Detection of Infectious Respiratory Diseases Using Wearable Devices

University of Turku
Department of Computing
Master of Science (Tech) Thesis
Biomedical Engineering and Health Technology
August 2024
Inka Mustajoki

Supervisors:
Matti Kaisti
Katri Karhinoja

UNIVERSITY OF TURKU
Department of Computing

Inka Mustajoki: Detection of Infectious Respiratory Diseases Using Wearable Devices

Master of Science (Tech) Thesis, 53 p., 2 app. p.
Biomedical Engineering and Health Technology
August 2024

Wearable devices are consumer worn devices that can be used to measure the body's physiological responses. Wearable devices are often non-invasive, for example worn on the wrist or finger, and can be used to measure heart rate, steps, respiratory rate, and several other parameters.

Respiratory tract infections are common diseases, and their severity can vary from mild, like runny nose, to life-threatening pneumonia. In particular, the coronavirus that spread as a pandemic in 2020, caused plenty of deaths and lots of grief to society. The purpose of this thesis is to find out, if the data from wearable devices can be used to detect a respiratory infection even before its symptom onset.

Three different anomaly detectors were applied to three publicly available datasets. Resting heart rate, step count, heart rate variability and temperature were used as parameters.

As a result, the impact of infection on the resting heart rate varied considerably between individuals. In general, sensitivities of the detectors were low, around 20%. In addition, different combinations of model parameters were tested to optimize the results, and with possible overfitting, sensitivity increased up to 50%.

The possible explanation for the low sensitivity of the detectors is that physiological alterations caused by infections are similar with other physiological reactions, for example stress. With more complex models and using several measurable parameters at the same time, it might be possible to identify respiratory infections more reliably with wearable devices.

Keywords: Wearables, Respiratory Infection, COVID-19, Anomaly Detection

TURUN YLIOPISTO
Tietotekniikan laitos

Inka Mustajoki: Detection of Infectious Respiratory Diseases Using Wearable Devices

Diplomityö, 53 s., 2 liites.
Lääketieteellinen tekniikka ja terveysteknologia
Elokuu 2024

Puettavat laitteet ovat laitteita, joita tavallinen kuluttaja voi käyttää fysiologisten vasteiden mittaamiseen. Usein puettavat laitteet ovat ei-invasiivisia, esimerkiksi ranteessa tai sormessa pidettäviä, ja niillä voi mitata sykettä, askeleita, hengitystaajuutta sekä useita muita muuttujia.
Hengitystieinfektiot ovat yleisiä sairauksia, ja niiden vakavuus voi vaihdella nuhasta hengenvaaralliseen keuhkokuumeeseen. Erityisesti vuonna 2020 pandemiaksi levinnyt koronavirus aiheutti paljon kuolemantapauksia sekä yhteiskunnallisia haasteita. Tämän diplomityön tarkoituksena on selvittää, voiko puettavista laitteista saadulla datalla tunnistaa hengitystieinfektion jopa ennen sen oireiden alkua.
Työssä käytettiin kolmea julkisesti saatavilla ollutta tietoaineistoa (*engl. dataset*) ja tunnistukseen käytettiin kolmea erilaista algoritmia. Parametreinä käytettiin leposykettä, askeleiden määrää, sykevälivaihtelua sekä ihon lämpötilaa. Havaittiin, että ihmisten välillä on suuria eroja siinä, tapahtuuko leposykkeessä muutosta infektion aikana. Yleisesti kaikkien algoritmien herkkyys oli matala, noin 20 %. Herkkyyden nostamiseksi algoritmeille annettiin erilaisia parametrien yhdistelmiä, millä herkkyys saatiin mahdollisesti ylisovittaen nousemaan jopa 50 %:iin.
Algoritmien heikkoa herkkyyttä selittää se, että infektioiden aiheuttamat puettavilla laitteilla mitattavat fysiologiset muutokset voivat sekoittua muihin fysiologisiin muutostiloihin, esimerkiksi stressiin. Monimutkaisemmilla tunnistimilla ja käyttäen useampaa mitattavaa parametria saman aikaisesti, puettavilla laitteilla olisi mahdollista tunnistaa hengitystieinfektioita luotettavammin.

Asiasanat: puettava teknologia, hengitystieinfektio, COVID-19, poikkeavuuksien tunnistus

Contents

List of acronyms

AUC Area Under the Curve

bpm Beats per Minute

COVID-19 Coronavirus disease

ECG Electrocardiography

FPR False Positive Rate

FSM Finite State Machine

HR Heart Rate

HRV Heart Rate Variability

LAAD Long Short-Term Memory Networks-based Autoencoder for Anomaly Detection

LSTM Long-Short Term Memory

PCR Polymerase Chain Reaction

PPG Photoplethysmography

RHRAD Resting Heart Rate Anomaly Detection

RHR Resting Heart Rate

ROC Receiver Operating Characteristic

RR Respiratory Rate

SCG Seismocardiography

SpO_2 Peripheral oxygen saturation

TPR True Positive Rate

1 Introduction

1.1 Wearables

Wearable devices (later wearables) can be defined as non-invasive, relatively cheap, customizable devices designed to monitor users' physiological parameters. Although typically focusing on non-invasive technologies, some wearables are minimally invasive like glucose sensors. For the purpose of this thesis, only non-invasive wearables are considered.

Wearables come in various forms such as wristbands, rings, or even piece of clothing offering users a convenient way to track parameters affecting their health (See Figure 1.1). Wearables can be used to several purposes: monitoring basic physiological parameters such as heart rate (HR), continuous tracking of sporadic health events such as arrhythmias or sleep apnea, detecting activity levels and habits or even pre-symptomatic infections, or more advanced prediction like mortality or pulmonary disease getting worse [1].

Common parameters measured by wearables include HR, step count, energy consumption and sleeping patterns. Additionally, more detailed metrics like heart rate variability (HRV), respiratory rate (RR), dermal temperature, stress levels, oxygen saturation (SpO_2) and blood pressure [2] is measurable using wearables.

As wearables become more advanced and accessible, they are increasingly integrated into daily life, empowering individuals to actively manage and track their

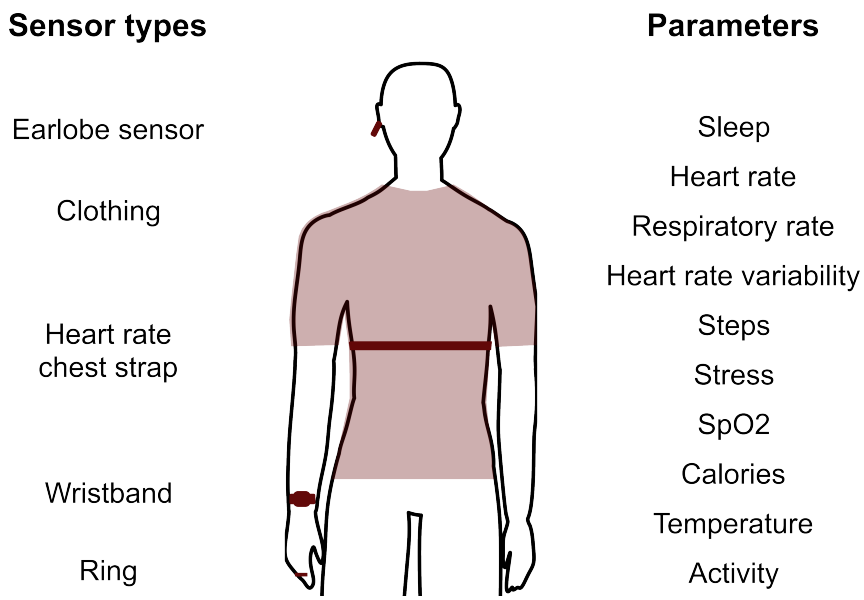

Figure 1.1: **Wearable sensor types and measurable parameters.**

health and possibly prevent complications of diseases. Many athletes (or individuals interested in fitness) use wearables for recovery and exercise planning to optimize the results [3]. Besides of individual's own will to use wearable and track life, wearables are also used in healthcare. For example, semi long-term cardio monitoring can be used in the healthcare sector for irregular arrhythmia detection [4] or sleep disorder definition.

Wearables mainly employ well known techniques like photopletysmography (PPG), electrocardiography (ECG), accelerometers and gyroscopes. PPG uses light to monitor changes in volume of blood in semi-superficial arteries. The light is transmitted to the skin and the intensity of reflected light describes blood volume. From PPG signal, HR and SpO_2 can be gathered relatively accurately. In addition, RR, HRV are commonly used parameters that can be extracted from PPG signal. [5]. Blood pressure can also be calculated from PPG signal utilizing for example pulse transit time or pulse wave velocity. However, these methods are not accurate enough for medical purposes, so measuring blood pressure using PPG is not commonly standardized method [6–8]. ECG measures the electrical activity of the heart. Traditionally ECG

measurement has 12-leads (total of 10 electrodes) or 6-leads (with 3 limb electrodes) which are placed to chest and limbs but especially wrist-worn wearables (e.g. smart watches) utilizes 1-lead ECG with two limb electrodes. Heart rhythm associated parameters like HR and HRV, and respiratory rate can be extracted from ECG signal. Accelerometers and gyroscopes are used to step count calculation, and along with other signals, to energy consumption calculation and sleep tracking. Besides standard fitness trackers, there are wearable devices for cardio health monitoring. Those devices utilize ECG and seismocardiography (SCG), which is based on measuring the movement of the chest caused by the heart beating. [9]. SCG provides information about heart function and may help with cardiac event detection.

Although technologies of wearables have developed recently and wearables have been generalized, many challenges still remain in wearables. Firstly, even though wearables are cheap compared to hospital monitors, the newest and the best smartwatches and smartrings can cost hundreds of euros [10]. Secondly, data quality is a considerable problem among wearables. Accuracy of heart rate signal from PPG can decrease up to 30% during an activity, like walking, compared to rest [11], and accuracy improves when the device knows the current activity of the user [12]. In addition to inaccuracy caused by movement, data quality and accuracy are affected by tightness of the sensor and possible moisture, for example sweat, between skin and the sensor [13].

Diverse scale of brands and measurement sites leads to diverse data processing methods and algorithms, so data from two wearables can differ considerably [1] which makes future research using commercial wearable more difficult. Development of new, higher quality algorithms used in wearables is slow because the process requires large amounts of labeled data.

1.2 Infections

An infection denotes a state where a pathogen, such as a virus or bacteria, infiltrates to a human body, replicates, and triggers an immune response. Infections can cause different diseases or states like respiratory infections, such as common cold or influenza. [14]. They can also induce more severe conditions like acquired immune deficiency syndrome (AIDS) caused by human immunodeficiency virus (HIV) and can even trigger certain autoimmune disorders [15]. The most common viruses causing respiratory infections in humans are rhinoviruses, coronaviruses, and influenza viruses [14, 16]. In addition, some bacteria, for example Streptococcus pneumoniae and Mycoplasma pneumoniae, can cause respiratory infections but viruses are more common cause of respiratory infections than bacteria [17].

On average an adult goes through 2-5 respiratory infections annually [14]. Symptoms of respiratory infection can vary from mild sneezing and sore throat to high fever and severe cough [16]. In general, however, milder symptoms are more prevalent. Respiratory viruses that cause infections can spread via aerosols suspended in the air, larger aerosol particles emitted directly from an infected person (e.g., through sneezing or coughing), or through contact with contaminated surfaces such as the hands of infected individuals or commonly touched objects like handrails or doorknobs [16]. Using a face mask can reduce the spread of aerosols. In addition, good hand hygiene can prevent the transmission of viruses both from an infected person to surfaces and from surfaces to unaffected individuals. [18, 19].

The emergence of SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) in 2019 precipitated in coronavirus disease (COVID-19) resulting to pandemic disease. COVID-19 has caused over 7 million deaths worldwide [20]. The most common symptoms of COVID-19 are cough, headache, fever, and fatigue, but also tiredness and stomach problems are possible symptoms [21]. In the beginning of the COVID-19 pandemic vaccine development started, and relatively quickly a few

different vaccines were developed. COVID-19 vaccine prevents from viral spreading and protects individuals from severe disease and deaths [22, 23]. COVID-19 median incubation period is approximately 5 days, with 97.5% of infected individuals experiencing symptoms within 11.5 days (if symptoms developed), indicating potential infectivity before symptom onset [24].

Influenza, commonly known as flu or seasonal flu, is a viral disease caused by influenza viruses. The most prevalent form of influenzas is seasonal influenza, which varies from year to year, and occurs most commonly in the winter months, in Northern hemisphere from November to March. Annually, millions of people worldwide have influenza, and hundreds of thousands of people die because of influenza infection [25]. A vaccine against seasonal influenza is developed every year, which effectively prevents the transmission, serious symptoms and deaths of influenza [26]. The symptoms of influenza are similar to those of COVID-19: from runny nose to high fever, and usually last up to two weeks. The incubation period of influenza is shorter than in COVID-19, typically around two days. [25].

Common cold is usually caused by viruses, mostly rhinoviruses. Symptoms typically include sore throat, nasal congestion and cough. A complication of the common cold can be, for example, a sinusitis caused by a bacterium. The incubation time for different viruses is heterogenic. For example, symptoms may start as early as 12 hours after rhinovirus exposure. The symptoms may worsen for 2-3 days, and on average the common cold lasts about a week. [16].

During respiratory infections, the body undergoes physiological changes in response to activation of the immune system and those changes can be measured, for example via wearables. Elevated inflammatory biomarkers, such as C-reactive protein, are associated with an increase in resting heart rate (RHR) [27], and even slight deviations from normal body temperature can significantly elevate heart rate [28]. More specifically, heart rate can increase seven beats per minute (bpm) for ev-

ery degree celcius and even more in high fever. Respiratory rate also elevates while body temperature increases. [29]. In addition, HRV have been reported to decrease during COVID-19 infection [30].

1.3 Anomaly Detectors

One approach to observe abnormal heart rate is to use different anomaly detectors. In the time series data, anomalies are patterns that deviate from normal behavioral data and anomaly detection is an action to identify those patterns. The goal of anomaly detection is to find the anomalous behavior of the unseen data based on historical events. [31]. Anomalous region in wearable data could be just outlier caused by improper device use or actual physiological outlier like an infection.

Isolation Forest is a simple and efficient anomaly detector based on binary tree structures. In Isolation Forest detector flow, several binary isolation trees are randomly created from the data. For every generated tree, distances between the root of the tree and each node are calculated. The shorter the distance a point is from the root of the tree, the more likely it is an anomaly. Deviating points in the data are most likely those with the shortest distance number on average, i.e. in several trees in the forest they are closest to the root. Ideally, the amount of anomalies in the data is minimal and anomalies differentiate from the non-anomalous, normal, data. [32–34].

Elliptic Envelope is an anomaly detector that utilizes normal distribution, also known as Gaussian distribution, of the data for the detection meaning that the input data must be normally distributed. Anomalies from Elliptic Envelope are located to the tails of the univariate distribution or in multivariate case the edges or outside of the shape. Using the properties of the Gaussian distributed input data, an ellipse is fitted around the central mode using a robust covariance estimate of the data, and the Mahalanobis distance derived from this estimate is then used to set the

threshold for detecting outliers or anomalies. [34, 35].

Finite state machine (FSM) approaches anomaly detection via mathematical modelling. In a given time point, the machine can be only in one of the defined finite number states. In time series data, that means that the output of certain time point is in an unambiguous state. The flow of an FSM begins with defining an initial state. When an action occurs, such as receiving a new daily value, the FSM transitions to a new state which is based on the current state and the action. The process also allows the machine staying in the same state if the action does not trigger a transition. [36].

