Analysis of the Association Between COPD Patients' Health and RESpeck Data and Comparison of Manually Measured Breathing Rates and Breathing Rates Measured by the RESpeck

Merle Pauline Peters

4th Year Project Report
Computer Science and Mathematics
School of Informatics
University of Edinburgh

2023

Abstract

This project analyses whether the data collected by the respiratory monitoring device, RESpeck, can be used for estimating the health status of COPD patients and how reliable and informative the breathing rates measured by the RESpeck are in comparison to breathing rates measured by nurses. The health status of COPD patients is recorded as CAT scores. A prediction model using RESpeck and associated Rehab data from pulmonary rehabilitation exercises achieves an MAE of 3.3, an RMSE of 4.5 and an accuracy of 73% for predicting the CAT score of the next day. An associative model achieves an MAE of 3.0, an RMSE of 4.1 and an accuracy of 75% for estimating the CAT score on the same day as when the RESpeck and Rehab data were collected from. Statistical analysis of the hourly averages of the breathing rates measured by the RESpeck and nurses showed that RESpeck breathing rates were on average 0.9 bpm higher than the nurse breathing rates. An analysis of the information content and uncertainty coefficient indicates that the RESpeck breathing rates are more informative (2.75 nats in comparison to 2.34 nats) and that RESpeck breathing rates reduce uncertainty about the nurse breathing rates more than the nurse breathing rates reduce uncertainty about the RESpeck breathing rates.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

The first dataset used in this project was collected as part of an Innovation Project with NHS Borders (Study Reference Number: 21/BORD/IN01) entitled ”Remote Monitoring and Pulmonary Rehabilitation of COPD (and COVID-19 recovered) Patients in the NHS Borders Region”, and the study approval letter is included in Appendix A.

The WGH-DCN dataset was collected as part of a Quality Improvement Programme (QIP) study, and Caldecott approval was provided by Lothian Health Authority (PI - Dr Paul Brennan, Centre for Clinical Brain Sciences, University of Edinburgh).

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Merle Pauline Peters*)

Acknowledgements

Many thanks to my supervisor D. K. Arvind and the time and advice he gave me during regular meetings. For their technical help, I thank Sharan Maiya and Andrew Bates. I am also thankful for the knowledge and expertise provided by Mark Prior and Gordon Drummond with respect to the NHS Borders analysis and QIP analysis respectively. And finally, I thank my family and friends for their emotional and academic support.

Contents

Chapter 1

Introduction

Monitoring a patient's health is an important part of assessing how a patient responds to treatment and whether a disease becomes more severe and dangerous. The RESpeck device and the system of accompanying apps are designed as a real-time alarm system that monitors the development of a patient's health status and can be used to inform patients and their physicians when the patient's health worsens or to review changes in the patient's health at an appointment.

As part of this project, data from two studies was analysed. The overall aim was to evaluate the descriptiveness of the data collected from the RESpeck and accompanying apps.

The first study collected data from COPD patients. COPD is a prevalent pulmonary disease that can cause life-threatening exacerbations, so close monitoring of patients' health is important. During the study, the patients recorded CAT scores which give an indication of their health, performed pulmonary rehabilitation (PR) exercises which were recorded by one of the apps accompanying the RESpeck, and wore the RESpeck.

The hypothesis is that data collected by the RESpeck device and by recording PR exercises during a specific day can be used to estimate a patient's CAT score for that day and to predict the score for the following day. To test this hypothesis, various features of the given data and their correlation with the CAT score are examined, and association and prediction models are trained and evaluated for data from the same day and the day preceding the CAT score respectively.

In the second study, the breathing rate of post-operative patients was measured by the RESpeck and nurses. The breathing rate is an essential indicator of a patient's health status, nonetheless, nurses often take measurements during very short time periods [1] or not at all [2]. The RESpeck would provide a device that consistently measures the breathing rate without requiring much of the nurses' time. The aim of the analysis of this dataset is to investigate the differences between the automatic RESpeck measurements and the manual measurements by nurses. The hypothesis is that breathing rates collected by the RESpeck are more reliable and informative than the breathing rates measured by the nurses. To that end, the RESpeck and nurse breathing rates are analysed using methods from statistics and information theory.

1.1 Contributions

The main contributions of this project are

- a detailed analysis of the correlations of CAT scores with the breathing and exercise data as recorded by the RESpeck and accompanying apps,
- the development of machine learning models that can estimate the CAT score for a specific day given breathing and exercise data from the same day and given data from the previous data,
- experiments on patient clustering according to how patients usually choose CAT scores in an attempt to improve machine learning models,
- a comparison of breathing rates measured by the RESpeck and breathing rates measured by nurses using hypotheses testing, information content and uncertainty coefficients.

1.2 Report outline

The report is divided into five chapters. The remaining chapters are structured as follows

- **Chapter 2** introduces related literature concerning COPD, respiratory monitoring, identifying reliable RESpeck data and predicting the CAT score using RESpeck data.
- **Chapter 3** concerns the analysis of the NHS Borders data. It explores the data, describes how it was preprocessed, what features were considered and how they correlate with the CAT score, and evaluates how models based on these features perform at estimating the CAT score of the same day and predicting it for the next day.
- **Chapter 4** concerns the analysis of the QIP data. The data exploration, identification of reliable data and comparison of RESpeck and nurse data are described.
- **Chapter 5** summarises the results of the analyses and provides suggestions for future works.

Chapter 2

Background

This chapter presents background knowledge for understanding the main contents of this report. First, COPD is introduced as well as ways to measure COPD and what pulmonary rehabilitation is. Next, the respiratory monitoring device, RESpeck, is described and finally previous work related to the analysis of the NHS Borders dataset is presented. There is no previous work related to the analysis of the QIP dataset.

2.1 Chronic obstructive pulmonary disease (COPD)

Chronic obstructive pulmonary disease (COPD) is a respiratory disease. It is "a heterogeneous lung condition [...] due to abnormalities of the airways [...] and/or alveoli [...] that cause persistent, often progressive, airflow obstruction."[3] Common symptoms are dyspnea (shortness of breath), chronic coughs with or without phlegm, frequent chest infections and persistent wheezing [3, 4, 5].

The Global Initiative for Chronic Obstructive Lung Disease (GOLD) classifies COPD as a preventable and treatable disease. COPD is mainly caused by long-term exposure to noxious particles or gases due to tobacco smoking, indoor and outdoor air pollution and other environmental factors [3, 5]. Possible treatments include smoking cessation, pharmacological therapy, vaccines to reduce the risk of infections, and pulmonary rehabilitation exercises (see 2.1.3).

Even though COPD is preventable and treatable, it is one of the main causes of death. The World Health Organization has predicted COPD to be responsible for 8.6% of deaths by 2030 and hence to become the third leading cause of death after ischaemic heart disease and cerebrovascular disease [6]. The GOLD estimates that every year about three million people die due to COPD (6% of all deaths globally) and that it will be over 5.4 million by 2060 [3]. This rise in deaths due to COPD is expected because of an aging population and continued or increased exposure to COPD risk factors such as tobacco smoking [3, 6].

2.1.1 Measuring well-being of COPD patients

To diagnose COPD, patients are tested using spirometry [5]. This test measures the volume of air that a patient can breathe out with one forced breath [7]. The outcomes of spirometric assessments are not only used for diagnosing COPD but are also commonly used to guide decisions about COPD management and treatment [8]. However, the correlation between spirometry results and a patient's health status is considered too low so the GOLD recommends the use of formal symptom assessments such as the COPD assessment test (CAT) or the COPD Questionnaire (CCQ) [5].

The IPCRG Users' Guide to COPD "Wellness" Tools [9] assesses the quality of nine out of over 42 tools with respect to the validity/reliability, responsiveness, applicability to primary care population, ease of administration, applicability in practice and the support of other languages. It includes tools that "measure the health status or quality of life as well as tools that measure COPD features such as dyspnoea and breathing problems". Overall, the CCQ is ranked best followed by the Chronic Respiratory Disease Questionnaire (CRQ), Medical Research Council - Dyspnoea (MRC-D) and CAT.

2.1.2 COPD assessment test (CAT)

The NHS Borders dataset includes CAT data to document the patients' health status and Quality of Life. The CAT is a questionnaire that consists of eight questions. For each question the patient can give a score between 0 and 5. The questions are [10]:

1. I never cough. / I cough all the time.
2. I have no phlegm (mucus) in my chest at all. / My chest is completely full of phlegm (mucus).
3. My chest does not feel tight at all. / My chest feels very tight.
4. When I walk up a hill or one flight of stairs I am not breathless. / When I walk up a hill or one flight of stairs I am very breathless.
5. I am not limited doing my activities at home. / I am very limited doing activities at home.
6. I am confident leaving my home despite my lung condition. / I am not at all confident leaving my home because of my lung condition.
7. I sleep soundly. / I don't sleep soundly because of my lung condition.
8. I have lots of energy. / I have no energy at all.

The scores for the individual questions are summed up to give the overall CAT score for a patient. Hence, CAT scores range from 0 to 40. They are classified into four impact levels [11]. Based on the correlation between the CAT and the St George's Respiratory Questionnaire for COPD (SGRQ-C), items of the SGRQ-C were mapped to the CAT impact levels for better interpretation of the CAT scores (see Table 2.1 for a representative selection) [12].

