygen saturation of a sleeping person with sleep apnea may not be of same character as our data. It might also depend on severity of the sleep disorder. Therefore, we see it as a task for future work to compare our results against results from sleep studies.

In order to use our benchmarking method to calculate the real accuracy of a pulse oximeter, we see it as necessary to perform a statistical analysis of the variation of results. We know from the literature that 200 samples paired with a CO-oximeter, evenly spread between 70 and 100%, are sufficient to determine the quality of a pulse oximeter. What this means for comparing continuous streams of data was not an object for discussion in our research. Instead, we focus on the feasibility of testing pulse oximeters in a non-invasive, non-medical manner.

A third object of later investigation is the relation between accuracy and precision, and the ability to detect desaturations in apnea. Even though such studies exist [24], it is possible by a testing a high number of different pulse oximeters to look into whether the labeled accuracy from the manufacturer is a correct indicator for quality in apnea detection. In Section 11.3, we discussed a possible relation between precision and the pulse oximeter's ability to measure desaturations. To investigate the direct relation, however, additional research is needed that includes more subjects and more pulse oximeters.

For our work, we wrote several scripts to handle the recorded data. Nevertheless, some tasks we chose to do manually to control the results, and also because of the time limit of our research. Even though we have implemented all of the scripts needed for analysis, a complete software application(e.g., Android app) that include two sets of sensor data and performs the analysis mentioned in this paper would ease the implementation and use of our benchmarking method. Classifying the desaturation is particularly time consuming, and may produce incorrect apnea classification.

Appendix

Appendix A

Source Code

The source code used in processing can be found at github.uio.no/kennetaf/DMMS

Appendix B

Cooking Hacks

B.1 Procedures

In Figure 5.5 earlier in this paper, we present the options for data acquisition. In our work with the Cooking Hacks MySignals and data acquisition methods, we tried different approaches that we did not find necessary to include earlier in the main work of this paper. However, these findings may be important for future work with the CH pulse oximeter We also reference different sections in the MySignals online manual (OM). Therefore, our investigations of other options that is not previously mentioned are listed in this section.

B.1.1 TFT display, Option 8

On our first attempt to extract data from the pulse oximeter, we connected the CH board to a computer with an Serial/USB connection. As a reference, and to compare against the serial example(OM 7.1.2), we tested the second and last code example for the SpO_2 BLE sensor. It sends the data stream output to the on-board TFT screen on the MySignals HW. The procedure was as follows:

1. Connect the MySignals board to a computer through serial UART
2. Upload the code from Arduino IDE to the MySignals board
3. Disconnect MySignals from the computer
4. Connect MySignals to a power source, and turn on the MySignals
5. Place the SPO2 sensor on the right index finger
6. Turn on the SPO2 sensor
7. Observe the response on the TFT screen

As a note, the example code provided misses enabling of BLE module power, and the BLE init fails. It is fixed by *bitSET bit6:1*. By trying this option, we can make the following observation. Where our first try with the example code with a serial port failed to reconnect with the SpO_2 sensor immediately, this option of not having the MySignals unit connected to a laptop with the USB cable (using only a power adapter) results in the oximeter reconnecting after about 1 second and provides us with a new

SpO_2 estimation. The oximeter might still fail to reconnect, but where the first code example reconnects after periods of several minutes, this code example usually provides us with a new estimation about every 6th-7th second.

The attempts described below follow about the same approach as in the list above; code is uploaded to the MySignals hardware when needed, and instructions from the documentation are followed.

B.1.2 MySignals App, Option 1

Next we inspect Option 1. Using only the MySignals app to connect to the pulse oximeter, it does not involve a MySignals unit. Although it would be interesting to check the result using this option for extracting data, after testing we found that the option is not available. There was no indication of this in the documentation.

B.1.3 Bluetooth, Option 2, 3 and 4

Options 2, 3 and 4 require an internal understanding of the MySignals Pulse Oximeter and its Bluetooth LE profile to communicate and acquire data through the BLE module. As noted in the background of this paper, no documentation is available about the matter. Therefore, we did not see this option as being suitable at the time because of the assumed lengthy process of reverse engineering the MySignals library.