There are also plenty of other techniques used for anomaly detection. Statistical methods use z-score standardized data, and significance level of 0.01 or 0.05 is considered as anomalous. That kind of statistical approach is only suitable for one-dimensional data. In addition, several machine learning algorithms are used for anomaly detection. For example, supervised machine learning method k-Nearest neighbours calculates distances between data points and those distances are used for anomaly detection. [34]. Besides of k-Nearest neighbours, supervised machine learning algorithms like Support Vector Machines and logistic regression. Algorithms utilizing neural networks are unsupervised machine learning methods, and they are also suitable for anomaly detection. [37].

In general, multiple challenges are related to anomaly detection. Firstly, especially for supervised machine learning based detectors, high amounts of labeled data are needed to train the model sufficiently, and acquiring high-quality labeled data is time-consuming. Secondly, the definition of normal or anomalous is not straight forward. Threshold to determine if point suspected to be anomalous is for real anomaly should be considered carefully. That often requires domain-specific knowledge and can vary greatly between different contexts, making standardization of the detector thresholds difficult. For example, in detecting infectious periods from

data, threshold for suspicious physiological alterations must be carefully considered so that most of real diseases are detected but all possible changes are not considered as anomalies. Thirdly, generalization of the detector to other applications or fields might not be feasible. Models trained on certain datasets may not perform well on different datasets because data distributions vary depending on where the anomalies come from. [31].

1.4 Related work

During and following the COVID-19 pandemic, multiple research groups have investigated the impact of respiratory infections on physiological responses, such as RHR, overall daily heart rate, HRV, sleep duration, and step counts, all of which can be measured through wearables [38–49]. Beyond COVID-19, the detection of other potentially serious infections, including influenza-like illnesses, has also been explored [48, 50, 51]. Because COVID-19 and influenza usually cause more severe symptoms, they are easier to detect by wearables compared to common cold. On the other hand, usually milder symptoms of common cold might be confused with normal physiological changes. The idea and concept of these kind of studies is presented in Figure 1.2.

The most accurate methodology for detecting infections, optimally reflecting real-world conditions and needs, involves gathering data from volunteers who wear devices during their everyday activities and also report their sick days.

Several physiological parameters have been tested with COVID-19 or other illness detection using wearables. The most used of them are RHR [40, 42, 44, 45, 47, 48], HRV [43, 44, 46, 49], respiratory rate [43, 44, 46, 49], sleep [40, 42, 45, 48], and step counts/ activity level [40, 42, 43, 45, 47, 48]. These parameters are preferred because most wearables (like smartwatches) can measure them, and their patterns may alter during an infection.

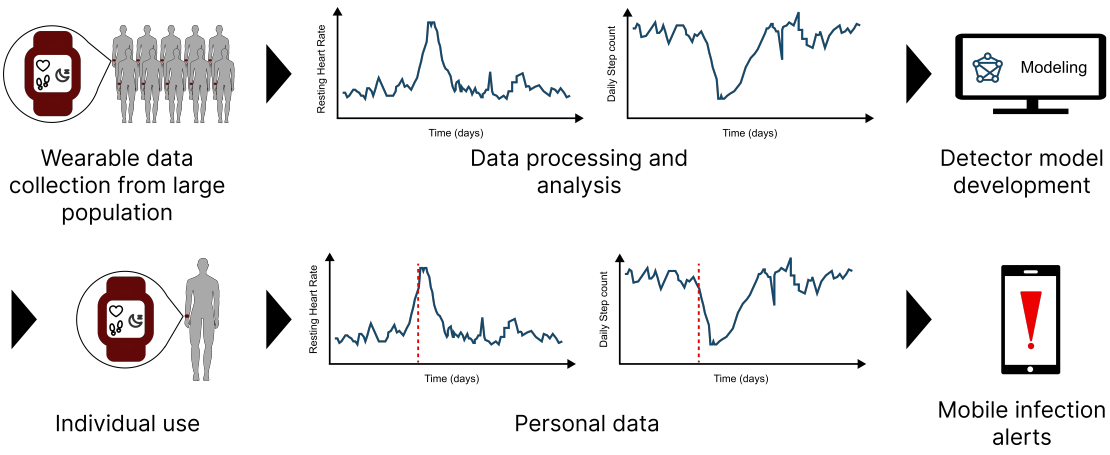

Figure 1.2: **Respiratory infection detection using wearables.** For working detectors, first longitudinal wearable data from large population is needed. That data is then analyzed and anomaly detectors are developed, possibly using machine learning. The models can be applied to individual level, and when anomalies are detected the user gets alert in their phones.

Most studies have been done using common wearables in the market, including Fitbit [45, 46, 48], Apple Watch [41], WHOOP [44], and Oura ring [43, 49]. In previous reports, wearables have been used either individually or in combinations [38, 47]. The number of participants and the ratio of healthy to positively tested individuals vary across studies, with participant numbers ranging from a minimum of 50 to several hundreds. Many of the studies have focused on only symptomatic participants of whom some had positive COVID-19 test, enabling comparisons between COVID-19 positive and negative cases.

Detection algorithms, which aim to identify anomalies or distinguish COVID-19 positive cases from negatives, utilize a range of machine learning methods. These include gradient boosting [44, 47], convolutional neural network [46], logistic regression [47], random forest classifiers [38, 43, 47], k-Nearest Neighbor [47], support vector machine [47], FSM [38] and Elliptic Envelope [45]. Some of these studies used their own detection methods [41, 49].

Comparative analyzes between COVID-19 and other viral respiratory infections

reveal interesting results. According to a study using WHOOP straps to monitor respiratory rate, COVID-19 could not be distinguished from non-COVID-19 infections [44]. The elevation in resting heart rate (RHR) was noted to be less significant among symptomatic people tested COVID-19 negative individuals when compared to COVID-19 positive cases [47]. Furthermore, the symptom period of COVID-19 was significantly longer than that of non-COVID-19 flu [48]. Physical alterations caused by COVID-19, including changes in RHR, HRV, RR and activity rate, could be detected on average 2.75 days prior to the positive test, whereas symptom onset was on average 1.98 days prior to positive test [43].

While respiratory rate appears to be relatively stable within individuals, intra-individual changes are noticeable. This variability makes it hard to set a universal threshold or detector, but creating one from individual data is still possible. [44]. Comparing changes in RR to changes in HR and HRV, respiratory rate had largest alterations around symptom onset [46] making RR efficient parameter for detection of COVID-19 and other respiratory infections.

RHR has been demonstrated to been elevated in respiratory infections [42, 45, 47]. However, the timing of RHR elevation varies across studies , ranging from occurring "near symptom onset" [45], to a few days before and after symptom onset [42], and even up to 13 days prior to symptom onset [47]. During COVID-19, RHR has been reported to increase by an average of 1.65 beats per minute above baseline [47].

Results on HRV changes during respiratory infection is not much studied. However, patterns in HRV have been detected to alter 7 days before and 7 days after COVID-19 diagnosis compared to study baseline [43].

As described in Table 1.1, different detectors and methods to detect COVID-19 or other respiratory illnesses has been explored. For instance, using the Gradient Boosting method and nocturnal respiratory rate, a study achieved a detection rate

Table 1.1: **Summary of previous studies.** PPV:Positive predictive value, NPV: Negative predicted value, CNN: Convolution neural network, ITA: Intelligent Testing Allocation, ML: Machine learning, AF: All-Frequency, AHF: All-High-Frequency, FHF: Fitbit-High-Frequency

Paper	n	Device	Disease	Parameters	'Detector'	Result
[40]	105	Fitbit & Apple watch	COVID-19	steps, RHR, sleep	??	Symptomatic COVID-19 positive and symptomatic COVID-19 negative participants were differentiable with high accuracy (AUC = 0.75) using wearable sensor metrics.
[41]	297	Apple Watch	COVID-19	HRV	-	-
[42]	41 + 85 + 1126	Smart-watch	COVID-19 & ILI	RHR, sleep, step count	-	-
[43]	74	Oura ring	COVID-19	HR, HRV, RR, activity rate	Random Forest	COVID-19 was detected on average 2.75 days before diagnostic test with 82% sensitivity, 63% specificity and AUC of 0.819.
[44]	271	WHOOP strap	COVID-19	RR (HRV, RHR)	Gradient Boost	Training set sensitivity 0.411, specificity 0.985, PPV 0.909, NPV 0.817
[45]	32+13 + 72	Fitbit	COVID-19	RHR, sleep, steps	-	-
[46]	2745	Fitbit	COVID-19	HR, HRV (night RR, HR, HRV)	CNN	AUC: 0.77 $\pm$0.018, 99% specificity $\rightarrow$0.259 sensitivity
[47]	1265	Smart-watch	COVID-19	RHR, steps	ITA + 5 different ML methods	Using only RHR-based features, AUC-ROC was lower compared to using only step-based features. Cross-validated AUC-ROC values for RHR vs step features were 0.64 vs. 0.67, 0.63 vs. 0.69 and 0.68 vs. 0.72 for AF, AHF and FHF training sets, respectively.
[48]	41+85	Fitbit	COVID-19 & ILI	steps, RHR, sleep	-	-
[49]	50	Oura ring	COVID-19 / fever	HR, HRV, RR, temperature	-	-
[52]	333	Fitbit	COVID-19	RHR, sleep, steps	Multivariate logistic regression	The model achieved AUC of 0.80 (interquartile range 0.73–0.86) in accurately classifying symptomatic individuals as either COVID-19 positive or negative

of 20% among positive individuals 2 days prior to symptom onset, rising up to 80% three days after symptom onset [44]. CuSum detector detected correctly 63% COVID-19 positives before and around symptom onset [45] and and the specificity and sensitivity of CuSum were 84% and 72%, respectively [38]. Random Forest Classifier for COVID-19 detection using Oura ring and HR, HRV, RR, temperature and activity levels as parameters achieved 82% sensitivity and 63% specificity. Notably, the model in a particular study performed best when incorporating all five parameters compared to excluding any individual parameter. [43]. Comparison between sensitivities and specificities in studies employing machine learning is presented in Figure 1.3. Overall specificities have been high in all studies but range of sensitivities have been wider.

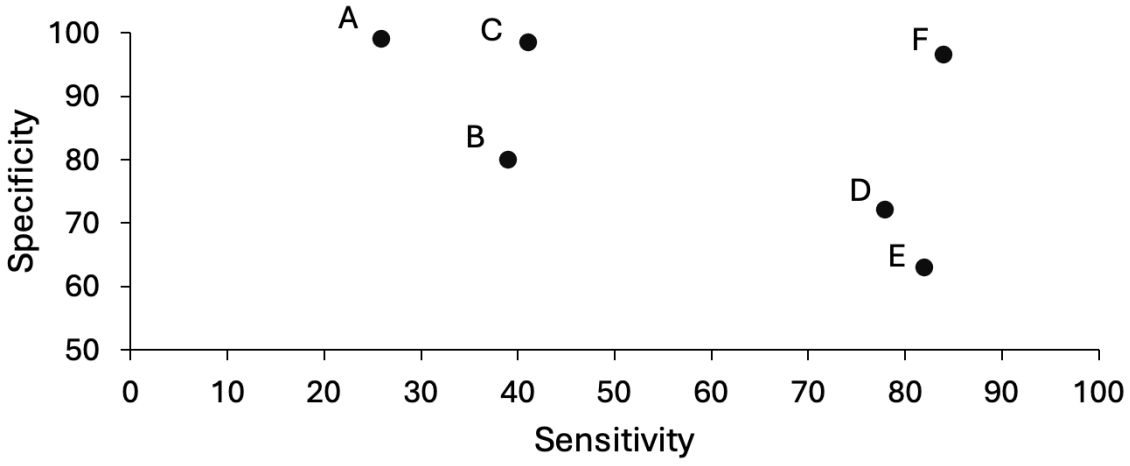

Figure 1.3: **Comparison between sensitivity and specificity in different studies. A** [46] Sickness prediction done with Convolutional neural network. **B** [52] COVID-19 detection done with Multivariate logistic regression with RHR. **C** [44] Healthy versus infected day classification training results. **D** [53] Passive COVID-19 detection using gradient boosting. **E** [43] COVID-19 detection using machine learning. **F** [50] Average of influenza and rhinovirus detection done using random forest.

In addition to studies conducted under free-living conditions involving numerous participants, some groups have pursued infection detection using wearables in controlled environment. In those studies, virus has been inoculated, ensuring that the

viral exposure time is strictly known (compared to 'natural' infection).

In a study from year 2021 [50], 31 healthy participants were inoculated with H1N1 influenza virus, while 18 participants received rhinovirus inoculation. Utilizing Empatica E4 wristbands, heart rate, dermal temperature, electrodermal activity, and movement data were collected for one day before and eleven days after H1N1 inoculation, and four days before and five days post-inoculation for rhinovirus. Symptom surveys, assessing both observable and non-observable symptoms, were administered twice daily, with viral infection confirmed through daily laboratory, polymerase chain reaction (PCR) testing. Predicting of the infection was done using Random Forest Classifier.

In a similar study from 2023 [51], 20 healthy participants were inoculated with H3N2 influenza virus. Bittium sensors were used to monitor single-lead ECG and movement data, from where heart rate and intra-beat interval data were extracted. Data was collected seven days prior and ten days after the inoculation. Participants also completed symptom surveys twice a day, utilizing a four-step scale to evaluate eight symptoms. Infection confirmation was achieved through twice-daily collection of blood samples and nasal swabs for PCR analysis. Predicting infection on each participant was done with semi-superivsed multivariable anomaly detection model.

Both of those controlled studies [50, 51] concluded that infection could be detected prior to symptom onset using machine learning models. With random forest classifier, H1N1 influenza was detected with 89% accuracy 36 hours after inoculation, and rhinovirus was detected with 88% accuracy within same time period. The random forest classifier was also able to accurately differ asymptomatic/non-infected cases from H1N1 and rhinovirus cases (accuracies 83% and 92%, respectively).

1.5 Research Motivation

Retrospective studies involving hundreds of participants have their limitations. Risks include participants' responsibility for data collection in free-living environments. Ensuring proper device usage, and accurately completing subjective symptom questionnaires that risk type can be minimized. [43, 48]. When classification relies solely on symptom questionnaires or self-administered tests, diseases may be misclassified as asymptomatic despite biological responses occurring [49]. Additionally, when selecting parameters, attention must be paid to whether changes are disease-induced or influenced by other factors, such as psychological or behavioral changes in activity levels during COVID-19 [40].

On the other hand, controlled viral exposure studies are valuable for defining illness detection methods. However, limited number of participants in these studies can restrict model validity, since there is big variance in physiological responses between individuals. Knowing the precise timing and dosage of viral exposure improves predictive accuracy, as demonstrated in [50]. Furthermore, outcomes are influenced by closely monitoring viral levels, employing structured symptom surveys, and controlling data collection procedures.

Respiratory infection detection research is mainly done in the past couple of years. Especially COVID-19 pandemic increased the number of studies in this field. Many research groups have tried to identify COVID-19 from other non-COVID-19 respiratory infections using wearable data, whereas some studies have detected any respiratory infection. Also, there is a knowledge gap whether an infection is detectable before or after its onset, raising the question, if the main goal is to detect the onset or the existence of infection.

Most of the studies about detecting respiratory infections using wearables have been pilot studies meaning that no further validation of data collection or detectors has been yet done. Additionally, significant differences between results of different

studies raise concerns about potential overfitting in some cases. In the future, multicenter studies and better validated and tested methods are crucial for more precise and accurate results. Also, most of the studies focused on just infected participants but data from healthy people would have given negative controls and more data about normal deviations in parameters.

To stop viral spreading, it would be important to detect infectious diseases early. There are sensitive PCR based laboratory tests and other laboratory tests to detect COVID-19 and influenza but they are expensive and time-consuming. On the other hand, there are at-home test kits available, but the their accuracy is lower [43, 45]. In addition, even home-tests are not performed if there are no symptoms or recent exposure to the virus. Therefore, there is a need for accessible and affordable infection detectors. These detectors would allow early detection of respiratory infections, enabling infected individuals to isolate and recover before their condition worsens. While fever can be easily detected through temperature measurements, wearable devices may provide additional tools to utilize this well known indicator.