Impact level	Description
low impact (0-9)	breathless several days a week, breathless walking up hills, most days are good in average week
medium impact (10-20)	breathless most days of the week, housework takes long or stop for rests, wheezing attacks a few days a month, a few good days in average week
high impact (21-30)	breathless walking on level ground, cough and/or breathing disturbs sleep, stops patient doing most things they want to do, ≥ 3 attacks of chest trouble in last year, no good days in average week
very high impact (31-40)	cannot move far from bed or chair, stops patient doing everything they want to do

Table 2.1: SGRQ-C items mapped to CAT impact levels [11, 12]

In a systematic review of the CAT in 2014 [13], studies assessing the CAT were considered with regards to the reliability, validity, responsiveness and minimum clinically important difference (MCID) of the CAT. Several studies showed that the CAT score stays consistent when evaluated under similar conditions. The studies further indicate that the CAT can reliably differentiate between healthy individuals and individuals diagnosed with COPD, infrequent exacerbators and frequent exacerbators, and exacerbation states and stable states. Overall, the CAT has been shown to provide a "reliable measurement of health status and [to be] responsive to change with treatment and exacerbations" [11].

The MCID is "the smallest difference in score in the domain of interest which patients perceive as beneficial" [14]. In the systematic review of the CAT in 2014 [13], four papers that tried to calculate the MCID for the CAT were reviewed. The MCID is concluded to be debatable as the individual studies get varying MCID values of 2 or up to almost 4 units. In 2018, the same four papers and an additional two papers were considered in a systematic review on clinically relevant differences in COPD health status [15]. Using triangulation, the MCID is estimated to be -2.54 for improvement and 1-2 for deterioration. However, in the individual studies that were considered the MCID for improvement varied between -1 and -4 for individual studies and there were only two studies that considered the MCID for deterioration.

2.1.3 Pulmonary rehabilitation

The American Thoracic Society (ATS) and European Respiratory Society (ERS) define pulmonary rehabilitation (PR) as "a comprehensive intervention based on a thorough patient assessment followed by patient-tailored therapies, which include, but are not limited to, exercise training, education, and behavior change, designed to improve the physical and psychological condition of people with chronic respiratory disesase and to promote the long-term adherence of health-enhancing behaviors." [16] PR programs usually last six to eighth weeks [3]. A key element of PR is exercise training which consists of endurance and interval training, resistance/strength training, upper and lower limb exercises, walking, flexibility exercises, inspiratory muscle training and

neuromuscular electrical stimulation [16, 3]. Its primary effect is an improvement in muscle function in COPD patients but also an increase in motivation to exercise in general, a reduction of mood disturbances such as anxiety and depression, a decrease of the symptom burden and improvements of cardiovascular functions [16, 3].

Despite several guidelines recommending the use of PR [3, 17, 18, 19] for COPD patients, environmental barriers, lack of awareness and misleading beliefs of the consequences of PR keep many eligible patients from attending PR programs [20]. Long distances to the next PR centre present a key barrier to patients because they have to spend long times travelling, they are often not able to travel independently and the access to public transportation and car parking is often restricted [21].

As an alternative to PR in centres, home-based PR might help circumvent these barriers. A review [22] of eighth studies comparing centre and home-based PR found no statistically significant differences in the exercise capacity and health-related quality of life. So home-based PR offers an alternative that is as effective as centre-based PR. Three of the reviewed studies reported the proportion of attended PR sessions. While the proportion of attended sessions for centre-based PR varied between 49% and 93%, it varied between 73% and 98% for home-based PR. This suggests that home-based PR might improve the attendance of PR programs. However, further research needs to be done.

2.2 Respiratory monitoring

There are several ways to monitor a patient's breathing. Traditional techniques are manual counting, spirometry, capnometry and pneumography. Manual counting is usually done by nurses but it is time-consuming and often inaccurate as measurements depend on monitoring periods and the patient's awareness [1][23]. In general, the traditional methods "often require cumbersome and expensive devices that may interfere with natural breathing" [24].

An alternative are wearable physiological monitoring systems that continuously collect data. Examples for these devices are RESpeck, RespiraSense and Sensecho. While the RESpeck and RespiraSense are small devices that are attached to the lower part of the chest, the Sensecho is a multi-sensor vest [25, 26]. The following section will consider the RESpeck in more detail.

2.2.1 RESpeck device

The RESpeck device is an accelerometer sensor that is worn on the left-side of the chest, just below the last rip. It is about 4 cm $\times$ 3.5 cm $\times$ 1 cm in size. Together with a group of mobile apps, the RESpeck is part of a home-based PR system. The Pairing App connects the RESpeck device with the phone, the AirRespeck App is used to access the sensor data from the RESpeck, the Rehab App is used to guide the patient through exercises and to record data related to the exercises, and finally the user can fill out the COPD assessment test (CAT) in the Diary App. A detailed description of the system

can be found in [27]. The CAT and PR were described in Sections 2.1.2 and 2.1.3 respectively.

2.2.2 Identifying reliable RESpeck data

The RESpeck measures a patient's breathing based on the movement of the chest. However, movement of the entire body and other environmental movements are detected by the RESpeck as well and might obscure the movement caused by the patient's breathing. Comparing the respiratory rate as measured by the RESpeck with the respiratory rate as measured by the nasal cannula, the model presented in [28] was trained to select reliable periods in the RESpeck data.

A disadvantage of taking nasual cannula data as the gold standard is that the nasual cannula itself is not infallible. As [28] mentions, some of the data from the nasual cannula was affected by the patient speaking, coughing or breathing through the mouth, or by displacement and disconnection of the cannula.

Testing the model on an unseen dataset resulted in about 50% of the RESpeck data being classified as reliable. For a particular patient, only 10% of the data was considered reliable. However, taking into account that 10% of an hour are 6 minutes, this is still much more time than what nurses usually have available for measuring the breathing rate of a single patient.

2.3 Previous work on predicting the CAT score using RESpeck data

The goal for the first dataset is to assess a patient's well-being based on RESpeck data. Considering that the CAT provides a reliable indication of COPD patients' health status (see 2.1.2), predicting CAT scores using RESpeck data is a significant step.

In 2016, Darius Fischer [29] built a first model to predict CAT scores based on the RESpeck data that was recorded while patients performed PR exercises. The results of his work were published in a summarising paper in 2022 [27]. The model is based on data collected from 31 COPD patients over 4.5 months. A model that always predicts the mean CAT score of all seen CAT scores is used as a baseline (MAE: 11.3, RMSE: 12.0). Features based on the activity level, breathing signal and breathing rate during breaks and exercises are considered as well as features describing the recovery after an exercise such as the length of rest periods. Because these features are based on the patient performing exercises, only CAT scores that were preceded by a valid exercise block within the last 48 hours were included in the analysis. An exercise block is considered valid if it consists of ten exercises and rest periods that are only related to one CAT score and include no NaN values. Hence, effectively only the data of 20 patients was investigated, ignoring 11 patients with invalid or missing exercise blocks.

The expressiveness of each feature was evaluated by calculating its correlation with the CAT score. Linear regression models based on individual features that were considered to be predictive of the CAT score achieved a model performance of up to 5.4 for the

MAE and 6.6 for the RMSE (corresponding to the *mean angles during rest* feature). Further models such as Ordinary Least Squares (OLS), Lasso, Ridge and Elastic Net regression, Artificial Neural Networks (ANN) and neural network ensembles were considered. An ensemble of six two-hidden-layer neural networks performed best (MAE: 2.8, RMSE: 4.6).

In 2022, Ioana Mihăilescu [30] revisited the problem of predicting CAT scores. An earlier version of the NHS Borders dataset which will be considered in this report was investigated there. The dataset included data from eleven COPD patients. However, two patients had no Diary data so they were excluded from the analysis. Overall, the data corresponding to 67 CAT scores was used.

The breathing rate during resting periods after some exercises was shown to have a high correlation with the CAT scores. As there were only few data points available for each exercise type, a combination of breathing rates during resting periods for all exercise types was used. A linear regression model based on the *breathing rate during rest* was trained and achieved an MAE of 5.9 and an RMSE of 6.8. These errors are comparable to the error values obtained for the linear regression model using only the *mean angles during rest* feature in [29] (MAE: 5.4, RMSE: 6.6). It was not possible to build a more complex model because of the limited amount of data.

Chapter 3

NHS Borders data

The NHS Borders dataset contains data recorded for COPD patients over several weeks. It contains RESpeck, Rehab and Diary data. The Diary data describes a patient's wellbeing as assessed by the patient themself. The goal is to find an association between the RESpeck and Rehab data with the Diary data which could eventually be used to estimate the Diary data on the basis of the other two.

3.1 Data exploration

NHS Borders is part of the Scottish public healthcare system, NHS Scotland. From 2021 onwards, NHS Borders has been collecting data from COPD patients wearing the RESpeck device over several weeks. Up to this point, 22 patients have been involved in the study. However, there is no data for PRB002, and PRB008 is not included in the data because they are still wearing the RESpeck so the data is incomplete and only minute-averaged RESpeck data is available. Alongside wearing the RESpeck, patients performed recorded Rehab exercises and filled in surveys about their well-being.

Concerning the patient ids, it should be noted that ids containing an 'X' usually refer to testing data that was artificially created. In this case however, PRX018 and PRX900 are real patients. Information on any special circumstances for the patients was collected for this study but the document in which they were stored was lost. Any circumstantial information mentioned here was obtained by specifically asking NHS Borders about unusual data in certain time periods.

3.1.1 RESpeck data

The RESpeck data consists of measurements taken at 12.5 Hz. Each measurement contains

- the interpolated phone timestamp,
- the RESpeck timestamp expressed as a Unix timestamp,
- a sequence number,

- the magnitude of acceleration along the x, y, and z axes,
- the breathing signal, calculated using the approach in [31],
- the breathing rate, calculated using the approach in [29],
- the activity level which is the Euclidean length of consecutive acceleration values,
- the activity type.