Nevertheless, even though we were not able to receive data, we did connect to the pulse oximeter using the example code in the documentation (OM 7.4.2) for Bluetooth connection. We therefore include some notes from our experience. The passkey for connecting to the MySignals is generated by the hardware and written out on screen. After connecting once with a computing device, the bondings have to be deleted from both the connected and MySignals devices. This can be done by adding *MySignals_BLE.deleteBonding()* at the start of the MySignals program. Then all bonding is deleted from the internal memory at start up. As we noted, however, software for data acquisition must be written once connected (typically server/client or master/slave), which requires internal MySignals BLE knowledge.

B.1.4 Other BLE Devices

We also checked other Bluetooth sensors accompanying the oximeter in the MySignals HW v2 kit to see if they had the same connection problem. In the documentation we find five different BLE sensors, including the pulse oximeter. However, three of them do not have the same connection type: the Body Scale BLE, the Glucometer BLE and the Body Scale BLE. If the ON button on the sensors is pressed, they connect to the MySignals platform once to deliver its test result, and turn off afterwards. The last sensor, Body Temperature BLE, did not accompany the kit, and we was not able to inspect it.

B.1.5 WiFi connection 5

Because of Option 1, in which we do not connect the pulse oximeter to a computer, we suspected that other connection options not including a serial cable might be able to solve the reconnection issue. Having the MySignals device connected to the computer with a cable might cause static and/or perturbations on the signal, either in the MySignals board/modules or on the Bluetooth connection. Having a wireless connection is also beneficial if we want the data stored on a remote device such as a database server.

For testing this option, the following setup was used: TCP/IP server on ubuntu OS, programming MySignals as a client, and connecting through Access Point on an Android device. We experienced a couple of issues when investigating this option. First, enabling the WiFi module after the BLE module results in errors in the BLE module, so we were not able to connect the MySignals to the pulse oximeter. Next, the delay after first connection from the MySignals to the ubuntu server must be at least 5 seconds, or else no values are received from the pulse oximeter. Last, we were not able to both collect data from the pulse oximeter and send it over WiFi. However, sending data from other cabled sensors (such as the body position sensor) over WiFi did not introduce any unexpected issues.

B.2 Coding

Attempts to acquire data through options other than serial port to computer were unsuccessful and did not provide us with the required results. As explained in Section 5.4, we ended up with reverse engineering the MySignals library to write new code for the serial port options fit for our purpose.

First, from the library we have a note about the description of methods. As we want to subscribe to the data stream of the SPO2 device, we inspect the following in the example code:

```
1 //To subscribe the SpO$_{2}$ measure write ‘‘1’’ in SPO2_HANDLE
2 char attributeData[1] = { 0x01 };
3 if (MySignals_BLE.attributeWrite(connection_handle_SpO$_{2}$,
4    1.7.2.0.4 SPO2_HANDLE, attributeData, 1) &=& 0)
5 { ...
```

Then we inspect the documentation and the source code library description of attributeWrite:

```
1 // write an attribute from a remote BLE device
2 // Function: write an attribute from a remote BLE device * att
     handle in
3 decimal.
4 uint16_t attributeWrite(uint8_t connection, uint16_t
5 atthandle, uint8_t * data, uint8_t length);
```

By looking into the attributeWrite function, we can learn that we are sending the data as payload to the SPO2 device, telling it we have

subscribed to its data stream. We may therefore assume that a description such as “write an attribute to the pulse oximeter” is better suited.