Early identification of respiratory tract infections, even before the symptom onset, and isolating the infected person and starting to rest would produce societal benefits. First, effective and accurate detection of infection using wearable devices would contribute to individuals having sufficient information about their own health, so that the disease can be treated either independently at home or in the context of medical care, and the duration and severity of the disease could be reduced. In addition, self-imposed isolation from other people has been found to reduce the spread of viruses, especially COVID-19, in the population [54]. Thus, the burden of medical treatment is also lightened, when there are fewer sick people and, thus, also seriously ill ones.

The purpose of this thesis is to determine, if it is possible to detect respiratory infections from wearable data. In this work, three publicly available high resolution

datasets are used. Infections are detected using three earlier published detectors. In addition, small pilot study with self-collected data was done to compare results with previous studies.

The research questions of this thesis are:

Q1. Is it possible to detect respiratory infections from wearable data?

Q2. Is it possible to detect respiratory infection onset from wearable data?

Q3. How sensitive/specific existing detectors are?

Q4. Are previously published results reproducible?

Q5. How the same detectors work with our own pilot data?

Generative AI (ChatGPT 3.5 and 4o) has been used in this thesis for rephrasing, spelling and with coding problems (for example figures).

2 Methods

2.1 Data description

Data 1 [45] The dataset consisted of 32 COVID-19 positive and 13 symptomatic COVID-19 negative participants collected between February 2020 and June 2020. COVID-19 negative participants had either influenza B, rhinovirus or some other unknown infection. Five of the COVID-19 positives had also non-COVID-19 illness in their data period. In the published dataset, there were also data from 73 healthy participants. In this thesis, Data 1 'positives' (people who reported some disease) and 'negatives' (healthy individuals = did not report any disease) are handled separately. Main anomaly detection analysis was done only with infected individuals ('positives'). The data was collected with Fitbit smartwatches. Heart rate was collected at 15 second resolution, step values at 60 second resolution, and sleep data as sleep stage intervals. In the published dataset, HR values over 200 bpm and under 30 bpm were removed as well as data duplicates. Timestamps were given as randomized date and time (YYYY-MM-DD HH:MM:SS). Symptom and diagnosis dates were given in supplementary Excel file for each participant. For subject ID AA0HAI1, three different measurements were given. Two of those clearly overlapped, so they were combined and considered as one measurement.

Data 2 [38] The dataset consisted of total 2123 individuals, with 84 reporting COVID-19 test date, and some of them reported also symptom onset date. Of those

84 COVID-19 positive participants, 49 participants used Fitbit for wearable data collection, and only the Fitbit users were considered for this thesis because then those were comparable with Data 1 and data collected with other devices (Apple watch) did not have high enough resolution. Sampling frequencies and resolutions were not reported in the paper, but they were similar to Data 1, since Data 2 study was extension of Data 1 study. In addition to COVID-19 infected individuals, there were 2039 participants who did not report any symptoms or COVID-19 tests. The lack of symptom data of these people led to the rejection of their data in this thesis.

Data 3 [49] The dataset consisted of 50 participants having fever (and other COVID-19 symptoms). The data was collected using Oura ring. Dermal temperature was collected at one minute resolution, RR at 30 second resolution, HR at mean per 5 min inter-beat interval, HRV RMSSD (root mean square of successive differences) per 5 minutes. RR, HR and HRV were from PPG signal sampled with 250 Hz. In the published dataset all the data was presented in a 1-min resolution linear interpolation of the original data.

2.2 Data pre-processing

For all pre-processing, analysis, and detectors, Python version 3.10.9 was used. In addition following versions of packages were used: Numpy 1.23.5, Pandas 1.5.3, Scikit-learn 1.2.1 and SciPy 1.10.0.

For Data 1 data pre-processing, available code published in [39, 45] was used. RHR was computed by first resampling heart rate and step data to one minute resolution (original 5 second and 1 minute resolution, respectively). 'Resting' was determined as time where steps were 0 at least 12 consecutive minutes. Then NaN values were dropped and rolling mean over 400 samples (400 minutes $\approx$ 6.5 hours) was taken, and finally the data was resampled to one-hour resolution.

For Data 2 pre-processing, before RHR computations, one additional step was done. Step files were resampled to 1-minute resolution and all NaN values were replaced with zeroes. Otherwise, data pre-processing steps were the same as with Data 1.

Further, for daily variation reduction seasonal decompositioning was done. After decomposition, measured parameters were z-score standardized. Seasonal corrected and standardized data was used for statistical analyses. In detector performance analyses, data were used as raw RHR or z-score standardized.

In Data 1 all sickness dates (symptoms, diagnoses, recoveries) in supplementary data were checked. First, all dates not in the data period were dropped. Then, the beginning of the disease was considered as the first symptom or diagnose day and recovery was determined to be 7 days after beginning of the disease. The disease onset could have occurred at the very beginning, in the middle, or near the end of the data period. One participant did not have any sickness dates and three participants had two sickness periods that were handled as separate sickness events.

For Data 2, symptom onset date was considered as disease onset if possible, otherwise test date was taken for the disease onset.

For Data 3, there were originally no timestamps in the data, but, there was a mention that the data is sampled with 1-minute resolution. For seasonal decomposition, timestamps were needed so artificial timestamps starting at 2024-01-01 00:00:00 were created. Raw data was smoothed with rolling mean with 400 samples as other datasets. RHR was computed by averaging 5 smallest values for each hour. Dermal temperature, HRV and RR were composed to hour resolution by averaging 60 consecutive samples. After computing hourly values, same decomposition and standardizations were done.

As described in the paper [49], the symptoms began 20 days prior to the end of the data, or maximum of the 45 days from the beginning of the data. Using that

information, disease onset was determined to be 20 days so 479 hours (samples) before data ends.

For statistical analysis, pre-detection period was time period from three days prior to disease onset to disease onset. Post-detection period was time period from disease onset to six days after disease onset. For baseline, all time points where the timestamp was earlier than 14 days before the disease onset and later than 21 days after the disease onset were selected. This was based on incubation time of the respiratory viruses and sufficient recovery time, meaning there should not be any incident of the same infection in the baseline period.

After data pre-processing, all three datasets were harmonized so that data were resampled to one-hour resolution and RHR values were calculated. Additionally, all disease onsets were collected.

Data 1 had 44 participants and on average 83.5 (standard deviation 25.0) days of data each (See Table 2.1). On average, participant had 43.6 (19.7) days data before disease onset and 39.9 (24.3) days data after disease onset. Mean number of days of the baseline data was 47.5 (26.3). Data 2 had on average 162.8 days of data from 47 events. Mean of 100.6 (53.9) days before and 62.2 (63.9) days after the disease onset. Mean length of baseline data was 127.4 (71.0) days. The structure of Data 3 was different, since all of the 50 events should have had same amount of data (20 days) after the disease onset. Average days of data per event was 61.6 (7.8) days and of that 41.7 (7.8) days was before the disease onset. Length of baseline data was on average 27.7 (7.8) days. In Data 1, three individuals had two different diseases within the data period, and one did not have any disease within data period (but somehow it was in positive dataset).

Table 2.1: **Data description**. Mean days (standard deviation) of data per participant. **n**: number of participants. **Prior**: Time period before symptom onset **Post**: Time period after symptom onset **Total**: Total length of data **Baseline**:time points earlier than 14 days before the disease onset and later than 21 days after the disease onset

	n	**Prior**	**Post**	**Total**	**Baseline**
Data 1	44	43.6 (19.7)	39.9 (24.3)	83.5 (25.0)	47.4 (26.3)
Data 2	47	100.6 (53.9)	62.2 (63.9)	162.8 (73.2)	127.4 (71.0)
Data 3	50	41.7 (7.8)	19.9 (18e-13)	61.6 (7.8)	27.7 (7.8)

2.3 Own pilot dataset

Our own dataset consisted of wearable data from three individuals. Data were collected with Oura rings, which were advised to be used as much as possible during days and nights. Individuals were asked to report all days that they were sick, especially symptom onset dates. All raw data was downloaded from Oura website. Heart rate, sleep and daily activity data files were used.

RHR was calculated from night-time RHR and daytime RHR. Night-time RHR was given in 5-minute resolution and night-time was calculated using Oura's own assesment of bedtime. Daytime RHR was as well collected with 5-minute resolution, but only HR data with activity/status label 'rest' was accepted for RHR. Notably, daytime RHR was added to software about two months after the start of data collection, so before that only night-time data was usable. Data from daytime and night-time were then merged. For further data pre-processing, rolling mean of 80-100 samples (approximately 7 hours, depending on how densely data were sampled, number of samples depended on the total length of data) and finally the data was resampled to determined resolution (5min/1h/6h/12h/1d/3d). After pre-processing, our own data were harmonized to previously published datasets.

2.4 Detectors

Statistical tests were used to find out changes between different data periods. Averages of every parameter in all time periods (baseline, pre-detection, post-detection) for every participant/sickness event were computed. Comparisons were made visually using boxplots and statistically using t-tests, assuming the null-hypothesis that the two groups were similar. P-values less than 0.05 were considered as statistically significant.

RHRAD (Resting Heart Rate Anomaly Detection), a detector that uses Gaussian density estimation for outlier detection, was implemented using *EllipticEnvelope* from Sklearn package [55] using parameters random_state = 42 and support_fraction = 0.7. For contamination, default value 0.1 was first used for detector performance evaluation, but later contamination was one of the parameters tested in range 0.01 and 0.5. RHRAD identified outliers by detecting extreme points in the dataset's distribution. The detector calculated the distance of each observation from the overall mean, considering both univariate and multivariate outliers. [45]. The outliers were observed for one participant/event at the time. Anomaly score came from detector's *desicion_function*. More negative values mean more anomalous. The function used threshold 0 for anomaly definition.

The second detector used was Isolation Forest. It utilized decision trees for unsupervised anomaly detection, and was implemented using *IsolationForest* function from *sklearn* package [55] with random_state = 0. For initial analyses contamination 0.05 was used but as in RHRAD, hyperparameter tuning was later on done. [38].

The third detector was Night signal, originally published in [38], which was based on FSM. Average RHR of the day (originally average RHR overnight) was compared to median of average daily RHR values for all previous nights. Red alarm (detected anomaly in this thesis) was raised, if average daily RHR was over the

median + certain threshold (in the original paper for red alarms 4 bpm) for at least two consecutive nights. Original publication had also yellow alarm with at least one day where threshold value 3 bpm over median.

One-hour resolution RHR was given to Night signal as input. First it computed daily mean and then missing days were filled with two next or prior day average. Then median of average daily RHR values for all previous nights was computed. Potential red alarms/anomalies were days where daily average exceeded the past days median plus the threshold. For the actual anomaly detection, all potentials were checked and if there was a potential alarm for at least two consecutive days, an alarm was raised (= anomaly is detected).

For detector performance analysis, RHR data (no standardization or seasonal correction) was input to RHRAD and Isolation Forest. For Figure 3.5 the results were resampled from hourly-resolution to daily-resolution by taking (nan)mean of all values of the day. Threshold for anomalous day was mean detector score below 0.

True positives were detected anomalies within infectious time period (for example from 3 days prior to symptom onset to recovery day (7 days after onset)). False negatives were time points detected not-anomalous in infectious time period. False positives were detected anomalies outside the infectious time period and true negatives correctly not-anomalous detected time points outside of infectious period. True positive rates (TPR) and False positive rates (FPR) were computed for ROC (Receiver operating characteristic) curves and AUC (Area under the curve) computations.

In addition to ROC curves and AUC values, detector performance was evaluated using sensitivity, specificity and accuracy. Sensitivity describes the rate of successful positive anomaly findings in the detection period, whereas specificity describes successfully negative detected points. Accuracy describes correctly predicted values

among all predictions. ROC represents detector's performance graphically using TPR and FPR with changing thresholds, and AUC describes the overall performance of the detector by quantifying the area under the ROC curve.

For testing Night signal reproducibility with Data 2, the algorithm, along with the detection window of -21 to 0 days relative to disease onset, was executed as detailed in the original publication [38]. In that case true positives were defined as participants who received at least one red alert in that detection window and false negatives were the participants who was not alerted with red alert in that period.

For detector hyperparameter tuning for RHRAD and Isolation Forest, contamination parameter (proportion of outliers in the dataset) values tested were 0.01, 0.05, 0.1, 0.2, 0.3, 0.4 & 0.5 to find out the best detector execution. In addition, different data resolutions (for input data to detector) were tested: 1-minute, 1-hour, 6-hours, 12-hours, 1-day, 3-days. Also, several time periods/windows for infectious period were tested: -3 to 0, -5 to 0, 0 to 3, 0 to 5, 0 to 10, -3 to 7 and -14 to 21 days to symptom onset.

2.4.1 LAAD

Deep learning anomaly detector using Long Short-Term Memory Networks-based autoencoder (LAAD) [39] used baseline RHR data for training. In contrast to the original publication of the model, as an infection period days -3 to 6 related to the disease onset was used instead of days -7 to 21. Days -20 to -10 were non-infectious period and all days after the disease onset and infectious period were recovery period. Baseline data was split to training data (95 %) and validation data (5%). All data were z-score standardized using the properties of the training data. The test data, consisting of all data not in training or validation, were further divided to to normal (non-infectious period) and anomaly (infectious period). In summary, baseline data was all data before 20 days of disease onset and it was used

for training and validation. As 'non-infectious' test data, days -20 to -10 was used and as 'infectious' test data days -3 to 6 around disease onset was used.

Next, the data was windowed to 8-sample batches (eight samples used for one detection) and reshaped to tensor format. Then training data were augmented (total size of augmented data was eight times the size of the baseline data) using seven different techniques: scaling, rotations, permutation, magnitude-wrapping, time-wrapping, window-wrapping and window-slicing. The actual autoencoder-decoder system was based on Long-Short Term Memory (LSTM), and it's purpose was to find anomalies from RHR data given in time series. The model training was stopped at the optimal point to prevent overfitting. Four LSTM layers were used in the final anomaly detector: RepeatVector, TimeDense and 128 hidden neurons for both the encoder and the decoder. Mean squared error was used as reconstruction error calculation and ADAM algorithm was used for optimizing the learning process. The model has been described in more detail previously [39]. LAAD performance was evaluated through performance metrics: sensitivity, specificity, precision, recall and accuracy.

3 Results

When daily averages were combined within all three datasets, it was possible to review if the infection caused visible alteration in RHR. Overall, RHR raised just before and peaked a few days after of the disease onset, when compared to the mean RHR of the previous 30 days. Data 2 had less significant peak around illness showed another peak at 25 days after disease onset, while the dispersion of the data was large. Figure 3.1 shows RHR daily averages of each datasets with 95% confidence intervals. It can also be noticed that deviation within the datas was huge, suggesting that some had really deterministic peak around disease whereas some had more flat RHR around the whole data period.