In Figure 3.1 the amount of RESpeck data is visualised. It can be seen that patients had the RESpeck for periods of 4 days (PRB106) to 15 weeks (PRB202). However, patients did not wear the RESpeck for the entire duration. This is expressed in the completeness value for each patient. Patients PRB004 and PRB106 actually only contain RESpeck data for a little less than 2 days or a couple of hours, respectively. Not taking these patients into consideration, patients had the RESpeck for periods of 38 days (PRB109) to 15 weeks (PRB202) with RESpeck data being actually recorded for 12 days (PRB109) to almost 9 weeks (PRB203). For how many complete weeks the RESpeck was worn can be seen in Figure 3.2. The RESpeck was given to patients most frequently for 7 to 9 complete weeks. This reflects the usual length of PR programs of 6 to 8 weeks (42 to 56 days). However, the RESpeck was actually only worn for 1 to 8 complete weeks with missing time periods of up to several days.

3.1.2 Rehab data

The Rehab data consists of recorded exercise sessions. For each session, there is the average and standard deviation of the breathing rate, the start and end timestamp, and a name. Additionally, there is information on the individual exercises performed during the session. For each exercise in the session, there is

- a list of breathing rates,
- the standard deviation of the breathing rate,
- the percentage correctness,
- timestamps of the start and end of the exercise,
- timestamp of the end of the resting time, and
- the exercise id.

Exercise IDs refer to the type of exercise performed (see Table B.1). According to a physiotherapist at NHS Borders, patients consistently rate PR_SIT_TO_STAND to be the most intense exercise and patients were discouraged from using the Rehab app to perform walking exercises (PR_WALKING) because of reoccurring crashes of the app during that exercise. So PR_WALKING exercises that were recorded nevertheless are removed from the analysis.

The amount of recorded sessions and exercises per patient can be seen in Figure 3.3. From the figure it is apparent that sessions are made up of varying amounts of exercises. For patient PRB007, the number of exercises exactly equals the number of exercise sessions. Closer inspection of this patient shows that each sessions consists of exactly

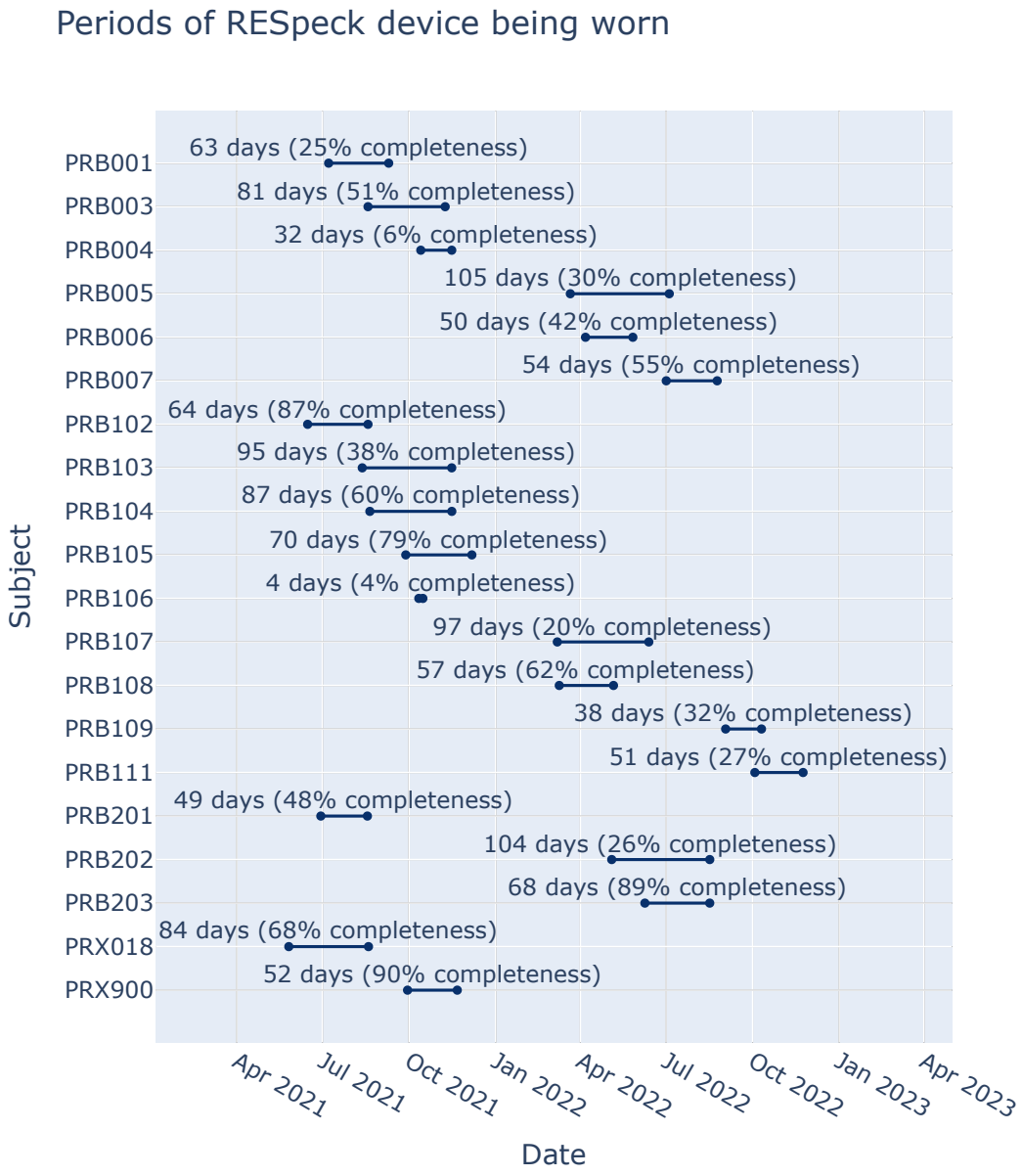


Figure 3.1: Amount of RESpeck data

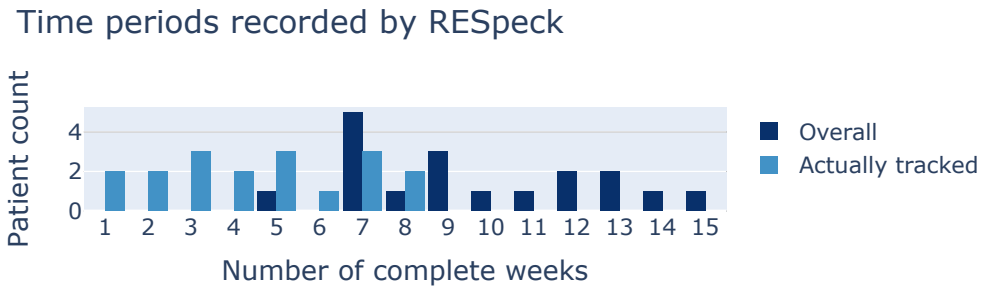


Figure 3.2: Histogram of tracked weeks

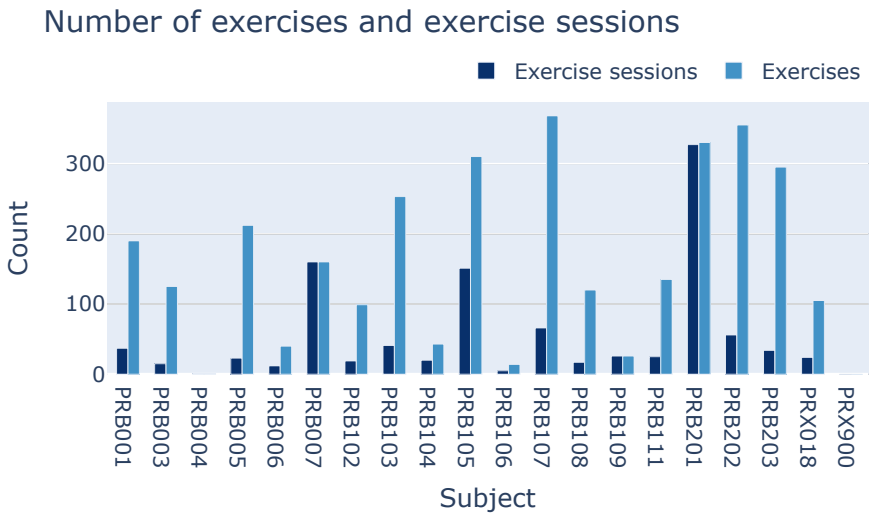


Figure 3.3: Amount of Rehab data

one exercise. However, some of these sessions happen so closely in time that they might be considered one session. Other patients like PRB107 have a very high exercise per session ratio. Looking more closely at PRB107 shows that most exercise sessions are clearly separated in time while some might again be considered one session because of their closeness in time.

Because exercises are grouped together into session inconsistently, exercises should be considered separately or by grouping them together according to specific time constraints, e.g. taking all exercises of a specific day. In the following discussion exercises will usually be considered individually. If they are grouped together, the way in which they are grouped will be described.

Patient PRX900 has no exercise data. Patients PRB004 and PRB106 have only very little exercise data consisting of 1 and 14 exercises respectively. The remaining patients have between 26 and 368 exercises (PRB109 and PRB107 respectively).

3.1.3 Diary data

Patients were asked about their well-being using the CAT via the Diary app. As discussed in Section 2.1.2, the resulting CAT score lies between 0 and 40 where 0 corresponds to good health and 40 corresponds to poor health (see Section 2.1.2). CAT scores were reported by 19 patients. Patient PRB001 did not report any CAT scores.

Figure 3.4 presents an overview of the CAT scores chosen by patients. This overview and graphs plotting the CAT scores in time for each patient (see Appendix D) form the basis for the following analysis.