The important events identified in the preliminary experiments resulted in an implementation such as the one shown below:

```
1
2  #include <MySignals.h>
3  #include <MySignals_BLE.h>
4
5  /* Write here the MAC address of BLE device to find */
6  char MAC_SPO2[14] = "00A05004182F";
7
8  /* Global variables */
9  uint8_t available_spo2 = 0;
10 uint8_t connected_spo2 = 0;
11 uint8_t connection_handle_spo2 = 0;
12 uint8_t pulse_spo2 = 0;
13 uint8_t spo2 = 0;
14 uint8_t reset = 0;
15
16 #define SPO2_HANDLE 15
17 #define BLE_DEBUG 1
18
19 /* Setup of the modules */
20 void setup()
21 {
22
23   MySignals.begin();
24
25   Serial.begin(115200);
26
27   MySignals.initSensorUART();
28   MySignals.enableSensorUART(BLE);
29   MySignals.initBodyPosition();
30
31   //Enable BLE module power -> bit6: 1
32   bitSet(MySignals.expanderState, EXP_BLE_POWER);
33   MySignals.expanderWrite(MySignals.expanderState);
34
35   //Enable BLE UART flow control -> bit5: 0
36   bitClear(MySignals.expanderState, EXP_BLE_FLOW_CONTROL);
37   MySignals.expanderWrite(MySignals.expanderState);
38
39   /* Why disable and enable the unit again? */
40   //Disable BLE module power -> bit6: 0
41   bitClear(MySignals.expanderState, EXP_BLE_POWER);
42   MySignals.expanderWrite(MySignals.expanderState);
43
44   delay(500);
45
46   //Enable BLE module power -> bit6: 1
47   bitSet(MySignals.expanderState, EXP_BLE_POWER);
48   MySignals.expanderWrite(MySignals.expanderState);
49
```

```
  delay(1000);

  MySignals_BLE.initialize_BLE_values();
  if (MySignals_BLE.initModule() == 1)
  {
    if (MySignals_BLE.sayHello() == 1)
    {
      MySignals.println("BLE init ok");
    }
    else
    {
      MySignals.println("BLE init fail");
      while (1){};
    }
  }else{
    MySignals.println("BLE init fail");
    while (1)
     {};
}}

void loop(){
/* Connect to the pulse oximeter */
  available_spo2 = MySignals_BLE.scanDevice(MAC_SPO2, 1000,
      TX_POWER_MAX);

  if (available_spo2 == 1) {

   if (MySignals_BLE.connectDirect(MAC_SPO2) == 1) {
     connected_spo2 = 1;
     connection_handle_spo2 = MySignals_BLE.connection_handle;

/* Subscribe to stream */
     char attributeData[1] = {0x01};
     if (MySignals_BLE.attributeWrite(connection_handle_spo2,
       SPO2_HANDLE, attributeData, 1) == 0){
        unsigned long previous = millis();
        uint8_t noEvents = 0;

/* Continuous data reading */
      do   {
        uint8_t eventRet = MySignals_BLE.waitEvent(1000);
        if (eventRet == BLE_EVENT_ATTCLIENT_ATTRIBUTE_VALUE) {

/* Read value from Serial, and output as csv */
         spo2 = MySignals_BLE.event[13];
         spo2 &= 0b01111111;
         MySignals.disableMuxUART();
         Serial.print(millis());
         Serial.print(F(","));
         Serial.print(spo2);
         Serial.print(F(","));
         Serial.print(pulse_spo2);
         MySignals.enableMuxUART();

```

```
103         noEvents = 0;
104 /* Handles disconnection and continuous unwanted events */
105        } else {
106         if(eventRet == BLE_EVENT_CONNECTION_DISCONNECTED)
107         break;
108         noEvents++;
109         if(noEvents > 5)
110         break;
111        }
112      }
113      while(1);
114
115 /* Error handling */
116
117     } else {
118      MySignals.println("Error subscribing");
119     }
120   }else{
121     connected_spo2 = 0;
122
123     MySignals.println("Not Connected");
124   }
125  } else if (available_spo2 == 0) {
126   //Do nothing
127  }else{
128
129   MySignals_BLE.hardwareReset();
130   MySignals_BLE.initialize_BLE_values();
131   delay(100);
132 }}
```

B.3 Data Quality

In this section we discuss an odd experience that results in outliers thought our research.

In this section we discuss an odd experience that results in outliers throughout our research. Depending on delays and sample rates, the values we get from the pulse oximeter vary. Inserting a delay $\geq$5000ms (0.2Hz) between each attempt to receive data from the pulse oximeter always results in values within the probable SpO_2 range of the tested person. A faster sample rate results in values displayed in Figure 5.8. However, when we remove all delays and disable all modules other than the BLE module, we achieve a sample rate between 8Hz and 10Hz. Then we also experience the disappearance of outlier values. As a result, if sample rate of where the BLE module of the MySignals ask for values (loop through the code) from the oximeter is <0.2Hz, and >8Hz, no outliers are present in the data. In our benchmarking we use the module for the accelerometer, which slows down the MCU from 3Hz to 6Hz. Therefore, outliers are present in the data in all of our experiments.

We did not find the cause for this behavior. However, we can speculate

that it has something to do with the disconnection and reconnection implemented in the pulse oximeter example code, which is meant to avoid overflow values or redundant data in the serial buffer.