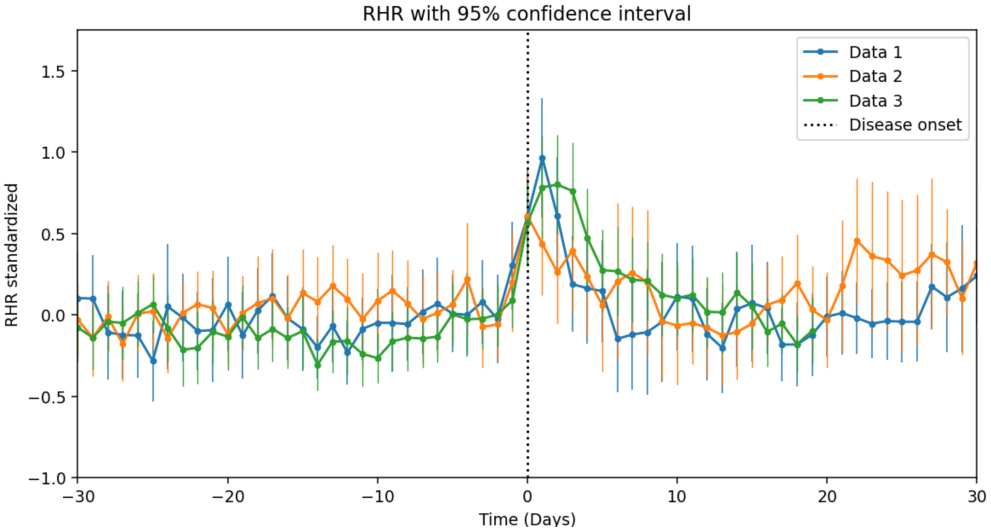

Figure 3.1: **RHR change around disease onset.** Daily average over the whole datasets. Overall, in all three datasets, RHR seems to have peak around the disease onset. The peak is just after day 0 but the raise starts before onset. Error bars show the 95% confidence interval of the day.

3.1 Statistical analyses

Two-tailed t-tests were used for feature comparisons between different time periods. The smallest variations in all three datasets were in baseline distributions, while both pre-detection and post-detection periods' distributions were much wider. Figure 3.2 shows alterations in resting heart rate in all three datasets. In all three datasets, change from baseline to post-detection period was statistically significant (p-value < 0.05). Post-detection period RHR was higher than baseline RHR. Only in Data 3 the difference between pre-detection and post-detection periods was statistically significant. All RHR comparison p-values are shown in Table 3.1.

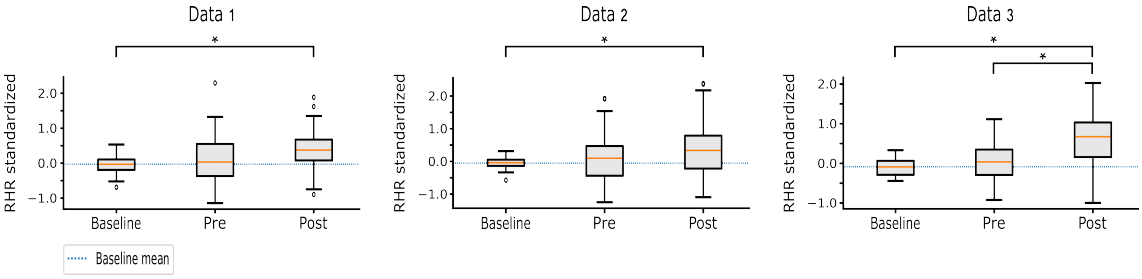

Figure 3.2: **Resting heart rate distributions across time periods.** Differences between distributions in different time periods. Baseline: time period before 14 days prior and 21 days after disease onset (symptom onset/diagnosis date). Pre: Pre-detection time, days -3 to 0 from disease onset. Post: Post-detection time, days 0-6 respectively from disease onset. * p-value < 0.05.

Besides of RHR, Datas 1 & 2 had daily step counts. Variation during baseline daily step count was small in Data 1 and Data 2 whereas pre- and post-detection periods variation was larger. Especially Data 2 pre-detection step distribution was wide. Statistically significant decrease was found in both of those datasets between baseline and post-detection periods. In addition, there was a statistically significant decrease in Data 2 from pre-detection period to post-detection period (see Figure 3.3A). P-values of baseline versus pre-detection period were not significant (see Table 3.1).

In addition to RHR, Data 3 included temperature and HRV data. Temperature

Table 3.1: **Time period feature comparisons.** All baseline versus post-detection comparisons were statistically significant. In Data 3, Temperature was the only statistically significant parameter in all time period comparisons. There were no significant changes between baseline and pre-detection times in any other parameter than temperature. **Bl**: Baseline. **Pre**: Pre-detection time - 0-3 days prior to the disease onset. **Post**: Post-detection time - 0-6 days after the disease onset. Statistically significant ($p < 0.05$) values are bolded.

		RHR	**Steps**	**Temperature**	**HRV**
Data 1	Bl vs Pre	0.152	0.515	-	-
	Bl vs Post	**<0.05**	**0.001**	-	-
	Pre vs Post	0.117	0.077	-	-
Data 2	Bl vs Pre	0.258	0.406	-	-
	Bl vs Post	**<0.05**	**<0.05**	-	-
	Pre vs Post	0.063	**<0.05**	-	-
Data 3	Bl vs Pre	0.081	-	**0.008**	0.406
	Bl vs Post	**<0.05**	-	**<0.05**	**<0.05**
	Pre vs Pots	**<0.05**	-	**0.003**	**<0.05**

was increased statistically significantly from baseline to pre-detection, baseline to post-detection and pre-detection to post-detection (see Figure 3.3 B). HRV decreased significantly from baseline and pre-detection to post-detection. However, no statistically significant changes in HRV were found between baseline and pre-detection time.

3.2 Anomaly detection

RHRAD and Isolation Forest used the properties of the entire data period to detect anomalies. Figure 3.4 represents all detected RHRAD and Isolation Forest anomalies for all three datasets. Plenty of false positives were detected for every participant with both detectors. There were no significant differences between RHRAD and Isolation Forest in the number of detections. False positives were detected evenly through the whole data periods. Some individuals did not have any correctly detected anomalies but only anomalies being outside of infection period. For further analysis, it is worth to notice that some individuals have quite little

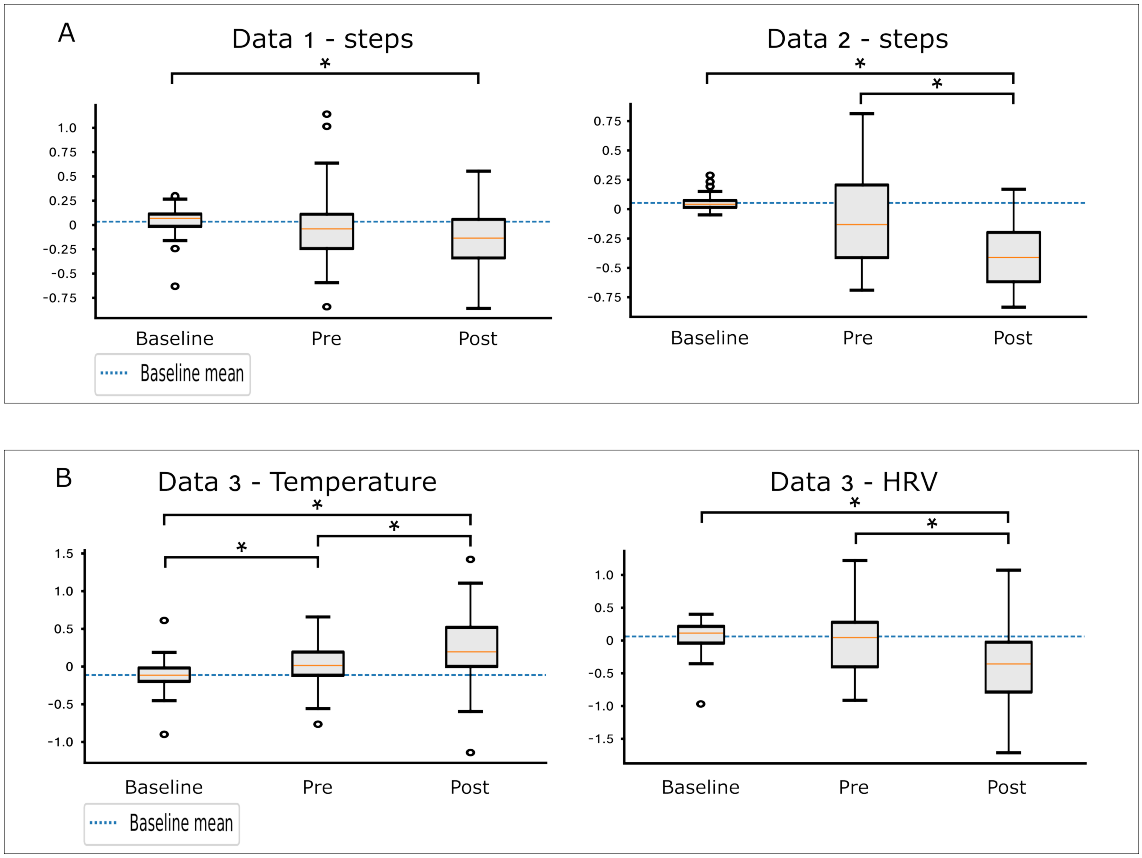

Figure 3.3: **Step, HRV and temperature distributions across time periods.** Baseline: time period before 14 days prior and 21 days after disease onset (symptom onset/diagnosis date). Pre: Pre-detection time, days -3 to 0 from disease onset. Post: Post-detection time, days 0-6 respectively from disease onset. **A.** Data 1 & Data 2 Step distributions across time periods. Steps trend to decrease from baseline in post-detection period. **B.** Data 3 Dermal temperature and HRV distributions across time periods. Temperature differed significantly in all three time periods. HRV differed significantly from baseline to post-detection period and from pre-detection period to post-detection period.

sensor data around reported infection time making anomalies impossible to detect (see the length of red infection line in Fig 3.4).

RHRAD and Isolation Forest could have used high-resolution data, in this thesis one-hour resolution, while Night signal required only one value per day for anomaly detection. For visualization purposes, daily detections were made by averaging RHRAD and Isolation Forest anomaly score and the day was alerted if average score of the day was negative. Threshold for Night signal anomaly was at least two

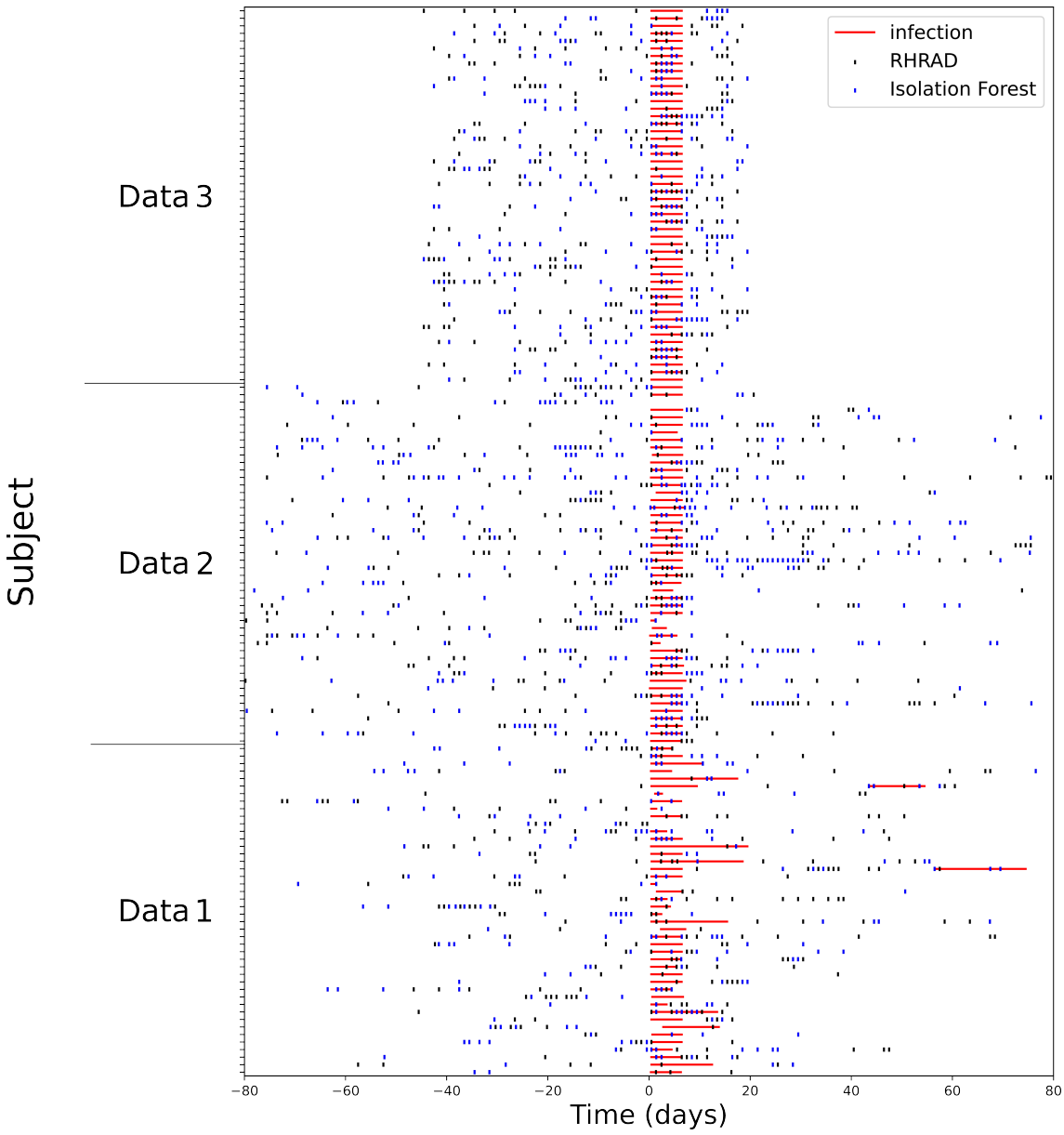

Figure 3.4: **Infection periods and detected anomalies for all datasetes.** RHRAD and Isolation Forest used as detectors. Anomalies thresholded to daily level with threshold 8 positive anomalies within a day to consider the day as positive. Contamination values 0.1 and 0.05 for RHRAD and Isolation Forest respectively were used as detector hyperparameter. Infections starts from symptom/disease onset and ends to recovery period which is 7 days or self-reported (Data 1). Data is limited to 80 days around disease onset for both directions. The plot indicates, that there were numerous false positives, and time period around infection did not differ significantly from other time points.

consecutive days RHR average over 4 bpm above the median of all past days.

Figure 3.5 shows one decent and one improper example of detector performance for each dataset. In proper examples, the clear elevation in RHR around disease onset (Fig 3.5 left column, decent examples) was detected by all three detectors (RHRAD, Isolation Forest and Night signal). Example of successful detection is shown Figure 3.5 left column. There was a fine cluster around infectious period in each example participant but still few false positives appeared. Right column of Figure 3.5 demonstrates that the infection was not detectable in every participant. In Data 2, there were many false positives from all detectors but no clear cluster. On the other hand, in Data 1 and Data 3, there were only few false positives but the data showed no change during the infectious period.

For performance evaluation purposes, ROC curves and corresponding AUC values were computed for RHRAD and Isolation Forest detectors and for each three datasets and all datasets combined. For detections, data were in one-hour resolution and infectious window of days -3 to 7 from symptom onset was used. Overall, AUCs were close to random guess (AUC = 0.5) in all cases. Highest AUC was by RHRAD in Data 2 (AUC = 0.601) and lowest in Data 1 by RHRAD (AUC = 0.552). ROC curves and AUC values are represented in the Figure 3.6. The performance of both detectors were similar, either of them was not remarkably better than the other.

Using parameters used in previous studies (data resolution 1-hour, infectious window days -3 to 7 around symptom onset and contamination values 0.1 for RHRAD and 0.05 for Isolation Forest) detector performance was not very effective. Sensitivities were low, from 0.098 (Data 3 Isolation Forest) to 0.248 (Data 3 Night signal) and specificities high, from 0.902 (Data 2 Night signal) to 0.958 (Data 3 Isolation Forest), and, in addition, accuracies were high (see Table 3.2).