Patients PRB004, PRB005, and PRB102 only have very few CAT scores. For a COPD patient, patient PRB003 has relatively low scores with only one score being above 10. A closer investigation of the data shows no weird jumps so the data seems to be consistent within itself. Patient PRB103's scores are too constant and too low, being

almost all 1, to be a likely reflection of the patient's health. The CAT scores for patient PRB106 are only few, irregular and jump between values of 7 and 29. Patient PRB107 and PRX900 chose values from a broad range. However, considering the development of scores over time there are only few jumps. The largest of these jumps for PRB107 is from 11 on March 14, 2022, to 27 on March 17, 2022. On speaking with NHS Borders, the jump was explained with a virus infection at that time. Overall, there are many scores over a duration of 3 months for PRB107 and almost 2 months for PRX900 so it seems reasonable that scores of a broad range are obtained. The CAT scores for patient PRB201 are largely consistent but there is a jump down to 15 on June 21, 2021. Patient PRX018's scores are mostly consistent except for the initial score of 4 which is significantly different from the other scores and 2 months away from the other scores. This score will be considered as an outlier and removed. Patients PRB111 and PRB202 contain scores of 0 that do not seem to belong to the data as they do not fit in with CAT scores close by. These outliers most likely originate from mistakes when recording the CAT scores, especially as there is often another CAT score on the same day. Apart from that, the scores of PRB111 suddenly jump up to 31 and 23 at the end of October which corresponds to an infection with COVID-19 according to NHS Borders. Patient PRB202 has an outlier score of 2 on May 19, 2022, which originates from the initial assessment of the patient where the physiotherapist showed them how to use the app.

Overall, the CAT scores of 14 patients look reasonably reliable when removing the discussed outliers.

3.2 Data preprocessing

In conclusion of the data exploration, data for 13 patients is used in the analysis. See Table C.1 for a summary of the available data and which patients' data is included. The RESpeck data for PRB201 contains an entry from January 1, 1970, 00:00:00 which corresponds to a unix timestamp of 0. This datapoint was removed. As described above, all PR_WALKING exercises were removed because of frequent app crashes during this exercise. Further, all 0 CAT scores, the 2 CAT score for PRB202 and the initial score of 4 for PRX018 were removed. Patients with only few CAT scores are not included in the analysis because the CAT scores are too few and vary too much in some cases to evaluate their reliability.

3.2.1 Associating RESpeck and Rehab data with CAT scores

To understand how the RESpeck and Rehab data reflect a patient's well-being, the association between them and the CAT scores was investigated. To this end, each CAT score has to be associated with some RESpeck and some Rehab data.

For this association, two different approaches were taken. In the first approach, each CAT score is associated with RESpeck data and exercises from the day on which the CAT score was recorded. This includes data from after the CAT score was recorded. The hypothesis underlying this approach is that the CAT score reflects the well-being of a patient during the entire day, independent of the time of day when it was recorded. The second approach takes all RESpeck data and exercises from the day prior to the

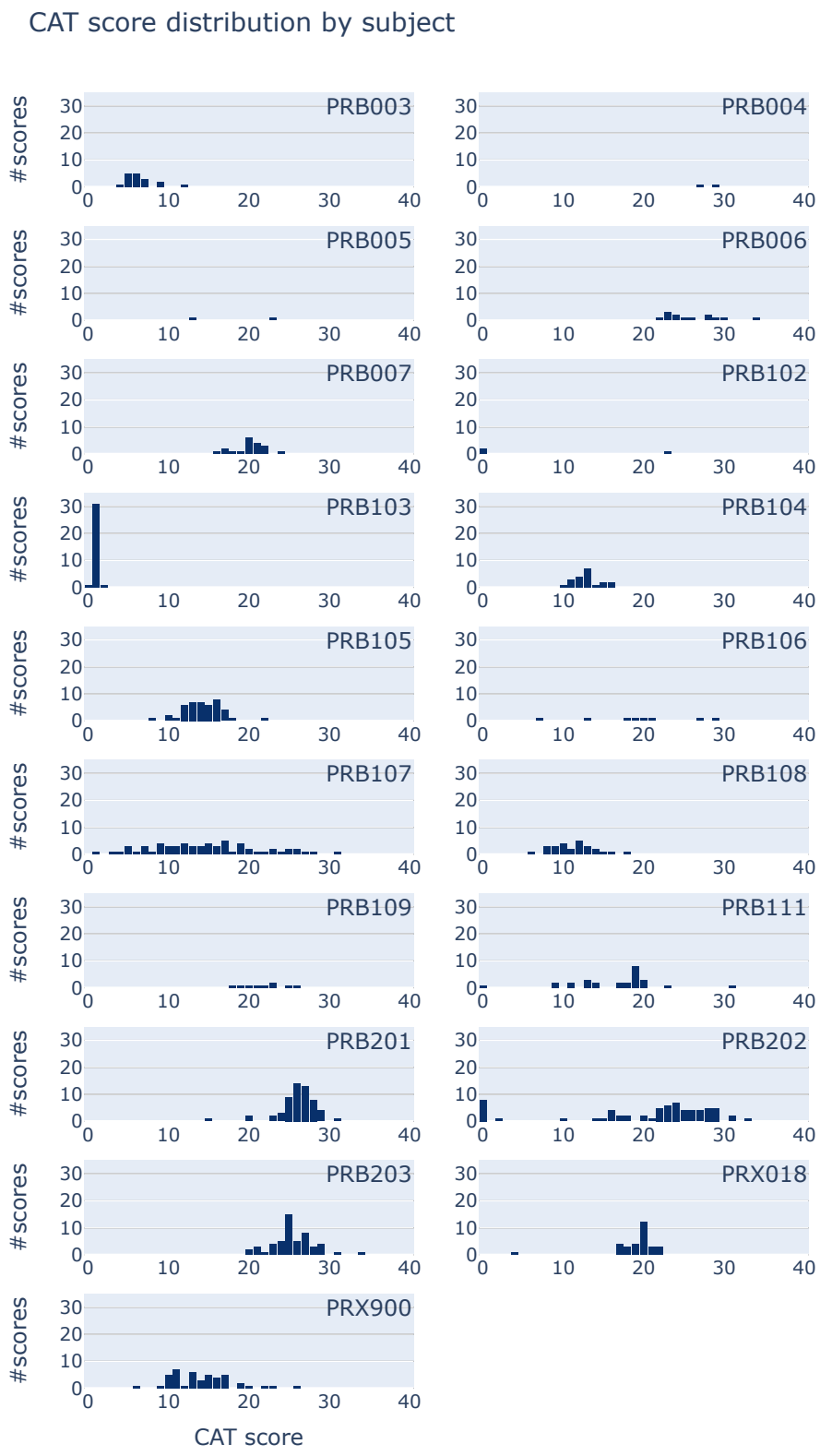


Figure 3.4: Distribution of CAT scores by patient

day the CAT score was recorded. The hypothesis underlying this approach is that the RESpeck and Rehab data of a certain day are predictors of the patient's well-being, the CAT score, on the next day. Both approaches are based on taking data from 24 h windows into consideration because patients were asked to report CAT scores on a daily basis. Overall, there are then 431 sets of CAT scores with corresponding RESpeck and Rehab data.

3.2.2 Identifying reliable RESpeck data

A model for identifying reliable RESpeck data was introduced in Section 2.2.2. The code corresponding to this model lacked documentation and was outdated, so it had to be analysed thoroughly and rewritten.

The model assesses the data by using a sliding window. For each window, the following features are considered:

- standard deviation of the breathing rate,
- mean difference between consecutive values of the breathing rate,
- number of non-NaN values of the breathing rate,
- time between maximum and minimum values of the breathing rate,
- standard deviation of the breathing signal, and
- mean time between consecutive peaks of the breathing signal.

The values for these features are fed into the model which returns either `True` or `False` classifying the data within the time window as reliable or unreliable, respectively. If any of the features can not be evaluated, the model can not assess the reliability for the data in this window.

The model was trained on 20s windows that overlap by half a window, so each datapoint is assessed twice by the model (apart from the datapoints in the first and last 10s). Within the QIP data analysis, the window sizes and overlaps are analysed in detail. For this dataset, the default parameters of 20s windows with 50% overlap are kept. For features based on the RESpeck data, three types of RESpeck data are considered. 'All' RESpeck data corresponds to all the collected RESpeck data without filtering for reliable data. 'Strictly reliable' RESpeck data refers to the subset of RESpeck data where both windows that a datapoint lies in are considered reliable by the model. Including datapoints as well for which only one of the windows is considered reliable is referred to as 'somewhat reliable' RESpeck data. Filtering for reliable data drastically reduces the amount of data that can be used, so considering multiple levels of reliability allows to investigate how much filtering is needed to find an association with the CAT scores.

3.3 Correlations with the CAT score

To estimate the CAT score, different features from the RESpeck and Rehab data are considered. The descriptiveness of each feature is measured by the Pearson correlation. Alternative correlations would be the Spearman and Kendall correlations but the Pearson correlation is the most common and has been used in previous works related to this project [29, 30]. Correlation coefficients with high magnitudes (i.e. that are close to -1 or 1) suggest a strong linear relation between the feature and CAT score while correlation coefficients close to 0 suggest a weak or non-existing linear relation. The p-value associated with the correlation describes how likely it is that the correlation arose by chance. A significance level of $\alpha = 0.05$ will be used, so p-values below α show that there is sufficient evidence that confirms the existence of a linear relationship for the correlation to be considered significant.

Overall, 115 features are included in the analysis. These are considered for data from the same day as the CAT score and data from the previous day. Appendix E shows an overview of the correlation coefficients and corresponding p-values of all features.