B.3.1 Sampling Rate

As mentioned in the section above, by enabling body position module on MySignals we are slowing down the sampling rate/data acquisition from the pulse oximeter, and activating modules is therefore a direct reason for outliers in data. In order to try to understand this behavior, we can inspect the following quote from the MySignals documentation:
"Q: Can I use all the sensors at the same time?
A: In the case of MySignals SW, yes you can. In the case of MySignals HW the Arduino processor is limited, so you can not manage all the sensors, wireless communication and others features at the same time. You should select a correct combination of the options available. Check the documentation for that."

Even though it is possible to enable more than one module at a time, we can assume that the limited processor ability does not allow it if our goal is to receive only data from the different modules, not including bad values or outliers.

Appendix C

Benchmarking Protocol Documents

C.1 Benchmarking Protocol

Benchmarking instructions

1. Go through prearrangements, and fill out requested documents
   (a) Health Statement: Answering yes to any questions about health condition automatically excludes the test subject from the test population.
   (b) Test subject may add additional information about the physical condition in the relevant section in the Event Document.
2. Go through description section with test subject.
3. Carry though a brief testing period, where the test subject practice both breath-holding and lowering the baseline oxygen saturation. Observe the monitor.
4. Begin the benchmarking. Remember to register events in document

Prearrangements:

- Translate "Test Subject Instructions" document to native language of the test subject if needed/possible. The degree of the subjects understanding of the instruction may affect the test results.
- Equipment:
  - Room with couch, bed or similar for testing
  - One reference pulse oximeter
  - One or more pulse oximeters as object for testing
  - Read usa instructions for each oximeter. Note information about placement.
  - Devices with installed methods for data acquisition, such as computer or smart phones
  - Stop-watch
  - Pencil and document for registering events (Document C.4)
- Remove nail polish if present
- Both test subject and test manager read the instructions. Make sure test subject knows what is meant by each event.
- Register:
  - Fill in "Statement of Privacy and Personal Condition" document.
  - Record location and altitude.
  - Roughly estimate temperature in room.
  - Register oximeter's finger locations or picture of setup. (It is important that the test oximeters to be located at the same hand as on the reference oximeters.)

Descriptions

Description of the breathing procedure:

We are testing pulse oximeters and their ability to measure oxygen changes in the blood. Basically, test subjects are holding his or her breath for a specified time period, with the intention to lower the oxygen saturation. The rate of the fall is determined by initial oxygen level, at the time breath holding starts.

To achieve the best result, you should therefore try to lower the oxygen saturation upfront of breath holding. You are not expected to be able to do this, but it would have a positive influence on the test results if possible. While holding your breath, it is important that you should not push yourself to an extent where you feel dizzy or unwell during the test period. But holding your breath after exhalation is more unpleasant then regular breath being held.

Note that the benchmarking process can be terminated at any stage, at your own will. Even though there is no indication that holding one's breath for short periods of time inflicts physiological damage, some persons might feel dizziness for several minutes afterwards. Studies show that without training it is very rare for persons to be able to hold their breath until they faint. Studies also indicate that the breaking point of your breath holding is mainly determined by your own strength of will.

Explain concepts

- **Calm breathing**, imagine that your breathing is as calm as when you relax or sleep. Try shallow or slow breathing. For best results, and as a guideline, very shallow breathing might make you feel uncomfortable and breathless.
- **Hold breath**, from normal expiration (normal breath out without forcing air out of lungs). It might also be important not to hold your breath so long that you "burst". You should be able to control your breathing within moments after breath holding.
- **Contain the need for deep breaths**, if possible desist from deep breaths, as they will increase oxygen saturation back to normal.

Test procedures

Register in process:

- Register events on document, with time elapsed timestamp.
  - **Start** of breath held
  - **Stop** of breath held
  - **Position** if changed
  - **Movement**
  - **Other events** such as external disturbance or interference

Breathing Script

1. *Calm breathing*, 3 minutes
2. *Hold breath* on expiration at least 10 seconds
3. *Calm breathing*, 2 minutes
4. Repeat points 2 and 3, a total of 8 times.
5. End
    - Remember: 2 minutes calm breathing at the end.

C.2 Test Subject Instructions

Description of the breathing procedure:
We are testing pulse oximeters and their ability to measure oxygen changes in the blood. Basically, test subjects are holding his or her breath for a specified time period, with the intention to lower the oxygen saturation. The rate of the fall in oxygen level is determined by the initial level at the beginning of the breath held.