Data from healthy individuals (individuals without any disease during the measurement period) was used to analyze the amount of false positives. Specificity was

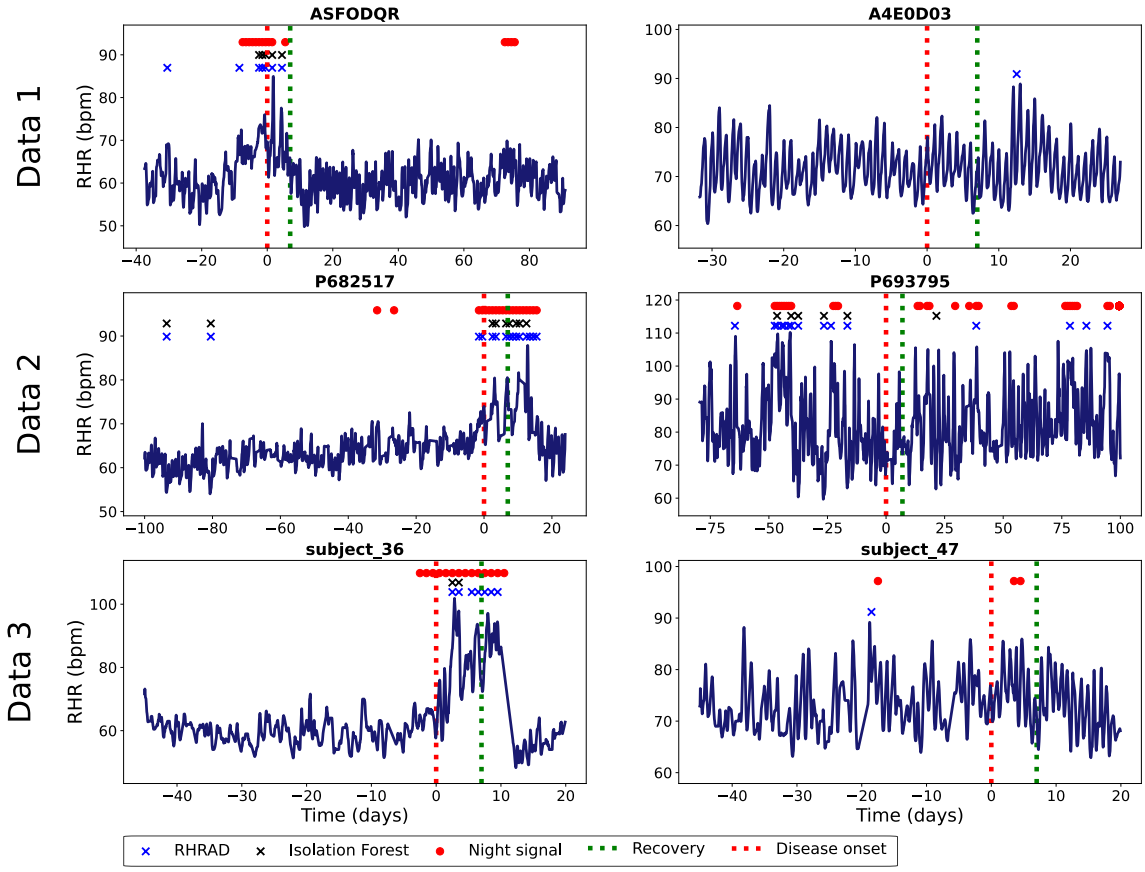

Figure 3.5: **Detector performance examples.** Left column shows almost optimal performance of all detectors in one subject from all datasets. Right column shows inadequate performance of all detectors in one subject in all datasets. Disease onset (symptom onset) is in day 0 and recovery is determined to be in day 7. For RHRAD and Isolation Forest, a day is considered as positive if mean value of anomaly scores is negative. Night signal threshold for alarm is two consecutive days RHR over 4 bpm over past median.

lower within healthy individuals compared to sick individuals, RHRAD 0.9 vs 0.910 and Isolation Forest 0.950 vs 0.955, respectively (see Table 3.3). However, accuracy among healthy individuals was higher compared to sick individuals. Similarly, AUC values were better using data from sick individuals compared to using all individuals as one dataset, RHRAD AUC 0.552 for only sick individuals and 0.504 for all population and Isolation Forest 0.562 for only sick individuals and 0.524 for all population (Figure 3.7).

When testing the reproducibility of Night signal using Data 2, sensitivity of 0.77

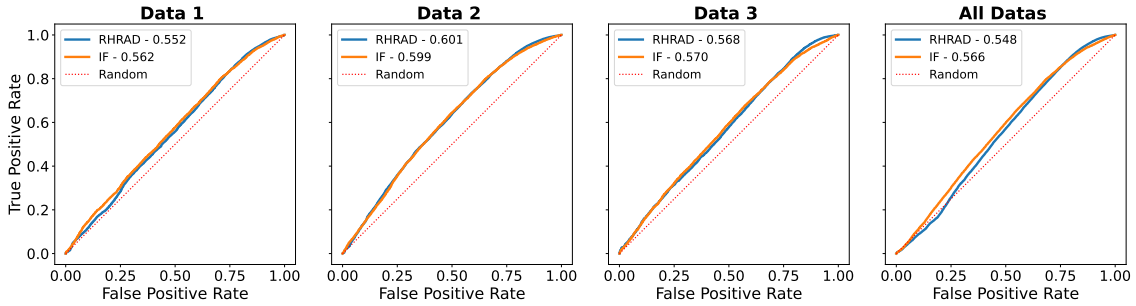

Figure 3.6: **RHRAD and Isolation Forest ROC curves in all of the datasets.** AUC value of RHRAD and Isolation Forest detectors are showed in the legend. For infectious period three days prior and 7 days post the symptom onset are considered.

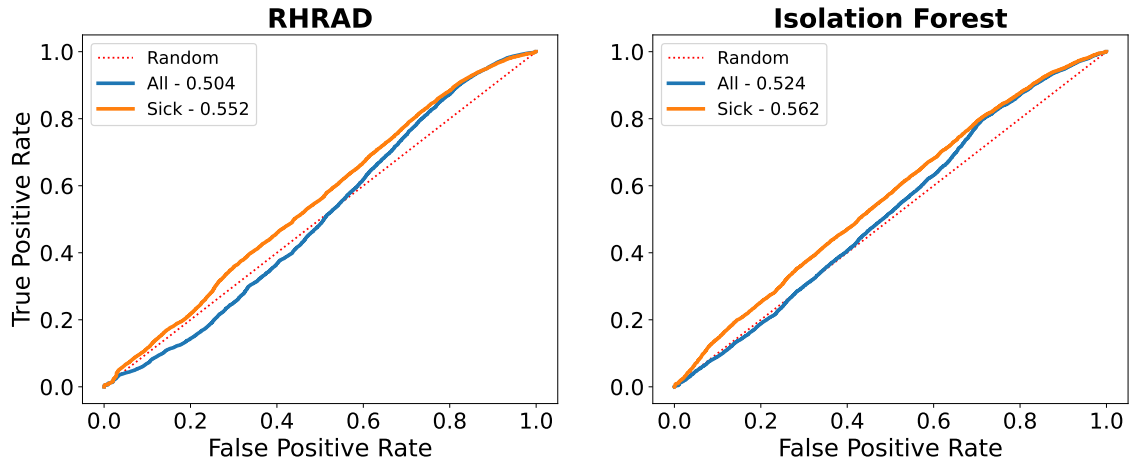

Figure 3.7: **ROC-AUC comparisons between Data 1 positives and sick + healthy.** Adding healthy individuals to the analysis, AUC values drops since more false positives. Results of both detectors are similar.

was achieved, consistent with the results reported in the original publication. This is a significant difference compared to the sensitivity of 0.2 presented in Table 3.2, where true positives and false negatives were defined differently than in the original publication. Additionally, a third approach was tested in which the detection window remained the same, from -3 to 7 days around disease onset, but true positives were defined as in the original publication. Under these conditions, sensitivity of 0.58 was achieved.

To find out best possible parameters for detectors, different combinations were

Table 3.2: **Detector metrics.** RHRAD and Isolation Forest sensitivity, specificity and accuracy of all datasets. Parameters were computed from binary classification using one-hour resolution data, disease window of days -3 to 7 around symptom onset and contamination 0.1 for RHRAD and 0.05 for Isolation Forest.

		Sensitivity	**Specificity**	**Accuracy**
RHRAD	Data 1	0.175	0.910	0.825
	Data 2	0.219	0.907	0.869
	Data 3	0.182	0.916	0.796
	All	0.192	0.909	0.841
Isolation Forest	Data 1	0.092	0.955	0.855
	Data 2	0.123	0.954	0.908
	Data 3	0.089	0.958	0.816
	All	0.100	0.955	0.874
Night signal	Data 1	0.213	0.922	0.842
	Data 2	0.204	0.902	0.865
	Data 3	0.248	0.921	0.812
	All	0.223	0.910	0.849

Table 3.3: **Data 1 metrics.** Data 1 detector performance metrics. **Sick:** Individuals having at least one disease in data period ('official' Data 1), n: 45. **Health:** Individuals who did not experience any disease in data period, n: 73. **All:** sick + health.

	RHRAD			**Isolation Forest**		
	Sick	**Health**	**All**	**Sick**	**Health**	**All**
sensitivity	0.175	-	0.175	0.092	-	0.092
specificity	0.910	0.900	0.903	0.955	0.950	0.952
accuracy	0.825	0.900	0.869	0.855	0.950	0.911

tested and evaluated. Tested parameters were data resolution from 1 minute to 3 days, detection window (just pre-symptomatic, just post-symptomatic and over symptom onset) and detector hyperparameter contamination. The three best and worst parameter combinations evaluated based on AUC of each detector are shown in Table 3.4. Regardless of the parameter combination tuning, AUC values were low. The highest AUC with RHRAD was 0.624 with 3 day resolution, days 0-10 from symptom onset as detection window and 0.2 contamination level. The best AUC with Isolation Forest was even lower, only 0.617 with following parameters: 3 day resolution, days 0-10 from symptom onset as detection window and 0.3 contamination level. The weakest AUC of both detector was 0.504, being close to random

Table 3.4: **Detector performance ranked by AUC.** The best values of each column of both detectors are bolded.

Parameters	AUC	Sensitivity	Specificity	Accuracy
RHRAD				
Resolution: 3d **Window**: days 0-10 **Contamination**: 0.2	**0.624**	0.435	0.814	0.778
Resolution: 3d **Window**: days 0-10 **Contamination**: 0.3	0.623	0.529	0.718	0.670
Resolution: 3d **Window**: days 0-5 **Contamination**: 0.3	0.622	**0.535**	0.709	0.699
	...			
Resolution: 1h **Window**: days -14-21 **Contamination**: 0.01	0.506	0.019	**0.992**	0.748
Resolution: 1h **Window**: days -3-0 **Contamination**: 0.01	0.506	0.021	0.990	**0.947**
Resolution: 1h **Window**: days -5-0 **Contamination**: 0.01	0.504	0.019	0.990	0.934
Isolation Forest				
Resolution: 3d, **Window**: days 0-10 **Contamination**: 0.3	**0.617**	0.517	0.7117	0.699
Resolution: 3d **Window**: days 0-5 **Contamination**: 0.3	0.616	**0.523**	0.710	0.698
Resolution: 3d **Window**: days 0-3 **Contamination**: 0.3	0.614	0.520	0.708	0.699
	...			
Resolution: 1h **Window**: days -14-21 **Contamination**: 0.01	0.505	0.018	**0.992**	0.747
Resolution: 6h **Window**: days -5-0 **Contamination**: 0.01	0.505	0.020	0.989	0.933
Resolution: 1h **Window**: days -5-0 **Contamination**: 0.01	0.504	0.018	0.990	**0.934**

guess. Overall, AUCs were better with lower-resolution (more averaged data and fewer data points) data and worse with higher-resolution data. In terms of sensitivity, higher contamination values improved sensitivity by reducing the number of false negatives. Conversely, lower contamination values increased specificity and accuracy by reducing the number of false positives. Ranges of sensitivity and specificity were large.

3.2.1 LAAD

Deep learning utilizing anomaly detector LAAD was used only with Data 1. LAAD algorithm worked successfully for 31/44 participants within Data 1. The rest 13 participants had too little training data because the disease onset happened too early (before day day 26 from the beginning of data collection) or some other issue occurred. As with RHRAD, Isolation Forest, and Night signal, LAAD detector performance deviated remarkably between subjects. Overall sensitivity, specificity and accuracy were 0.331, 0.866 and 0.606, respectively. The same example individuals for visual performance evaluation were used as with other detectors. Subject ASFODQR highlighted a clear anomaly period around the disease onset (see top of Figure 3.8). However, the anomalous period was much longer than determined -3 to 6 days around the disease onset. There were also few false positives along the data time period, but those were not as remarkable as around infectious period. LAAD was able to successfully detect all time points within the infectious period as positives leading to 0 false negatives. Also, there were 22 false positives. Infectious period in Subject A4E0D03 was not detectable using LAAD. LAAD found visually (see bottom of Figure 3.8) three rises in the data but only one of those was in the infection period. Resulting evaluation metrics for subject A4E0D03, there were 46 true positives, 10 false positives, 114 true negatives and 66 false negatives. The example subjects from Data 1 performed very differently from each other. Both of

them had false positives. The good example had no false negatives, whereas the bad example had total of 66 hours of false negatives. The good example had high sensitivity and moderate specificity and the bad example had low sensitivity and high specificity (Figure 3.9).

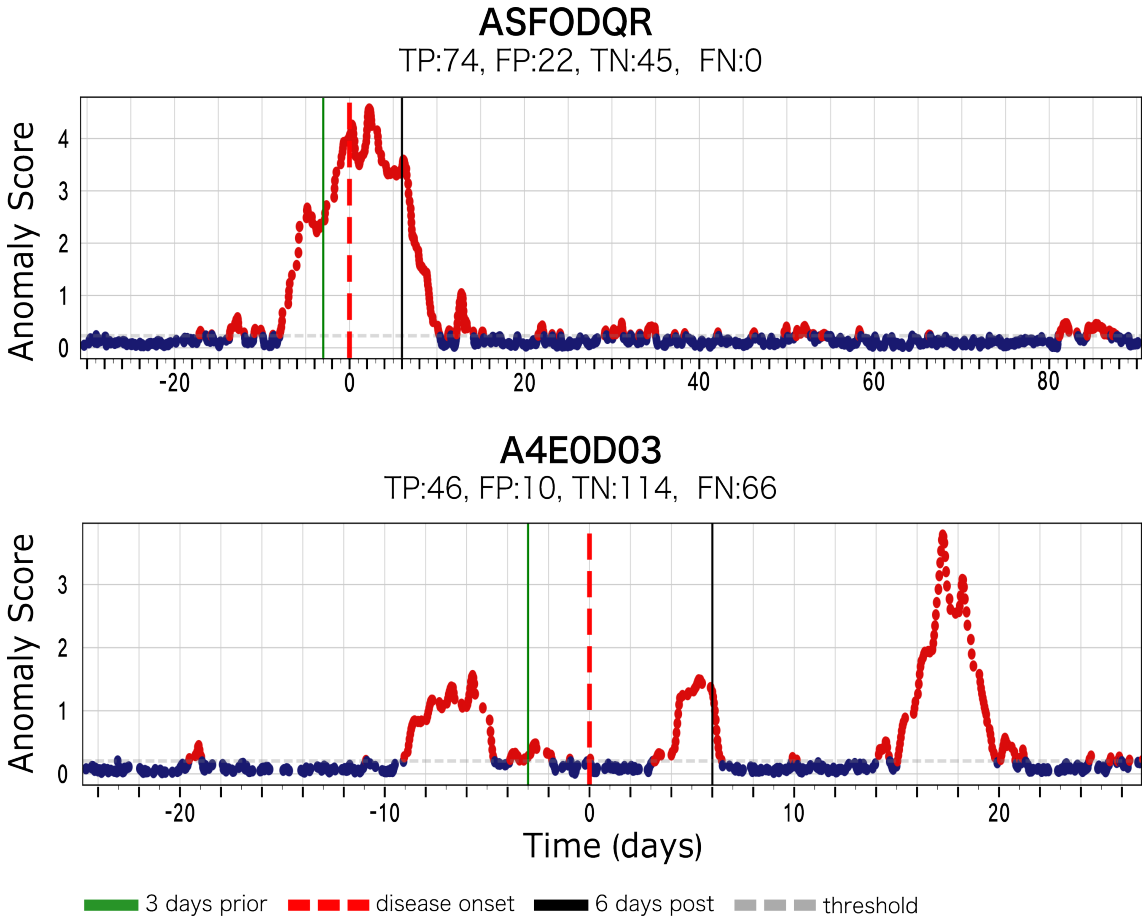

Figure 3.8: **LAAD execution with Data 1.** Same example subjects as in Fig 3.5. Subject ASFODQR had clear anomaly peak around sickness whereas subject A4E0D0E had three different high anomaly score areas but they were not related strictly to disease onset (day 0). True positives (TP) and false negatives (FN) were computed from the period between -3 and 6 days around the disease onset. True negatives (TN) and False positives (FP) were computed from time period -20 to -10 days from disease onset. Training data was all data prior -20 days to disease onset.