3.3.1 CAT scores and exercises

There are 9 types of exercises (excluding PR_WALKING). Exercises in the selected 24h window that are of the same type are considered together. For each exercise type, the following features are considered:

- `avgBreathingRate`: the average breathing rate,
- `avgCorrectness`: the average correctness of the exercises,
- `runThroughs`: how often this type of exercise was performed,
- `avgExerciseDuration`: the average exercise duration,
- `avgBreakDuration`: the average duration of the breaks following the exercises,
- `avgExerciseBreakDurationRatio`: the ratio of the average exercise and break durations,
- `actLevelDuringExercises`: the average activity level during the exercises,
- `actLevelDuringBreaks`: the average activity level during the break following the exercises.

In addition, features that summarise the exercises of the specified time period are included. These features are similar to the features for individual exercise types but instead of just considering the average of the breathing rates, correctness values, exercise durations and break durations, the standard deviations, minima and maxima of these values are considered as well.

The Pearson correlation coefficients of the CAT scores with these features are visualised in Figures 3.5 and 3.6 using RESpeck and Rehab data from the same day as the CAT score or the previous day, respectively. The figures show the correlation coefficients of the CAT scores with the described features for all exercise types, including `exercises`

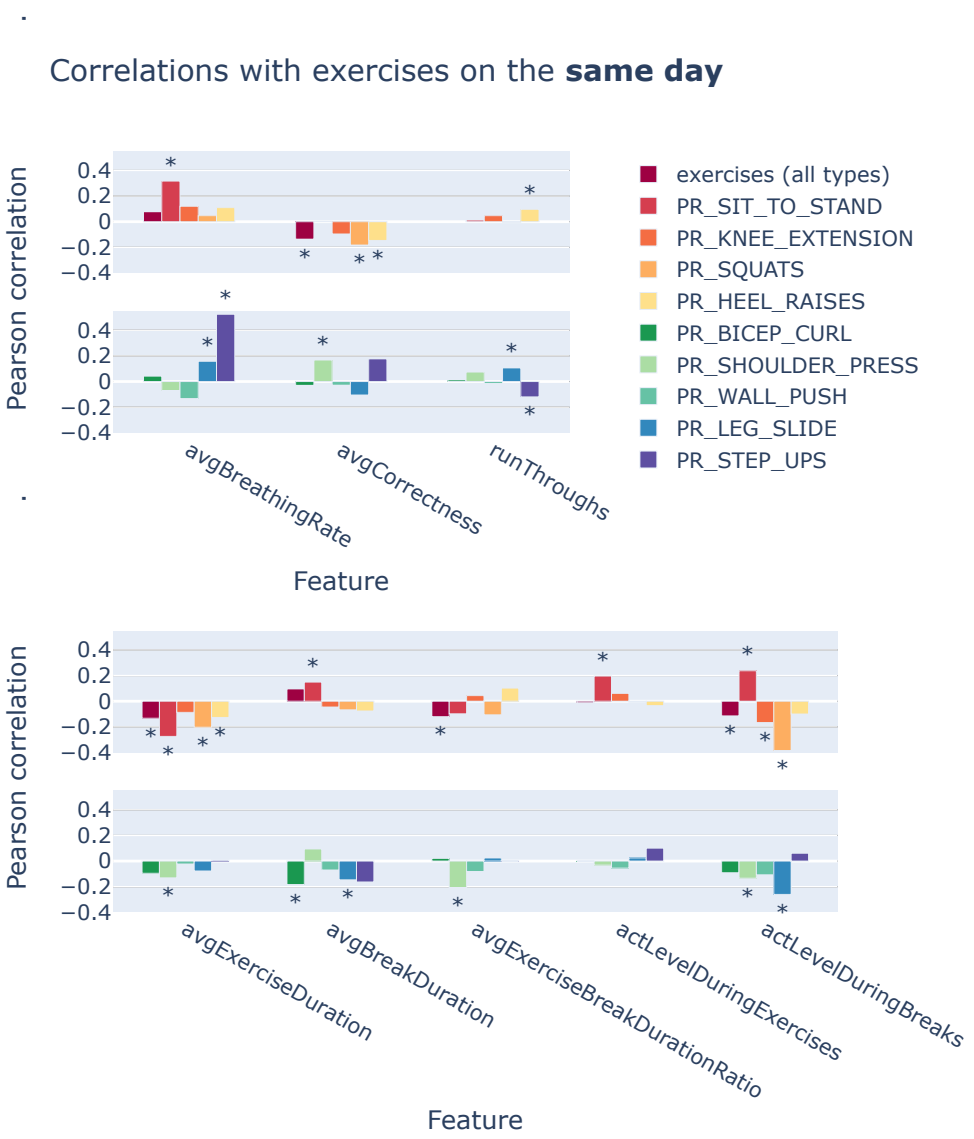


Figure 3.5: Correlations between CAT scores and exercises on the same day. The asterisk $*$ highlights significant correlations, i.e. where the p-value < 0.05.

which takes into account exercises of all types. The correlations with some of the `exercises` features that were not considered for individual exercise types are not shown here but will be discussed later.

A positive correlation implies that the corresponding feature decreases as the CAT score increases, i.e. when the patient feels worse. On the other hand, a negative correlation implies that the corresponding feature increases as the CAT score increases.

Some of the strongest positive correlation coefficients correspond to the breathing rate features. Intuitively, it makes sense for the average breathing rate to have a positive correlation with the CAT scores because it can be expected that patients are out of breath more easily when they are not feeling well. This agrees with the majority of correlation coefficients for the breathing rate features being positive, both for the features based on exercises from the same day and the previous day. In particular, the correlations of

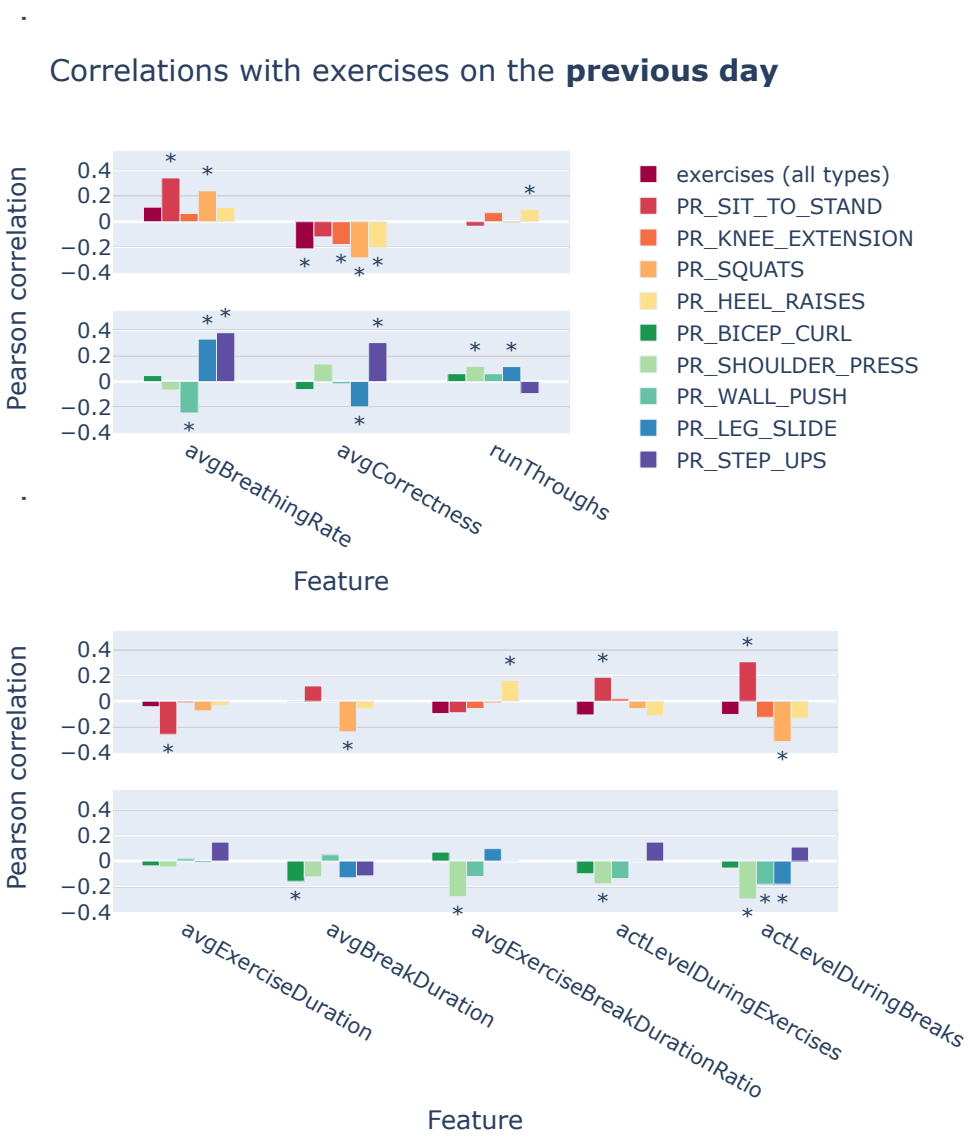


Figure 3.6: Correlations between CAT scores and exercises on the previous day. The asterisk $*$ highlights significant correlations, i.e. where the p-value < 0.05.

CAT scores with the average breathing rates for PR_STEP_UPS, PR_SIT_TO_STAND and PR_LEG_SLIDE exercises are strongly positive with correlation coefficients of up to 0.52 and statistically significant. For the exercises of the previous day, the average breathing rates of PR_WALL_PUSH and PR_SQUATS exercises also have a significant correlation with the CAT scores. Important to note is that the correlation of PR_WALL_PUSH_avgBreathingRate with the CAT score is negative with a correlation coefficient of -0.25. Hence, the average breathing rate during wall pushs being low relates to the patient feeling unwell the next day. Correlations with the average breathing rate for the other exercise types are not significant.