To achieve the best results, you should therefore try to lower the oxygen saturation of breath held upfront. You are not expected to be able to do this, but if possible, it would have a positive influence on the test results. While holding your breath, it is important that you do not push yourself to an extent where you feel dizzy or unwell during the test period. But holding your breath after exhalation is more unpleasant than regular breath holding.

Note that the benchmarking process might be terminated at any stage, at your own will. Even though there is no indication that your holding breath for short periods of time inflicts physiological damage, some persons might feel dizziness for several minutes afterwards. Studies shows that without training it is rare for persons to be able to hold their breath until they faint. Studies also indicate that the breaking point of your breath holding is mainly determined by your own strength of will.

Concepts

- **Calm breathing**, imagine that your breathing is as calm as when you relax or sleep. Try shallow or slow breathing. For best results, and as a guideline, very shallow breathing might make you feel uncomfortable and breathless.
- **Hold breath**, from normal expiration (normal breath out without forcing air out of lungs). It might also be important not to hold your breath so long that you "burst". You should be able to control your breathing within moments after breath holding.
- **Contain the need for deep breaths**, if possible desist from deep breaths, as they will increase oxygen saturation back to normal.

Breathing Script

1. *Calm breathing*, 3 minutes
2. *Hold breath* on expiration at least 10 seconds
3. *Calm breathing*, 2 minutes
4. Repeat points 2 and 3, a total of 8 times.
5. End
    - Remember: 2 minutes calm breathing at the end.

C.3 Physical Health Statement

The purpose of this document is to avoid any health risks and to ensure that the inclusion and exclusion conditions are followed.

Subject ID: _____
Age: _____
Sex: _____

Answer yes or no to the following questions(answer yes if in doubt):
Pregnant? _____
Smoker? _____
Any lung disease or respiratory problems? _____
Illness causing low level of oxygen in the blood? _____
Any heart condition? _____
Any brain issues? _____
Other health conditions or illnesses at the moment? _____

Name: _____

Date and Signature:

_____

C.4 Event Document

exp id = [ ], subject id = [ ], Altitude ≅ _____

Oximeters finger
location:_____

Ax = Apnea x, Ix = Incident x,

Hendelse	Start m:s	End m:s	Note
Start			
A1 - 3 min			
A2 - 5 min			
A3 - 7 min			
A4 - 9 min			
A5 - 11 min			
A6 - 13 min			
A7 - 15 min			
A8 - 17 min			
end 19 min			

Bibliography

[1] American Society of Anesthesiologists. *ASA Physical Status Classification System*. 2018. URL: https://www.asahq.org/resources/clinical-information/asa-physical-status-classification-system (visited on 19/06/2018).

[2] American Sleep Apnea Association. *Central Sleep Apnea*. 2018. URL: https://www.sleepapnea.org/learn/sleep-apnea/central-sleep-apnea/ (visited on 08/06/2018).

[3] Steven J Barker and Nitin K Shah. 'The effects of motion on the performance of pulse oximeters in volunteers (revised publication)'. In: *Anesthesiology: The Journal of the American Society of Anesthesiologists* 86.1 (1997), pp. 101–108.

[4] Richard B Berry et al. 'Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events: deliberations of the sleep apnea definitions task force of the American Academy of Sleep Medicine'. In: *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine* 8.5 (2012), p. 597.

[5] National Center for Biotechnology Information. *Systemic Circulation (Blood Circulation)*. 2018. URL: https://www.ncbi.nlm.nih.gov/pubmedhealth/PMHT0023062/ (visited on 19/06/2018).

[6] BITalino. *BITalino*. 2018. URL: http://bitalino.com/en/ (visited on 18/02/2018).

[7] J Martin Bland and Douglas G Altman. 'Measuring agreement in method comparison studies'. In: *Statistical methods in medical research* 8.2 (1999), pp. 135–160.

[8] J Martin Bland and DouglasG Altman. 'Statistical methods for assessing agreement between two methods of clinical measurement'. In: *The lancet* 327.8476 (1986), pp. 307–310.

[9] Mallory M Chan, Michael M Chan and Edward D Chan. 'What is the effect of fingernail polish on pulse oximetry?' In: *Chest* 123.6 (2003), pp. 2163–2164.