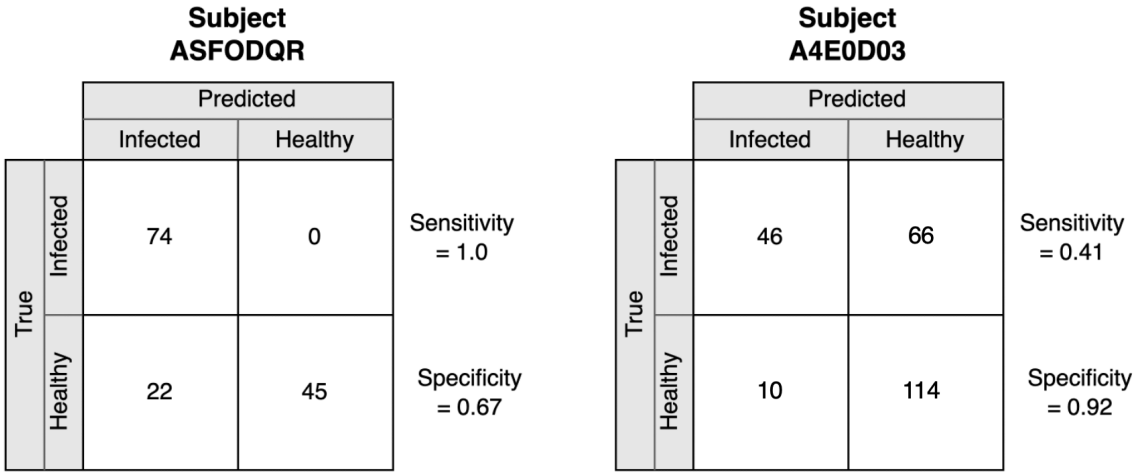

Figure 3.9: **Data 1 LAAD metrics.** LAAD execution comparison as metrics between two example subjects.

3.3 Pilot study evaluation

Our own dataset was collected with Oura rings from three subjects. Two of them collected data over half a year, each having one respiratory infection at the end of the collection. The third subject collected data for three months, and experienced five respiratory infections during that time. RHR and detected anomalies are shown in Figure 3.10. For both subjects 1 and 2, but especially for subject 2, several false positives were detected by all three detectors. Fewer anomalies in total were detected for subject 3, and most of them were during the infectious period. Overall, with parameters of 1-hour data resolution, days 0-5 after symptom onset for detection window and 0.1 contamination, sensitivities were 0.215 and 0.213, specificities 0.908 and 0.908, accuracies 0.864 and 0.864 and AUC 0.561 and 0.560 for RHRAD and Isolation Forest, respectively.

The same (hyper)parameter optimization tests used in previously published datasets were applied also for our own pilot data. The best AUCs for detection were found within a 0-3 day window after symptom onset, using a 3-day data resolution and high contamination values, AUC 0.659 for both detectors. Figure 3.11 shows

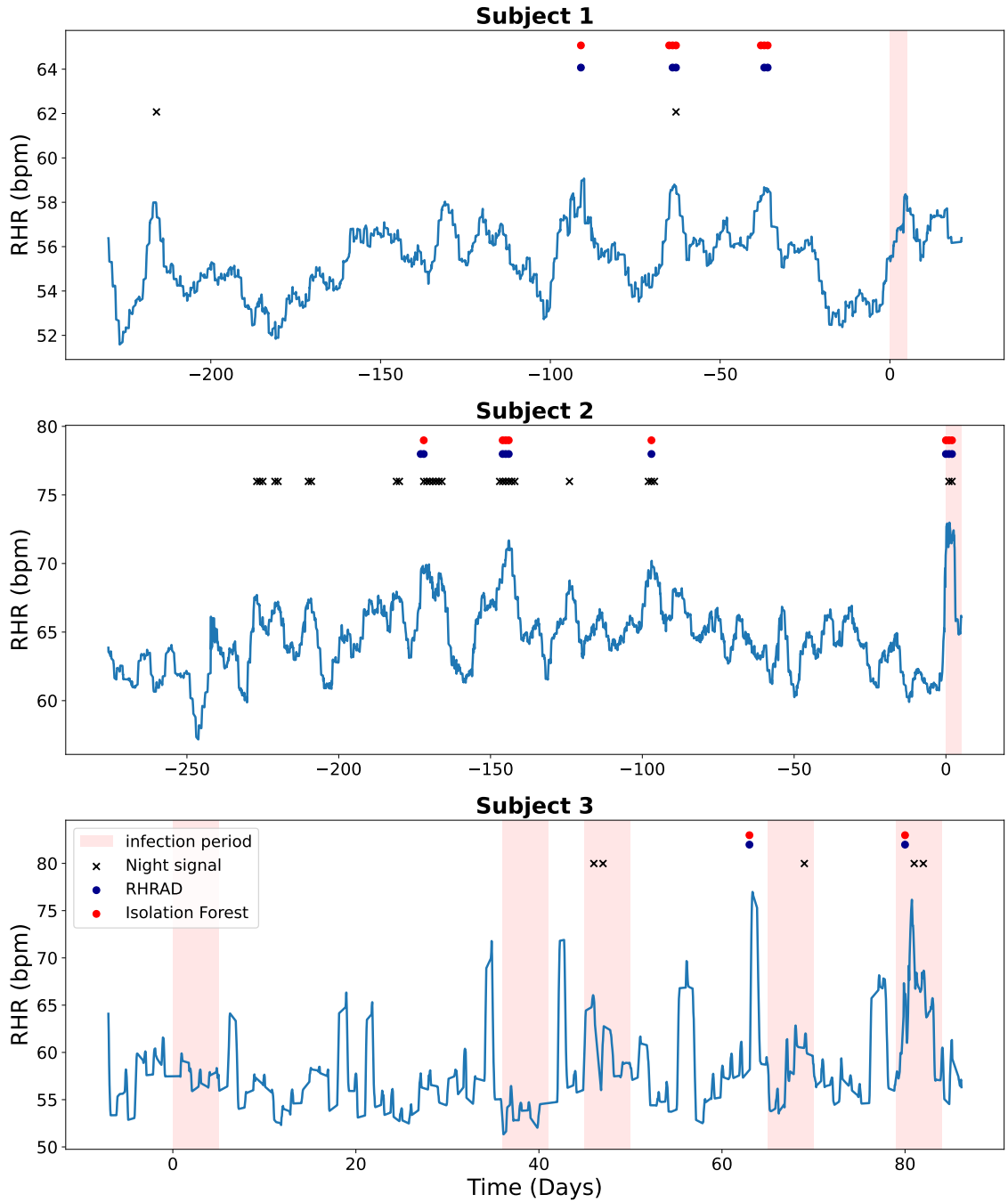

Figure 3.10: **Pilot data.** RHR in hourly resolution for each three subject. Sickness windows in red spans, days 0-5 from symptom onset. For anomaly detectors (RHRAD and Isolation Forest) 0.1 contamination was used.

the RHR data for each subject, with detections made using the 'best' parameters.

In contrast, for RHRAD the worst AUCs were observed with a 3-day data resolu-

tion when the detection window was either only before symptom onset or during the days around symptom onset, AUC 0.487 with days -3 to 7 as detection window and 0.1 contamination and AUC 0.446 with days -3 to 0 as detection window and 0.4 contamination level. For the Isolation Forest, the lowest AUCs were observed with a 1-day resolution and a detection window up to 10 days after symptom onset, AUC 0.471, or from days -3 to 0, using a 3-day data resolution, AUC 0.446 with 0.4 contamination level and AUC 0.445 with 0.3 contamination level. Similar to other datasets, the sensitivity and specificity ranged widely. Table 3.5 presents the three best and three worst parameter combinations by AUC for both detectors.

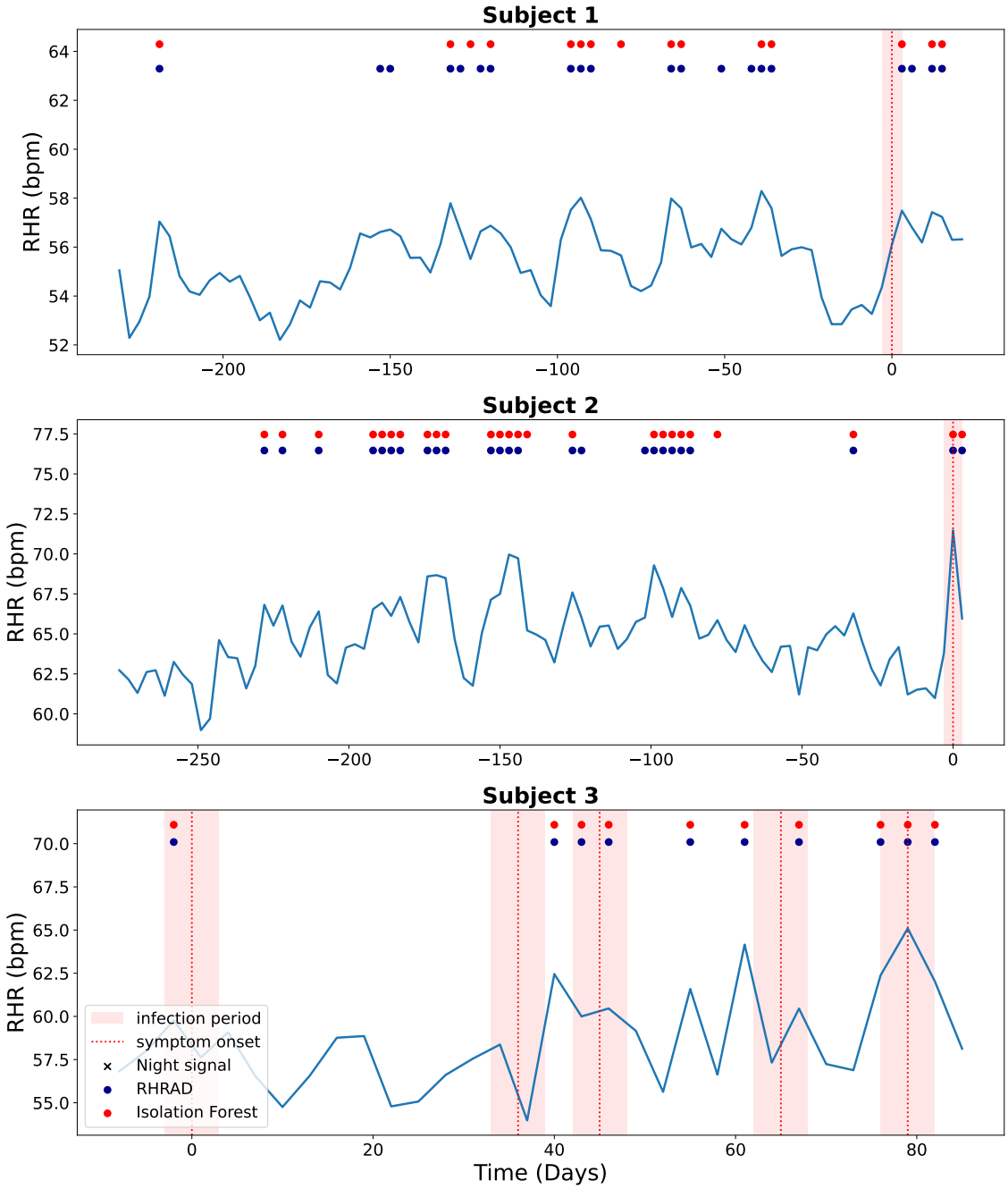

Figure 3.11: **Pilot data with 'optimal' parameters.** Data was averaged to three-day resolution, detection window was days -3 to 3 days around symptom onset and for detections, contamination level 0.5 was used. For subject 3 more true positives were detected than false positives. For subjects 1 & 2, plenty of false positives detected by both detectors but also true positives.

Table 3.5: **Pilot data Detector performance ranked by AUC.** The best values of each column of both detectors are bolded.

Parameters	AUC	Sensitivity	Specificity	Accuracy
RHRAD				
Resolution: 3d **Window**: days 0-3 **Contamination**: 0.5	**0.659**	**0.800**	0.517	0.531
Resolution: 3d **Window**: days 0-3 **Contamination**: 0.4	0.656	0.700	0.612	0.616
Resolution: 3d **Window**: days 0-3 **Contamination**: 0.4	0.655	0.500	**0.811**	**0.796**
	...			
Resolution: 3d **Window**: days -3-7 **Contamination**: 0.1	0.487	0.28	0.693	0.645
Resolution: 3d **Window**: days -5-0 **Contamination**: 0.4	0.476	0.357	0.593	0.578
Resolution: 3d **Window**: days -3-0 **Contamination**: 0.4	0.446	0.300	0.592	0.578
Isolation Forest				
Resolution: 3d, **Window**: days 0-3 **Contamination**: 0.5	**0.659**	**0.800**	0.517	0.531
Resolution: 3d **Window**: days 0-3 **Contamination**: 0.4	0.656	0.700	0.612	0.616
Resolution: 1d **Window**: days 0-3 **Contamination**: 0.2	0.656	0.500	**0.812**	**0.800**
	...			
Resolution: 1d **Window**: days 0-10 **Contamination**: 0.5	0.471	0.454	0.494	0.490
Resolution: 3d **Window**: days -3-0 **Contamination**: 0.4	0.446	0.300	0.592	0.578
Resolution: 3d **Window**: days -3-0 **Contamination**: 0.3	0.445	0.200	0.692	0.668

4 Discussion

The purpose of this thesis was to explore if COVID-19 or other respiratory illness is detectable from user wearable data. Furthermore, the aim was to investigate whether respiratory infection can be detected even before the symptom onset. Three public datasets were found from previously published articles. Datasets 1 & 2 [38, 45] had long-time free-living data from mainly COVID-19 positive individuals but also some cases of influenzas and common colds (e.g. rhinovirus infection). Data 3 [49] was from individuals who experienced COVID-19 related fever. Three different anomaly detectors were used to find the sickness from RHR data: RHRAD [45], Isolation Forest [38] and Night signal [38]. Furthermore, LAAD [39] was tested with Data 1 for further analysis. In addition, the same detectors were tested with pilot our own Oura data to find out if the methods are applicable for other datasets also.

None of the previously published studies, whose data were used in this work, presented statistical comparisons of resting heart rate between different time periods. However, in the original publication of Data 1 [45], step count and sleep duration before and after the symptom onset were compared. In the statistical comparisons of this thesis, temperature was the only parameter that could find differences between baseline and pre-infection period. That may suggest that during the pre-infection period, either no other physiological responses occurred or individual variation was so significant that differences could not be detected without more complex models. All other parameters (RHR, step counts and HRV) showed statistically significant

difference between baseline period and post-detection period (days 0-6 after symptom onset). Based on that, the detectors might be able to detect diseases decently if the detection/infection window was days 0-6 after symptom onset and all other days within -14 to 21 days around the symptom onset were discarded (to reduce false positives caused by the disease).