The correlation coefficients for the average correctness features are mostly negative. The strongest negative correlations are found for PR_HEEL_RAISES, PR_SQUATS and when considering all exercise types together. These correlations are significant

both when considering same day's and previous day's data. Using data from the day prior to the CAT score being reported, the correlations are stronger and in addition to PR_HEEL_RAISES, PR_SQUATS and considering all exercise types together, PR_LEG_SLIDE's and PR_KNEE_EXTENSION's average correctness values have a significant negative correlation with the CAT scores as well. Intuitively, it can be expected that exercises are done less accurately when a patient is feeling worse, so negative correlations would be expected. However, for PR_SHOULDER_PRESS and PR_STEP_UPS the average correctness has a positive correlation with the CAT scores. In the case of PR_SHOULDER_PRESS the correlation is significant considering the same day's data and in the case of PR_STEP_UPS it is significant for the previous day's data. These positive correlations might be connected to patients performing exercises more slowly or only exercises with which they feel more comfortable when they are unwell.

There are significant positive correlations for the number of run throughs with the CAT scores for PR_LEG_SLIDE and PR_HEEL_RAISES independent of whether the same day's or previous day's data is used. Using the same day's data, there is as well a significant negative correlation for PR_STEP_UPS. Using the previous day's data, there is another positive correlation for PR_SHOULDER_PRESS. Examining the data more closely, it can be seen that an exercise of a specific type is either done or not done on a certain day and rarely repeated, so `runThroughs` is an indicator of whether an exercise of the specific type was done or not. Following on from that, a positive correlation indicates that exercises of this type are done more often when the patient is feeling unwell while a negative correlation indicates that the patient is doing the exercise less often when feeling unwell.

The CAT score is negatively correlated with the duration of exercises on the same day for all exercise types both when considering them individually and together. The correlation is significant for PR_SIT_TO_STAND, PR_SQUATS, PR_HEEL_RAISES, PR_SHOULDER_PRESS and when considering all exercise types together. The correlations are negative with correlation coefficients between -0.27 and -0.13. For PR_SIT_TO_STAND, the correlation is strongest. As this is often considered to be the most strenuous exercise type and a negative correlation indicates that the exercise duration is reduced when the patient is feeling unwell, this makes sense intuitively. Considering exercises of the previous day, correlations of the exercise duration with the CAT score are much weaker. The correlation coefficient is positive for PR_STEP_UPS but the p-value is 0.24 so this could well be due to chance. The only strong correlation between CAT scores and the duration of previous day's exercises remains for PR_SIT_TO_STAND. With a correlation coefficient of -0.26 and a p-value of 0.003 this correlation is significant and strongly negative.

Correlations between average break durations and CAT scores vary. For most exercise types and when considering them together the data does not show any significant correlation. The break durations after PR_BICEP_CURL exercises are negatively correlated with CAT scores both when considering exercises on the same or the day before the CAT score. PR_LEG_SLIDE also has a significant negative correlation when considering exercises of the same day. PR_SQUATS has a significant negative correlation when considering data from the previous day. These negative correlations

indicate that breaks become shorter when patients are unwell. Patients might try to power through exercises and not make their usual breaks to get it done with quickly when they are not feeling good. For PR_SIT_TO_STAND the correlations are the other way around. It is significant when considering exercises from the same day as the CAT score and even though it is insignificant for exercises of the previous day, there is a tendency towards a a positive correlation there as well. This might be explained by PR_SIT_TO_STAND being considered one of the most demanding exercises so while patients might try to reduce break times for other exercises to be done with it more quickly when feeling unwell, this exercise might be so exhausting that they have to take longer breaks.

The ratio of exercise and break durations is high when the duration of an exercise is relatively long compared to the duration of the following break. The ratio is low if it is the other way around. Hence, a positive correlation with CAT scores indicates that exercise durations become longer in relation to break durations when a patient is feeling unwell. A negative correlation indicates that break durations become longer in relation to exercise durations. For most exercise types, there are no significant correlations so the data does not show any clear linear patterns. Considering all exercise types together, there is a significant negative correlation for exercises of the same day as the CAT score. So overall, exercise durations are reduced in relation to break durations on the day that a patient is feeling unwell. For PR_SHOULDER_PRESS this correlation is particularly pronounced and becomes even stronger when considering previous day's exercises with correlation coefficients of -0.21 and -0.28. For PR_HEEL_RAISES exercises on the day before the correlation is inverted indicating that exercise durations are increased in relation to break durations on the day before the patient is feeling unwell.

The final two groups of features describe the average activity level during exercises or the breaks following these exercises. Generally we might expect correlations between the average activity level and CAT scores to be negative such that the activity level goes down as the patient is feeling worse. Considering the activity level during exercises, this is only reflected in the significant negative correlation for PR_SHOULDER_PRESS exercises on the previous day. Most of these correlations are insignificant. Considering the activity level during breaks, many exercise types and all exercise types considered together have significant negative correlations with the CAT score on the same and next day. Interestingly, for PR_SIT_TO_STAND the correlations for activity levels during exercises and breaks and with CAT scores for the same and next day are all positive and significant.

Overall, the PR_SIT_TO_STAND features have the highest amount of significant correlations, closely followed by features based on PR_SHOULDER_PRESS and PR_LEG_SLIDE. Considering all exercise types together results in a similar amount of significant correlations when exercises of the same day as the CAT score are used. For exercises of the previous day, considering all exercise types there are almost no significant correlations with the CAT scores.

In addition to the exercises visualised in Figures 3.5 and 3.6, more features related to exercises of all types are considered. For the breathing rate, correctness, exercise duration and break duration, there are not only features for the average but also for the maximum,

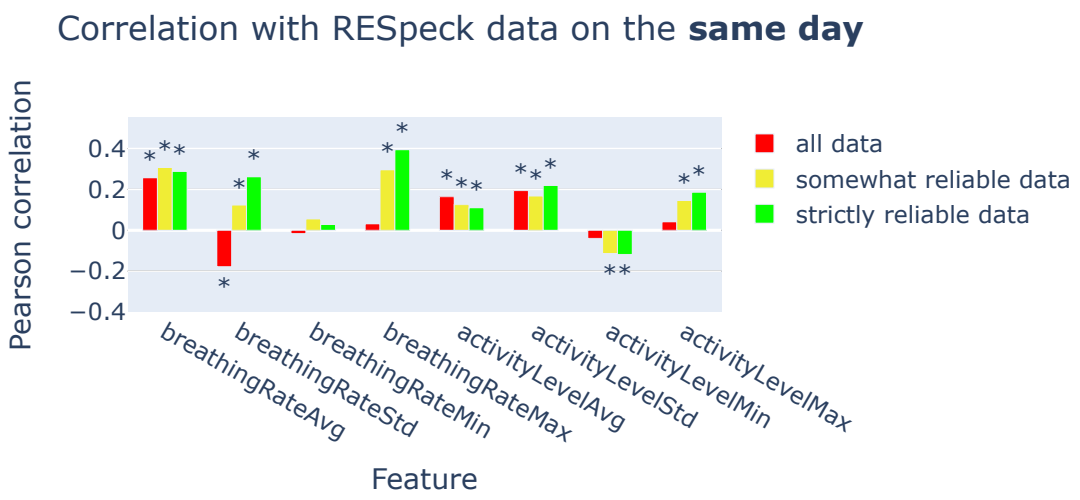


Figure 3.7: Correlations between CAT scores and RESpeck data on the same day. The asterisk $*$ highlights significant correlations, i.e. where the p-value < 0.05.

minimum and standard deviation. The features `exercisesMinBreathingRate` and `exercisesStdCorrectness` have significant positive correlations with the CAT score for exercise on the same and previous day with correlation coefficients from 0.13 to 0.23. The features `exercisesMinBreakDuration`, `exercisesStdExerciseDuration` and `exercisesMinCorrectness` have significant negative correlations for both days with correlation coefficients of -0.15 to -0.22. For exercises on the same day, `exercises MaxExerciseDuration` has a significant negative correlation with $\rho = -0.16$. For exercises on the previous day, `exercisesStdBreathingRate` has a significant negative correlation with $\rho = -0.19$.

3.3.2 CAT scores and RESpeck data

The remaining features involve only data collected by the RESpeck. As before, data from the same day as the CAT score and data from the previous day is examined. From the available REspeck data, the breathing rate and activity level are used as features. For each of them, averages, standard deviations, minima and maxima are considered. These features are calculated using all RESpeck data and only using "somewhat reliable" and "strictly reliable" data (see 3.2.2).

Figures 3.7 and 3.8 show the correlation coefficients for these features with the CAT score for data from the same day and data from the previous day respectively. Overall, it can be noticed that more correlations with RESpeck data from the previous day are significant than correlations with data from the same day. Further, a pattern similar to steps going upwards (downwards for negative correlations) can be observed when comparing the features for "all data", "somewhat reliable data" and "strictly reliable data" where correlations are strongest for "strictly reliable data". This indicates that the model used for filtering for reliable data is good at reducing noise which might obscure the correlation with the CAT score. The opposite pattern can be observed for

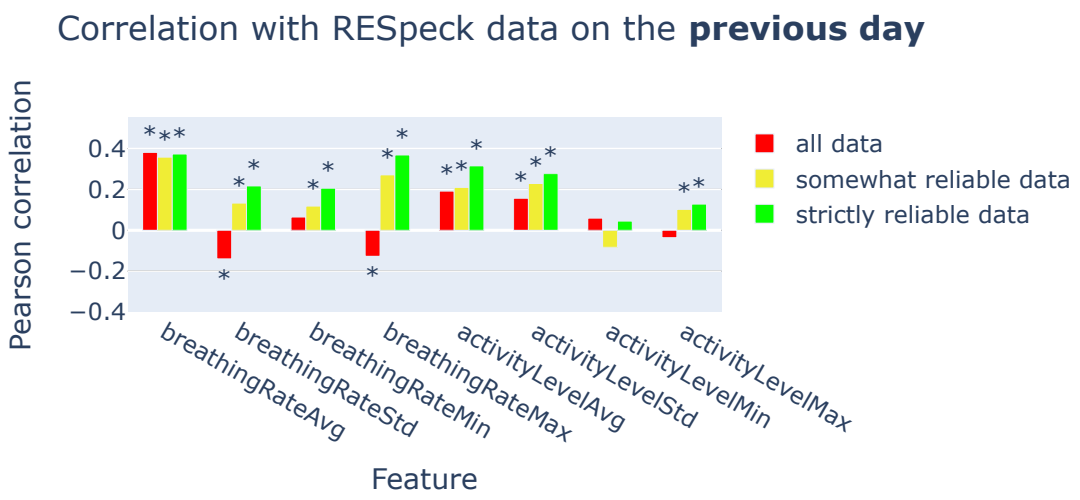


Figure 3.8: Correlations between CAT scores and RESpeck data on the previous day. The asterisk $*$ highlights significant correlations, i.e. where the p-value < 0.05.

the average activity level when considering data from the same day as the CAT score.