[10] SJ Choi et al. 'Comparison of desaturation and resaturation response times between transmission and reflectance pulse oximeters'. In: *Acta Anaesthesiologica Scandinavica* 54.2 (2010), pp. 212–217.

[11] Charles J Coté et al. 'The effect of nail polish on pulse oximetry.' In: *Anesthesia and analgesia* 67.7 (1988), pp. 683–686.

[12] Robert O Crapo et al. 'Arterial blood gas reference values for sea level and an altitude of 1,400 meters'. In: *American Journal of Respiratory and Critical Care Medicine* 160.5 (1999), pp. 1525–1531.

[13] Oxford Living Dictionary. *Invasive*. URL: https://en.oxforddictionaries.com/.

[14] Ramon Farré et al. 'Importance of the pulse oximeter averaging time when measuring oxygen desaturation in sleep apnea'. In: *Sleep* 21.4 (1998), pp. 386–390.

[15] John R Feiner, John W Severinghaus and Philip E Bickler. 'Dark skin decreases the accuracy of pulse oximeters at low oxygen saturation: the effects of oximeter probe type and gender'. In: *Anesthesia & Analgesia* 105.6 (2007), S18–S23.

[16] Robert R Fluck et al. 'Does ambient light affect the accuracy of pulse oximetry?' In: *Respiratory Care* 48.7 (2003), pp. 677–680.

[17] US Food, Drug Administration et al. 'Pulse oximeters—premarket notification submissions [510 (k) s]: guidance for industry and food and drug administration staff'. In: *US Department of Health and Human Services* (2013).

[18] Svein-Petter Gjøby. 'Extensible data acquisition tool for Android'. MA thesis. 2016.

[19] Cooking Hacks. *MySignals HW v2 - eHealth and Medical IoT Development Platform for Arduino*. 2018. URL: https://www.cooking-hacks.com/mysignals-hw-ehealth-medical-biometric-iot-platform-arduino-tutorial/ (visited on 22/05/2018).

[20] Harvard. *Apnea Hypopnea Index (AHI)*. 2011. URL: http://healthysleep.med.harvard.edu/sleep-apnea/diagnosing-osa/understanding-results (visited on 12/02/2018).

[21] helsenorge.no. *Snorking og søvnapné*. 2018. URL: https://helsenorge.no/sykdom/sovnproblemer/snorking-og-sovnapne (visited on 22/05/2018).

[22] *Medical electrical equipment – Part 2-61: Particular requirements for basic safety and essential performance of pulse oximeter equipment*. Standard. Geneva, CH: International Organization for Standardization, Apr. 2011.

[23] *Medical electrical equipment – Part 2-61: Particular requirements for basic safety and essential performance of pulse oximeter equipment*. Standard. Geneva, CH: International Organization for Standardization, des 2017.

[24] Michael S Lipnick et al. 'The Accuracy of 6 Inexpensive Pulse Oximeters Not Cleared by the Food and Drug Administration: The Possible Global Public Health Implications'. In: *Anesthesia & Analgesia* 123.2 (2016), pp. 338–345.

[25] Fredrik Løberg. 'Measuring the Signal Quality of Respiratory Effort Sensors for Sleep Apnea Monitoring'. MA thesis. 2018.

[26] Mark R Macknet et al. 'The accuracy of noninvasive and continuous total hemoglobin measurement by pulse CO-Oximetry in human subjects undergoing hemodilution'. In: *Anesthesia & Analgesia* 111.6 (2010), pp. 1424–1426.

[27] Nox Medical. *NOX T3*. 2018. URL: http://www.noxmedical.com/products/nox-t3-sleep-monitor (visited on 18/02/2018).

[28] Center for Medicare & Medicaid Services. *Type I, Type II, Type III Sleep Monitors, CMS AASM Guidelines*. 2018. URL: https://clevemed.com/cms-aasm-guidelines-for-sleep-monitors-type-i-type-ii-type-iii/ (visited on 08/06/2018).

[29] QJW Milner and GR Mathews. 'An assessment of the accuracy of pulse oximeters'. In: *Anaesthesia* 67.4 (2012), pp. 396–401.

[30] Yvonne Ng et al. 'Oxygen desaturation index differs significantly between types of sleep software'. In: *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine* 13.4 (2017), p. 599.