Confidence intervals of daily RHR were large, indicating that not all individuals experienced deterministic raise in RHR during an infection, and that makes detector performance unreliable. However, for some individuals, anomalous periods were visibly present in the RHR data. This has also a notable effect on the performance of detectors, since detectors work better for those individuals who had visible deviation in their data compared to those individuals whom data was more flat. Additionally, it is worth noting that in general the variances among individuals were more prominent than the differences between detectors.

Major challenge of these type of self-annotated, free-living studies is the quality of the data. They are influenced by the fact that the participants have measured themselves with their own devices (i.e., they have used their own devices as they normally would), which may not follow recommended measurement practices. In Data 1 and Data 2, there were instances where individuals have missing periods (up to tens of days), presenting challenges in the analysis. Additionally, the duration of device wear throughout the day may vary, affecting the computation of RHR. In the data processing, features were averaged multiple times, which results in a distortion of the actual data. In addition, respiratory rate data was included in Data 3, but it was corrupted in multiple subjects (started accumulating), which prevents the use of respiratory rate in generalizing the findings of this study (see appendix Figure A.1).

The overall sensitivity of any used detector and dataset was relatively low, from 8.9% to 24.8% (Tables 3.2 and 3.3). Sensitivity describes the number/ratio of cor-

rectly detected anomalies within the infectious period. When anomaly detection is done with one-hour resolution data, false negatives are understandably raised because it is unrealistic to have raised RHR for ten consecutive days days, even though the person were severely ill. Or, on the other hand, if the whole disease window is marked as anomalous by the detectors, rest of the data period (health) should have very little deviation which would also be ambiguous.

Specificity gives information about false positives, i.e. how well healthy period is detected to be not-anomalous. Specificities of the detectors in different datasets were high, around 90%, meaning that number of false positives identified compared to whole health data period was quite low. High specificities can be explained by the prevalence of healthy period data, which increases the true negative rate. In a long-term detection case like this, specificity typically decreases when sensitivity is increased, as more data points would generally be classified as anomalous, in which case the healthy interval values would also be detected as (false) positives. This pattern was observed across various detector value combinations tested in this thesis, but specificity stayed relatively high, over 70%.

In the initial analysis of detector performance, all anomalies were evaluated on a daily basis. A day was considered anomalous if the average anomaly score was negative or if more than eight hours within that day were classified as anomalous in binary classification. This approach reduces the number of incorrectly classified days, because there had to be several anomalous hours per day, so individual outliers did not cause anomalous classification of that day. Although compressing the detector from hourly to daily resolution may not always be the most practical, there are potential applications. For instance, a detector could signal if the previous day was anomalous compared to past, and then suggest adjustments in activity levels or closer monitoring of one's health condition. However, determining sensitive enough threshold is not easy and more research should be done.

From the parameter combination test, the best parameters performed significantly better compared to the weaker parameters. The best AUCs were achieved using data averaged to 3-day resolution and high contamination values. Using those parameters in real detection cases with small disease windows could lead to overfitting and unrealistic results. Using only low-resolution data, the disease is only detected if data batches fits perfectly, and disease is long enough. This approach also reduces false positives, because short anomalous periods are not shown, also leading to higher AUC.

The detectors used in this study were relatively simple. With more complex detectors, trends and infections could be detected more sensitively, but this would require consideration of common pitfalls in machine learning, such as data leakage to the model prematurely [56]. By taking care of good machine learning practice, with slightly more complex methods utilizing machine learning might be able to detect infections even before symptom onset, if the infection at all causes a physiological response detectable by wearable devices before the onset of symptoms.

RHRAD is based on the Elliptic Envelope, which assumes that the input data is normally distributed. However, when looking at data of each individual, only a few showed Shapiro-Wilk test p-values below 0.05, indicating that their distributions were approximately normal. It is doubtful why the original publication used a detector that assumes normally distributed input data and at the same time the data of a single individual is not normally distributed. This may affect the performance of the detector by degrading it. On the other hand, the results of RHRAD were very similar compared to Isolation Forest, whose operation does not require the data to be normally distributed.

Isolation Forest is, at least theoretically, optimal for this kind of anomaly detection, especially because the ratio of disease periods to healthy periods is small.To have effective Isolation Forest detector, anomalous regions need to deviate signifi-

cantly from healthy data. This suggests that the main challenges in this kind of anomaly detection are more due to the properties of the data and physiological factors, rather than the detector, as the performance metrics for all detectors used in this thesis were relatively similar. Additionally, it is important to consider how Isolation Forest could be adapted to alert abnormalities in real time, as the goal is to identify infections using wearable devices real time.

FSM based detector is somewhat useful for detecting sick or anomalous days from time-period data in real time. However, there are some issues to consider for optimizing the performance of Night signal algorithm. First, comparing single averaged day value to median of all past data might not be the optimal. For example, if a person is sick twice in a relatively short period, the overall median is increased compared to healthy baseline median. For solution, some kind of sliding window approach, for example four to six weeks of data, might be more optimal than the whole past of the data, especially if the amount of data is high. Secondly, it is arguable if using just one value per day (or night) is the best and the most useful way to detect infections. Averaging daily data does not require high-resolution data, making it practical and easy to implement. However, using more densely sampled data, such as one value every six hours, would provide more detailed information about a person's health status and detect variations better. This approach would require considering daily variations and potentially establishing new thresholds or states to account for these changes.

Additionally, using more than one parameter, for example activity levels, respiratory rate or HRV, in the detection algorithm could enhance model performance. In one previous study [52] combining RHR, daily activity, and sleep AUC of 0.72 (95% CI 0.64 - 0.80) with sensitivity 0.36 (0.22-0.50) and specificity 0.95 (0.92-0.97) was achieved using binary classification (not further specified). In comparison, using only daily RHR they got AUC 0.52 (0.41-0.64), sensitivity 0.39 (0.22-0.56) and

specificity 0.80 (0.75-0.86). That would implicate that using model with multiple parameters could improve AUC when detecting infections. However, in that study only COVID-19 was targeted, and test data was only days 0-7 after the symptom onset, i.e. there is no evidence of pre-symptomatic identification with both single and multiple parameters.

With the most reasonable parameters, like hourly sampled data and infection window days -3 to 7 around the symptom onset and relatively low contamination value, AUC was low, but accuracy was still fairly high. It is probably due to heavily imbalanced data; the amount of data within infectious periods had significantly less data than healthy periods for almost all the participants analysed (see Table 2.1 for baseline lengths in days compared to 3 - 10 days of disease period). Thus, looking just the accuracy does not tell the truth about the used anomaly detectors. To solve this problem and improve the performance of the detector, there are a few things that could be tried. First, the data input to the detectors could be processed differently. For example, using moving window instead of the whole data period for detector. Secondly, algorithms (detectors) should be chosen to be suitable for heavily imbalanced data. [57].

In this thesis, the achieved detector sensitivities were insufficient for reliably identifying respiratory infections. As previously mentioned, increasing sensitivity might reduce specificity, so a sufficient balance between the two values must be achieved and evaluated. Sensitivity over 50 % would correctly identify half of the time points during the disease, which could be sufficient with high-resolution data. However, achieving higher sensitivity could increase amount of false positives. High sensitivity is crucial to ensure the user can take necessary actions following the alert. On the other hand, too many false positives, for example a few per week, might lead to alarm fatigue. It is a phenomenon especially in the health care, where the user, like nurse, ignore the alerts due to too many false alerts [58]. That can cause true

positives to be missed, reducing the practical value of the detector. Therefore, the development of the detectors must consider the frequency of predictions and the thresholds for alerts. Practically, it is more beneficial for users to receive notifications before symptoms appear, as they will likely notice illness once symptoms begin, which makes the development process even harder.

About results in publications providing data. In the original publication of Data 1 [45], RHRAD and similarly working HROSAD (heart rate over steps anomaly detector) and RHR-diff, which detects anomalies from 28 day sliding window and standardized residuals, were used for detection. It was shown that 88% of individuals had elevation in RHR around the symptom date, if it was reported. Reported elevation was usually noticed some days prior to the symptom onset. The report highlighted two subjects in whom heart rate elevated clearly before the symptom onset, other started raising 15 and the other four days before symptom onset. It is unclear if the elevation 15 days before symptom onset was due to the infection since the incubation time of COVID-19 is at maximum reported to be 14 days [24]. Difference between COVID-19 and non-COVID-19 participant was not remarkable. Any sensitivity, specificity or other qualitative metrics is not stated. That raises questions about their data quality and detector performance [45].

In the original report of Data 2 [38], more detailed results about detector performance were published compared to Data 1 publication. In the study population, 80% of the infected participants received alerts (detected anomalies) either pre-symptomatically or within asymptomatic COVID-19 infection, and sensitivity was 77% for Fitbit users. The Night signal sensitivity with Data 2 in this thesis was only 20%, which is considerably lower than in the original publication. The difference can be explained by a different approach of defining true positives and false positives. In the original study, true positives were COVID-19 positive participants who received a red alert from the Night signal algorithm within the detection window.

The detection window was time period of 21 days before symptom onset or diagnosis date to the symptom onset or diagnosis date. False negatives were COVID-19 positive subjects who were not receiving any red alert within the same detection period. For COVID-19 negative or untested participants, true negatives were defined as green alerts correctly sent during non-COVID-19 periods, and false positives were red alerts incorrectly sent during these non-COVID-19 periods. For COVID-19 negative and untested participants, the algorithms were tested to get further specificity analysis. Both Night signal and online RHRAD achieved 87.7 % specificity, which is lower than healthy individual specificity tested in this thesis results. Furthermore, based on their analysis of Night signal, the amount of red alerts (anomalies) were the highest around days -2 to 8 relative to the symptom onset, and the peak was on day 3. This fact indicates, that the most significant RHR changes happen just after symptom onset, not so much before that.

According to the analysis of the original study, the sensitivity of online RHRAD with Fitbit data was 69% compared to findings in this thesis, offline RHRAD sensitivity 21.9 % (using just their Fitbit data). In the article there was also speculation of the cause of the false positives by online detectors: alcohol intake, mental stress, holiday season and even large meals. However, the detectors alerted more anomalies during the illness period compared to the normal period. [38].

The third data and report focused on monitoring COVID-19 related fever. It was observed that wearables can effectively detect fever, showing a significant increase in temperature from baseline to the symptom window. Using a single time-point temperature measurement proved appropriate when accounting for inter-individual variation and daily rhythms. The study established thresholds for day and night values and reported that 94% of participants who experienced fever, measured also fever-like days during the symptom window. Additionally, 58% of those who was not reporting actual fever still had fever-like days in the symptom window. Partici-

pants were categorized based on the presence of fever-like days during the detection window, revealing a statistically significant difference in heart rate and heart rate variability (HRV), although no difference was found in self-reported cases. The focus of the study on fever detection did not directly compare to our work. It is important to note for the results of this thesis, that it was not known, which of the 12 participants did not actually report a fever. This might decrease the sensitivity, despite 58% of the participants had fever-like moments. [49].

Physiological responses measured by wearable devices due to infection may not necessarily differ from other physiological special events. Even though detector finds an anomalous event from wearable data, the reason for that could be stress, high-intensity training or measurement error instead of illness.

For example in Norwegian HUNT study it is shown that regular insomnia might slightly elevate RHR levels [59]. In a short time-period exercising activates sympathetic nerve system and increase RHR but in longer period regular exercising can decrease RHR due improvement of physical performance [60]. Also, daily habits like caffeine consumption might temporarily elevate RHR [60] whereas smoking changes heart rate levels more permanently [61]. In addition, alcohol consumption decreases HRV even with small doses, and might have changes to resting heart rate but RHR alterations might not be as sensitive as HRV alterations [62]. There is a limited number of high-quality studies investigating the factors elevating RHR or reducing HRV. However, short-time cold exposure (like cold water swimming), sauna, travelling, high-altitudes, certain times of year (such as holidays) might have an effect to RHR or HRV.

In addition to self-regulated states / habits and respiratory infections, RHR might be elevated by many other diseases and physiological conditions. About physiological conditions, many cardiac diseases, for example arrhythmia, electrolyte disturbance, and diabetic ketoacidosis may alter resting heart rate. In addition to

diseases, many pharmaceuticals like anti-psychotic medicine might affect RHR. [60]

Utilizing location information and possible activity recognition, detection of infections might be improved. Recognized activity and locations may tell the detector whether the user has done something out of ordinary and physiological alterations are caused by that rather than respiratory infection. To achieve automatic detection of infection, more research and labeled data (with all possible extraordinary things labeled) is needed.

5 Conclusions

In general, all AUC values and sensitivities were low for any dataset and detector, meaning these methods are not yet suitable for detection of infections from wearable data. Sicknesses, especially COVID-19, can be to some extent detected after the symptom onset at least for more severe cases. However, existing methods and devices detect physiological alterations related to respiratory infections quite poorly before the onset of symptoms. Results from this thesis are not straight forwardly comparable to previously published results. However, metrics like sensitivity or specificity were not published earlier using data and detector combinations used in this thesis. For our own pilot data, the same methods and detectors worked similarly than for previously published datasets.

In conclusion, with more advanced data and methods, respiratory infections might be detectable even before symptom onset. Additionally, even if detection becomes possible, false positives would still occur, and eliminating them would be challenging. So, there is work to be done for the future.

References

[1] S. Canali, V. Schiaffonati, and A. Aliverti, *PLOS Digital Health*, no. 10, S. Mulvaney, Ed., e0000104, Oct. 13, 2022.

[2] A. Kamišalić, I. Fister, M. Turkanović, and S. Karakatič, "Sensors and Functionalities of Non-Invasive Wrist-Wearable Devices: A Review", vol. 18, no. 6, p. 1714,

[3] R. De Fazio, V. M. Mastronardi, M. De Vittorio, and P. Visconti, "Wearable Sensors and Smart Devices to Monitor Rehabilitation Parameters and Sports Performance: An Overview", *Sensors*, vol. 23, no. 4, p. 1856, Feb. 2023.

[4] D. Duncker *et al.*, "Smart Wearables for Cardiac Monitoring—Real-World Use beyond Atrial Fibrillation", *Sensors*, vol. 21, no. 7, p. 2539, Apr. 2021.

[5] D. K. Ming *et al.*, "Continuous physiological monitoring using wearable technology to inform individual management of infectious diseases, public health and outbreak responses", *International Journal of Infectious Diseases*, vol. 96, pp. 648–654, Jul. 2020.

[6] C. Qin, X. Wang, G. Xu, and X. Ma, "Advances in cuffless continuous blood pressure monitoring technology based on ppg signals", *BioMed Research International*, vol. 2022, no. 1, p. 8 094 351, 2022.

[7] X. Quan *et al.*, "Advances in non-invasive blood pressure monitoring", *Sensors*, vol. 21, no. 13, p. 4273, 2021.

[8] T. Tamura, "Cuffless blood pressure monitors: Principles, standards and approval for medical use", *IEICE Transactions on Communications*, vol. 104, no. 6, pp. 580–586, 2021.

[9] J. Zanetti and D. Salerno, "Seismocardiography: A technique for recording precordial acceleration", in *[1991] Computer-Based Medical Systems@m_ Proceedings of the Fourth Annual IEEE Symposium*, IEEE Comput. Soc. Press, pp. 4–9.

[10] J. Dunn, R. Runge, and M. Snyder, "Wearables and the medical revolution", *Personalized Medicine*, vol. 15, no. 5, pp. 429–448, Sep. 2018.