Correlations with the average breathing rate, average activity level and standard deviation of the activity level are generally significantly positive indicating that they increase when a patient is feeling unwell. The positive correlation for the average breathing rate agrees with the positive correlations that were identified for the average breathing rates during exercises. The maximum of the activity level also has a positive correlation with the CAT scores but only for "somewhat reliable" and "strictly reliable" data.

For the standard deviation and maximum of the breathing rate, an interesting pattern can be observed where the correlation is negative when including all RESpeck data and positive when filtering for more reliable data. Nevertheless, all correlations are significant according to their p-values. This might have something to do with what RESpeck data is treated as unreliable and what the patient is doing at that time.

The minimum breathing rate has no correlation considering RESpeck data on the same day as the CAT score but a strong positive correlation on the previous day for "somewhat reliable" and "strictly reliable" data. This positive correlation indicates that the minimum breathing rate goes up on the day before the patient is feeling worse.

The minimum activity level of "somewhat reliable" and "strictly reliable" data has a significant but weak negative correlation with the CAT score on the same day. When the CAT score of the following day is considered, there is no correlation. The weak negative correlation indicates that the minimum activity level tends to be smaller when the patient is feeling unwell. This might be because the patient is avoiding movement when feeling unwell.

In conclusion, the RESpeck data already provides good indicators of the CAT score without considering the Rehab data. High correlations identified here had correlation coefficients between about 0.2 and 0.4. Including the exercise data results in features

that have correlations up to the same range (flipping the sign for negative correlations). Only the feature PR_STEP_UPS_avgBreathingRate is above this range with a correlation coefficient of 0.52.

3.4 CAT score association and prediction

The CAT score can be seen as an indicator of a patient's well being. Hence, CAT score association and prediction corresponds to evaluating a patient's current well being and predicting the future well being. CAT score association will be based on using the features that were calculated from RESpeck and Rehab data of the same day as when CAT scores were reported. CAT score prediction will be based on using the features from the previous day. After preprocessing the data as described, 431 CAT scores remain so the association and prediction models will be based on 431 feature sets with their corresponding CAT scores. For training the models and comparing their performances, only 70% of the data (301 feature sets) are used. The remaining 30% (130 feature sets) are the test set on which the final model will be evaluated.

3.4.1 Baseline and measures of error

As in [30] and [29] a baseline model that always outputs the average of the seen CAT scores is used. The performances of the models developed in [30] and [29] will be considered for comparison purposes (see Section 2.3). However, [30] used an earlier and much smaller version of the dataset while [29] used an entirely different dataset.

To quantitatively assess model performances, the mean absolute error (MAE) and root mean squared error (RMSE) are used as in [30] and [29]. This is a common combination for assessing model performances because MAE is easy to interpret while RMSE is better for comparing models because it magnifies large deviations from the true score and reduces the error caused by small deviations. To include an error metric that focuses less on the mathematical properties of CAT scores and more on how they are used in practice, the accuracy of whether the true and predicted CAT scores lie in the same impact level group (see 2.1.2) will be used as an additional error metric. The dataset is relatively small so it can be expected that the model performances vary a lot depending on exactly which CAT scores are used for training and which for validation. Hence, 10-fold validation is used in the evaluation of all models to stabilise the error values.

For both CAT score association and prediction, the baseline model outputs the average CAT score. Hence, the baseline model performs the same for association and prediction. It has an MAE of 5.7, an RMSE of 6.8 and an accuracy of 47%. The MAE implies that the estimated CAT score is on average 5.7 units away from the true CAT score. On a scale from 0 to 40 that is an error of 14% relative to the scale. An accuracy of 47% implies that for more than half of the samples the estimated CAT score was within a different impact level group than the true CAT score. In [29], the baseline model achieved an MAE of 11.3 and an RMSE of 12.0. This indicates that in the dataset analysed there, the CAT scores were much more spread out and less close to the average CAT score. In [30], the baseline model had an MAE of 7.5 and an RMSE of 8.3. These

values are more similar but still significantly higher than for the baseline model used here.

3.4.2 Histogram-based Gradient Boosting Regression Tree

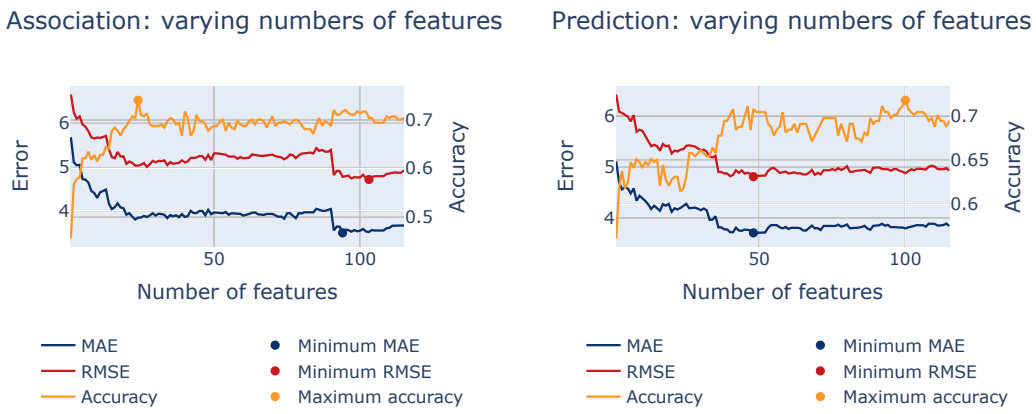
From the 431 feature sets, only 10 feature sets of the same day's data and 8 feature sets of the previous day's data do not contain NaN values. Because not every patient is doing every type of exercise on the day and the day prior to reporting a CAT score, this was to be expected. As a consequence, the model used for estimating the CAT score must be able to handle NaN values.

The histogram-based gradient boosting regression tree as implemented in `ensemble` in `sklearn` allows NaN values. The model is based on an ensemble of decision trees that are fitted using gradient descent. Gradient boosting is in general considered to be a powerful and popular machine learning algorithm [32].

The model is trained and evaluated with different feature subsets. First, all features are considered. Then, only those features for which a significant correlation with the CAT score was found, i.e. the corresponding p-value is below 0.05, are included. Finally, in an attempt to find the best set of features to use, all features are sorted according to the significance of their correlation and the correlation coefficient. To this end, the p-value is rounded to two decimal places and sorted in ascending order. Features with the same rounded p-value are sorted in descending order according to the absolute value of their correlation coefficient. The model is then trained and evaluated for the first n features of this list. Figures 3.9a and 3.9b show how well the model performs for varying n in CAT score association and prediction respectively. For CAT score association, the maximum accuracy is reached by choosing the best 24 features while the minimum MAE and RMSE are only reached for 94 and 103 features respectively. There is a distinct reduction in MAE and RMSE when going from chosing 90 features to 91 features. For CAT score prediction, the MAE and RMSE values initially decreases rapidly for a growing number of features. From about 40 features onward, both MAE and RMSE stay relatively consistent. The accuracy is much more volatile and reaches a clear maximum at 100 features.

For all feature subsets, a histogram-based gradient boosting regression tree was trained and evaluated using 10-fold validation. The models' performances are displayed in Table 3.1. For the CAT score association, including only significant features compared to including all features reduces the model's performance slightly. For the CAT score prediction, it does not really make a difference. As would have been expected, using only the best features according to RMSE and accuracy optimises the corresponding error metrics.