[31] ORA. *Polysomnography (PSG) – AKA Sleep Study or Sleep Test*. 2018. URL: http://www.chicagosleepapneasnoring.com/polysomnography/ (visited on 08/06/2018).

[32] MJ Parkes. 'Breath-holding and its breakpoint'. In: *Experimental physiology* 91.1 (2006), pp. 1–15.

[33] Paul E Peppard et al. 'Increased prevalence of sleep-disordered breathing in adults'. In: *American journal of epidemiology* 177.9 (2013), pp. 1006–1014.

[34] Michael T Petterson, Valerie L Begnoche and John M Graybeal. 'The effect of motion on pulse oximetry and its clinical significance'. In: *Anesthesia & Analgesia* 105.6 (2007), S78–S84.

[35] N Phattraprayoon et al. 'Accuracy of pulse oximeter readings from probe placement on newborn wrist and ankle'. In: *Journal of Perinatology* 32.4 (2012), pp. 276–280.

[36] Jeffrey J Pretto et al. 'Clinical use of pulse oximetry: official guidelines from the Thoracic Society of Australia and New Zealand'. In: *Respirology* 19.1 (2014), pp. 38–46.

[37] Naresh M Punjabi. 'The epidemiology of adult obstructive sleep apnea'. In: *Proceedings of the American Thoracic Society* 5.2 (2008), pp. 136–143.

[38] Asher Qureshi, Robert D Ballard and Harold S Nelson. 'Obstructive sleep apnea'. In: *Journal of Allergy and Clinical Immunology* 112.4 (2003), pp. 643–651.

[39] Scott A Sasse et al. 'Arterial blood gas changes during breath-holding from functional residual capacity'. In: *Chest* 110.4 (1996), pp. 958–964.

[40] John W Severinghaus and Poul B Astrup. 'History of blood gas analysis. VI. Oximetry'. In: *Journal of clinical monitoring* 2.4 (1986), pp. 270–288.

[41] John W Severinghaus and Karen H Naifeh. 'Accuracy of response of six pulse oximeters to profound hypoxia.' In: *Anesthesiology* 67.4 (1987), pp. 551–558.

[42] Hugo Plácido da Silva et al. 'Off-the-person electrocardiography: performance assessment and clinical correlation'. In: *Health and Technology* 4.4 (2015), pp. 309–318.

[43] Thomas Plagemann Stein Kristiansen Vera Goebel and Karl Øyri. *Event Modeling and Processing to Simplify Real-Time Analysis of Physiological Signals*. Tech. rep. University of Oslo, Norway, Oslo University Hospital, Norway, 2017.

[44] Kingman P Strohl and Murray D Altose. 'Oxygen saturation during breath-holding and during apneas in sleep'. In: *Chest* 85.2 (1984), pp. 181–186.

[45] A Thornton, W Ruehland, B Duce et al. 'ASTA/ASA commentary on AASM manual for the scoring of sleep and associated events'. In: *Australasian Sleep Technologists Association, Australasian Sleep Association, Version1* 7 (2010).

[46] Kevin Townsend. *Introduction to Bluetooth Low Energy*. 2018. URL: https://learn.adafruit.com/introduction-to-bluetooth-low-energy (visited on 29/06/2018).

[47] Nerraj Suri Vinay Sashidananda Abdelmajid Khelil. 'Quality of Information in Wireless Sensor Networks: A Servey'. In: *ICIQ (to appear)* (2010).

[48] David B Wax, Philip Rubin and Steven Neustein. 'A comparison of transmittance and reflectance pulse oximetry during vascular surgery'. In: *Anesthesia & Analgesia* 109.6 (2009), pp. 1847–1849.

[49] Stephanie Wilson et al. 'Comparing finger and forehead sensors to measure oxygen saturation in people with chronic obstructive pulmonary disease'. In: *Respirology* 18.7 (2013), pp. 1143–1147.

[50] Michael W Wukitsch et al. 'Pulse oximetry: analysis of theory, technology, and practice'. In: *Journal of clinical monitoring* 4.4 (1988), pp. 290–301.

[51] Pu Zhang, Baoyu Hong and Jing Chen. 'Design of Pulse Oximeter Simulator Calibration Equipment'. In: *World Congress on Medical Physics and Biomedical Engineering May 26-31, 2012, Beijing, China*. Springer. 2013, pp. 1533–1536.