[11] B. Bent, B. Goldstein, W. Kibbe, and J. Dunn, "Investigating sources of inaccuracy in wearable optical heart rate sensors. npj digital medicine, 3 (1), 18", *Collins, T., Woolley, SI, Oniani, S., Pires, IM, Garcia, NM, Ledger, SJ, & Pandyan, A.(2019). Version reporting and assessment approaches for new and updated activity and heart rate monitors. Sensors*, vol. 19, no. 7, p. 1705, 2020.

[12] R. K. Reddy *et al.*, "Accuracy of wrist-worn activity monitors during common daily physical activities and types of structured exercise: Evaluation study", *JMIR mHealth and uHealth*, vol. 6, no. 12, e10338, 2018.

[13] M. Orini, G. Guvensen, A. Jamieson, N. Chaturvedi, and A. D. Hughes, "Movement, sweating, and contact pressure as sources of heart rate inaccuracy in wearable devices", in *2022 Computing in Cardiology (CinC)*, IEEE, vol. 498, 2022, pp. 1–4.

[14] R. Eccles, "Understanding the symptoms of the common cold and influenza", *The Lancet Infectious Diseases*, vol. 5, no. 11, pp. 718–725, Nov. 2005.

[15] B. Sundaresan, F. Shirafkan, K. Ripperger, and K. Rattay, "The Role of Viral Infections in the Onset of Autoimmune Diseases", *Viruses*, vol. 15, no. 3, p. 782, Mar. 18, 2023.

[16] T. Heikkinen and A. Järvinen, "The common cold", *The Lancet*, vol. 361, no. 9351, pp. 51–59, Jan. 2003, ISSN: 01406736.

[17] N. Jain, R. Lodha, and S. Kabra, "Upper respiratory tract infections", *The Indian Journal of Pediatrics*, vol. 68, pp. 1135–1138, 2001.

[18] H. M. Ollila *et al.*, "Face masks to prevent transmission of respiratory infections: Systematic review and meta-analysis of randomized controlled trials on face mask use", vol. 17, no. 12, e0271517,

[19] S. F. Bloomfield, A. E. Aiello, B. Cookson, C. O'Boyle, and E. L. Larson, "The effectiveness of hand hygiene procedures in reducing the risks of infections in home and community settings including handwashing and alcohol-based hand sanitizers", vol. 35, no. 10, S27–S64,

[20] World Health Organization. "COVID-19 deaths | WHO COVID-19 dashboard", datadot. (2023), [Online]. Available: `http://data.who.int/dashboards/covid19/cases` (visited on 05/07/2024).

[21] C. Huang *et al.*, "Clinical features of patients infected with 2019 novel coronavirus in wuhan, china", *The lancet*, vol. 395, no. 10223, pp. 497–506, 2020.

[22] C. Dye, *The benefits of large scale covid-19 vaccination*, 2022.

[23] Y.-Z. Huang and C.-C. Kuan, "Vaccination to reduce severe covid-19 and mortality in covid-19 patients: A systematic review and meta-analysis.", *European Review for Medical & Pharmacological Sciences*, vol. 26, no. 5, 2022.

[24] S. A. Lauer *et al.*, "The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application", *Annals of Internal Medicine*, vol. 172, no. 9, pp. 577–582, May 5, 2020, ISSN: 0003-4819, 1539-3704.

[25] WHO / Khaled Mostafa. "Influenza (Seasonal)". (), [Online]. Available: `https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal)` (visited on 04/12/2024).

[26] L. J. Keilman, "Seasonal Influenza (Flu)", *Nursing Clinics of North America*, vol. 54, no. 2, pp. 227–243, Jun. 2019.

[27] S. P. Whelton *et al.*, "Association Between Resting Heart Rate and Inflammatory Biomarkers (High-Sensitivity C-Reactive Protein, Interleukin-6, and Fibrinogen) (from the Multi-Ethnic Study of Atherosclerosis)", *The American Journal of Cardiology*, vol. 113, no. 4, pp. 644–649, Feb. 2014.

[28] J. Karjalainen, "Fever and Cardiac Rhythm", *Archives of Internal Medicine*, vol. 146, no. 6, p. 1169, Jun. 1, 1986.

[29] M. M. Jensen, J. G. Kellett, P. Hallas, and M. Brabrand, "Fever increases heart rate and respiratory rate; a prospective observational study of acutely admitted medical patients", *Acute Med*, vol. 18, no. 3, pp. 141–143, 2019.

[30] C. A. Sanches, G. A. Silva, A. F. H. Librantz, L. M. M. Sampaio, and P. A. Belan, "Wearable devices to diagnose and monitor the progression of covid-19 through heart rate variability measurement: Systematic review and meta-analysis", *Journal of Medical Internet Research*, vol. 25, e47112, 2023.

[31] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey", *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[32] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection", *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 6, no. 1, pp. 1–39, 2012.

[33] W. S. Al Farizi, I. Hidayah, and M. N. Rizal, "Isolation forest based anomaly detection: A systematic literature review", in *2021 8th International Confer-*

ence on Information Technology, Computer and Electrical Engineering (ICITACEE), IEEE, 2021, pp. 118–122.

[34] S. Shriram and E. Sivasankar, "Anomaly detection on shuttle data using unsupervised learning techniques", in *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, IEEE, 2019, pp. 221–225.

[35] P. Rousseeuw and K. Driessen, "A fast algorithm for the minimum covariance determinant estimator", *Technometrics*, vol. 41, pp. 212–223, Aug. 1999.

[36] W. Zhang, Q. Yang, and Y. Geng, "A survey of anomaly detection methods in networks", in *2009 International Symposium on Computer Network and Multimedia Technology*, IEEE, 2009, pp. 1–3.

[37] A. B. Nassif, M. A. Talib, Q. Nasir, and F. M. Dakalbab, "Machine learning for anomaly detection: A systematic review", *Ieee Access*, vol. 9, pp. 78 658–78 700, 2021.

[38] A. Alavi *et al.*, "Real-time alerting system for COVID-19 and other stress events using wearable data", *Nature Medicine*, vol. 28, no. 1, pp. 175–184, Jan. 2022.

[39] G. K. Bogu and M. P. Snyder, "Deep learning-based detection of COVID-19 using wearables data", Infectious Diseases (except HIV/AIDS), preprint, Jan. 9, 2021.

[40] J. L. Cleary, Y. Fang, S. Sen, and Z. Wu, "A caveat to using wearable sensor data for COVID-19 detection: The role of behavioral change after receipt of test results", *PLOS ONE*, vol. 17, no. 12, A. Borri, Ed., e0277350, Dec. 30, 2022.

[41] R. P. Hirten *et al.*, "Use of Physiological Data From a Wearable Device to Identify SARS-CoV-2 Infection and Symptoms and Predict COVID-19 Diagnosis: Observational Study", *Journal of Medical Internet Research*, vol. 23, no. 2, e26107, Feb. 22, 2021.

[42] N. Marinsek *et al.*, "Measuring COVID-19 and Influenza in the Real World via Person-Generated Health Data", Infectious Diseases (except HIV/AIDS), preprint, May 30, 2020.

[43] A. E. Mason *et al.*, "Detection of COVID-19 using multimodal data from a wearable device: Results from the first TemPredict Study", *Scientific Reports*, vol. 12, no. 1, p. 3463, Mar. 2, 2022.

[44] D. J. Miller *et al.*, "Analyzing changes in respiratory rate to predict the risk of COVID-19 infection", *PLOS ONE*, vol. 15, no. 12, E. P. Scilingo, Ed., e0243693, Dec. 10, 2020.

[45] T. Mishra *et al.*, "Pre-symptomatic detection of COVID-19 from smartwatch data", *Nature Biomedical Engineering*, vol. 4, no. 12, pp. 1208–1220, Nov. 18, 2020.

[46] A. Natarajan, H.-W. Su, and C. Heneghan, "Assessment of physiological signs associated with COVID-19 measured using wearable devices", *npj Digital Medicine*, vol. 3, no. 1, p. 156, Nov. 30, 2020.

[47] M. M. H. Shandhi *et al.*, "A method for intelligent allocation of diagnostic testing by leveraging data from commercial wearable devices: A case study on COVID-19", *npj Digital Medicine*, vol. 5, no. 1, p. 130, Sep. 1, 2022.

[48] A. Shapiro *et al.*, "Characterizing COVID-19 and Influenza Illnesses in the Real World via Person-Generated Health Data", *Patterns*, vol. 2, no. 1, p. 100 188, Jan. 2021.

[49] B. L. Smarr *et al.*, "Feasibility of continuous fever monitoring using wearable devices", *Scientific Reports*, vol. 10, no. 1, p. 21 640, Dec. 14, 2020.

[50] E. Grzesiak *et al.*, "Assessment of the Feasibility of Using Noninvasive Wearable Biometric Monitoring Sensors to Detect Influenza and the Common Cold Before Symptom Onset", *JAMA Network Open*, vol. 4, no. 9, e2128534, Sep. 29, 2021.

[51] D. S. Temple *et al.*, "Wearable Sensor-Based Detection of Influenza in Presymptomatic and Asymptomatic Individuals", *The Journal of Infectious Diseases*, vol. 227, no. 7, pp. 864–872, Apr. 12, 2023.

[52] G. Quer *et al.*, "Wearable sensor data and self-reported symptoms for COVID-19 detection", *Nature Medicine*, vol. 27, no. 1, pp. 73–77, Jan. 2021, ISSN: 1078-8956, 1546-170X. DOI: 10.1038/s41591-020-1123-x.

[53] M. Gadaleta *et al.*, "Passive detection of COVID-19 with wearable sensors and explainable machine learning algorithms", *npj Digital Medicine*, vol. 4, no. 1, p. 166, Dec. 8, 2021.

[54] A. J. Kucharski *et al.*, "Effectiveness of isolation, testing, contact tracing, and physical distancing on reducing transmission of SARS-CoV-2 in different settings: A mathematical modelling study", vol. 20, no. 10, pp. 1151–1160,

[55] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[56] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science", vol. 4, no. 9, p. 100 804,

[57] V. S. Spelmen and R. Porkodi, "A review on handling imbalanced data", in *2018 international conference on current trends towards converging technologies (ICCTCT)*, IEEE, 2018, pp. 1–11.

[58] J. P. Keller Jr, "Clinical alarm hazards: A "top ten" health technology safety concern", *Journal of electrocardiology*, vol. 45, no. 6, pp. 588–591, 2012.

[59] M. Hauan, L. B. Strand, and L. E. Laugsand, "Associations of insomnia symptoms with blood pressure and resting heart rate: The hunt study in norway", *Behavioral sleep medicine*, vol. 16, no. 5, pp. 504–522, 2018.

[60] B. Olshansky, F. Ricci, and A. Fedorowski, "Importance of resting heart rate", *Trends in Cardiovascular Medicine*, 2022.

[61] G. Papathanasiou *et al.*, "Effects of smoking on heart rate at rest and during exercise, and on heart rate recovery, in young adults", *Hellenic J Cardiol*, vol. 54, no. 3, pp. 168–77, 2013.

[62] M. Romanowicz, J. E. Schmidt, J. M. Bostwick, D. A. Mrazek, and V. M. Karpyak, "Changes in heart rate variability associated with acute alcohol consumption: Current knowledge and implications for practice and research", *Alcoholism: Clinical and Experimental Research*, vol. 35, no. 6, pp. 1092–1105, 2011.

Appendix A Data 3 respiration rate data quality

Surprisingly many subjects in Data 3 had problems with respiration rate data quality (see Figure A.1). For some individuals, RR data started to accumulate at the beginning and then normalizes, while for some others, RR might have been normal at first and then started to increase or decrease abnormally. One subject (Subject 39) had strange period in the middle of the data period and in addition to RR, other parameters had strange behaviour too. One other subject's RR did not increasingly or decreasingly accumulate but remains weirdly consistent for over 20 days. All these issues could be due to the original data from the device being linearly interpolated to minute-level accuracy, with some missing periods or other issues. Such data cannot be reliably used for any analysis.

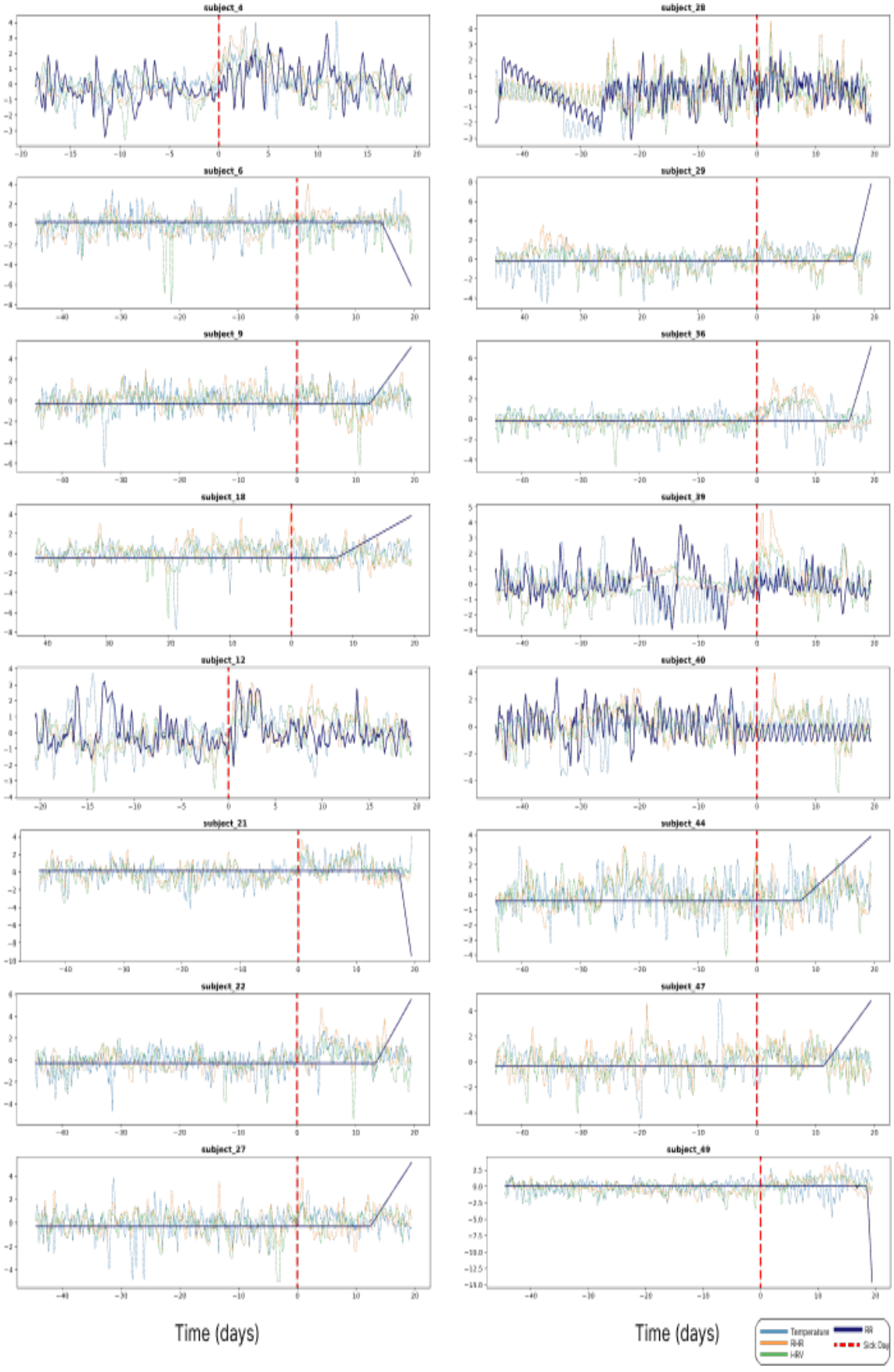

Figure A.1: **Problems in Data 3 RR.** Subjects 4 and 12 are examples of normal behavior of RR data and all of others have remarkable problems.