For CAT score association, the big change in accuracy from optimising according to RMSE to optimising according to accuracy aligns with the observations of Figure 3.9a. Whether using the best features according to RMSE or accuracy is better depends on whether it is more important that the estimated CAT score is close to the true CAT score or that they belong in the same impact level group. Considering that the MAE value, which describes the mean absolute distance between true and estimated CAT score,



(a) CAT score association

(b) CAT score prediction

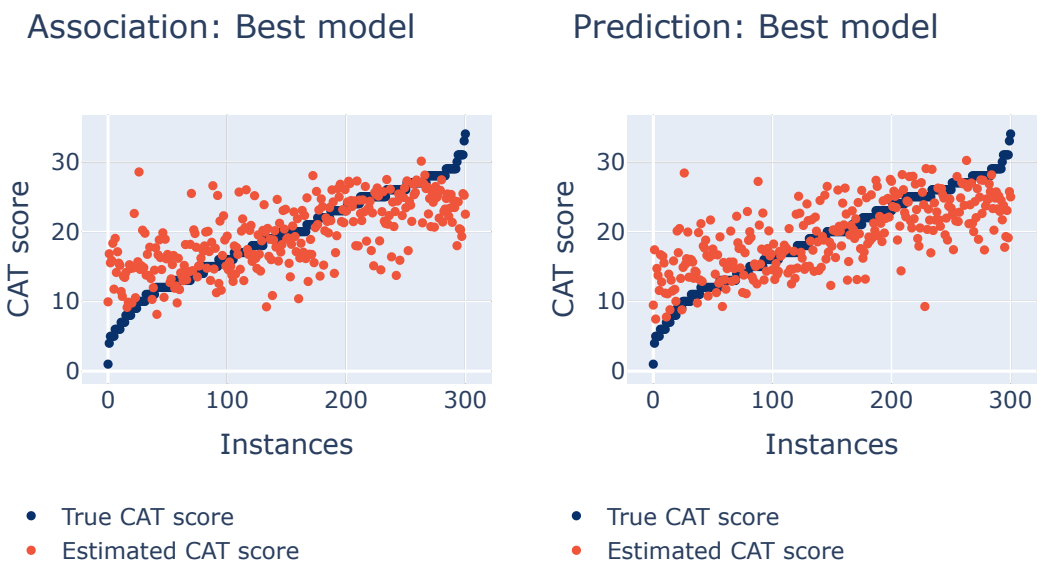
Figure 3.9: Training and evaluating the model with varying numbers of features

		MAE	RMSE	Accuracy
Baseline		5.7	6.8	47%
Association	all features	3.7	4.9	70%
	only significant features	4.0	5.3	69%
	best features according to RMSE	3.5	4.7	70%
	best features according to accuracy	3.8	5.0	74%
Prediction	all features	3.8	4.9	69%
	only significant features	3.7	4.8	70%
	best features according to RMSE	3.7	4.8	71%
	best features according to accuracy	3.8	4.9	72%

Table 3.1: Model performances

only changes by 0.3 units, the accuracy improvement by 4 percent points when using the best features according to accuracy seems more significant. Based on this argument, the model using the best features according to accuracy is the best model for CAT score association. Figure 3.10a illustrates CAT score estimates made by this model. The instances are sorted by the true CAT score so the upwards trend of estimated CAT scores illustrates a good correlation between true and estimated CAT scores. Quantifying this correlation, the correlation coefficient is 0.67 with a p-value of about 10^{-40}. Hence, the correlation is very significant and relatively strongly positive as would be expected.

For CAT score prediction, all models perform very similar with respect to MAE and RMSE values. There are only slight differences in accuracy values but based on these differences, the model using the best features according to accuracy is the best model with an MAE of 3.8, RMSE of 4.9 and accuracy of 72%. CAT score predictions made by this model are illustrated in Figure 3.10b. The correlation of the estimated and true CAT scores has a coefficient of 0.69 and a p-value of about 10^{-44}. So, as for CAT score association, the correlation is very significant and even slightly more strongly positive.



(a) CAT score association

(b) CAT score prediction

Figure 3.10: Model estimations

3.4.3 Final model's performance

Until this point, the model evaluation was based solely on using data from the training and validation sets. For a final evaluation of the best model, a test set was set aside in the beginning. On this test set, the model achieves an MAE of 3.6, an RMSE of 4.7 and an accuracy of 67% for CAT score association and an MAE of 3.5, an RMSE of 4.7 and an accuracy of 72% for CAT score prediction. The MAE and RMSE values are similar or even a bit better than the ones achieved previously. However, the accuracy for CAT score association is the worst one so far, excluding the baseline model's accuracy, even though feature selection was optimised for accuracy.

Histograms of the true and estimated CAT scores illustrate how the estimated CAT scores are smoothed out in comparison to the true CAT scores. In particular, the relatively low frequency of scores from 18 to 19 is not reflected in the estimated scores. Scores close to the impact level borders like these are important for accuracy but the used model, a regression model, does not put any emphasis on differentiating between scores of different impact levels. To put focus on the accuracy of the impact levels, a classifier that differentiates between these impact levels should have been used directly. As all considered models are regressors, the model selection should be based on MAE and RMSE values that are more closely related to the aim of regression. Evaluating the model using the best features according to RMSE on the test set achieves an MAE of 3.0, an RMSE of 4.1 and an accuracy of 75% for CAT score association and an MAE of 3.3, an RMSE of 4.5 and an accuracy of 73% for CAT score prediction. These are the best values so far. This demonstrates that models should be selected with respect to an error metric that is related to the intrinsic error metrics of the model. Further, it demonstrates that optimising with respect to reasonable error metrics results in better accuracy values than directly optimising with respect to accuracy.

Considering the CAT's minimum clinically important difference (MCID) for improvement is estimated to lie between -1 and -4 (see Section 2.1.2), models with MAE values of 3.0 and 3.3 are already reasonable CAT score estimators. Finally, the improvement of all error metrics in comparison to all previous models shows that the additional data from using the entire training set for training, and not splitting it into training and validation sets, improves the model's performance. This suggests that the model's performance might be significantly improved by training on more data.

In [29] and [30], the features are only built on the data preceding the CAT scores so the models are predictors. In [29], the best model has an MAE of 2.8 and an RMSE of 4.6 while in [30], the best model has an MAE of 5.4 and an RMSE of 6.6. Hence, this model's performance is similar to the model's performance in [29]. It is better than the model's performance in [30] which was to be expected because the data originated from the same study but since [30] more data has been collected.

3.4.4 Patient clustering

In an initial approach to the preprocessing of the data, no CAT scores were classified as outliers, so no values were removed. In addition, all patients that had at least one CAT score, some RESpeck data and some exercise data were included in the analysis. As a consequence, only patients PRB001, PRB002 and PRX900 were excluded from the analysis. At the point when this approach was taken, the data from PRB111 was not collected yet, so the analysis included data from 17 patients.

Considering all scores as valid, it can be observed in Figure 3.4 that patients differ greatly in how they chose CAT scores. Some patients only chose values from a narrow range and some patients use almost the entire range of values. The patients choosing values from a narrow range seem to have an absolute scale in mind, while patients choosing form a wide range seem to have a more relative scale in mind. A clustering method was applied to split patients into groups of similarly behaving patients. For each patient, their behaviour was quantified by calculating the average and standard deviation of the chosen CAT scores. Figure 3.11 illustrates the similarities and differences in which patients are choosing CAT scores. The axes show the average and standard deviation of the CAT scores of each patient while the size of each patient's marker visualises the amount of recorded CAT scores.

Because clustering algorithms aim to minimise the distance between elements in a cluster, without further constraints, the optimal clustering would be putting each patient in a separate cluster. However, small clusters have little data which makes them prone to overfitting, and the goal is not to consider patients individually but to group them together according to similarities in behaviour. Hence, a sensible number of clusters has to be decided in advance of applying a clustering algorithm. From inspection of Figure 3.11 (and ignoring the clustering suggested by the colours), a number of six clusters makes sense taking into account that small clusters should be avoided.

Applying KMeans clustering with six clusters and normalised features yields the clustering visualised by the colouring in Figure 3.11. Clusters 0, 2, 3, and 5 have low standard deviations and differentiate between patients by their CAT score ranges. Cluster 1

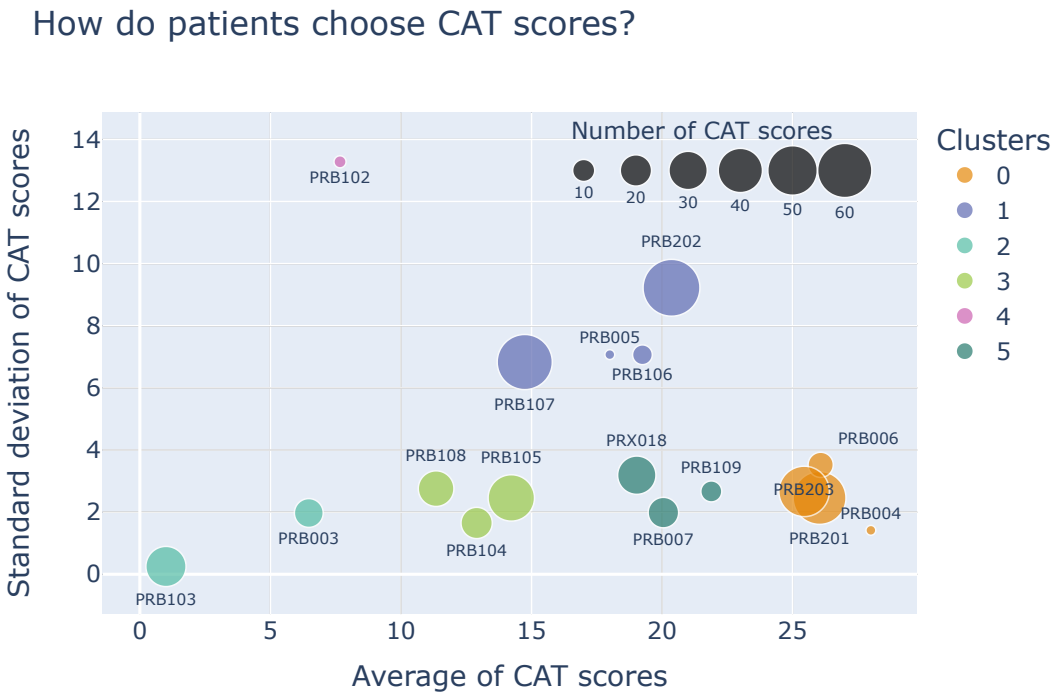


Figure 3.11: Patient profiles with almost no data cleaning

includes patients that choose CAT scores from a broad range. Cluster 4 only consists of patient PRB102 for whom only 2 CAT scores were recorded. Figure 3.12 shows what the same figure looks like with data processed as described in Section 3.2. Note that the ranges of the axes stay the same. The clusters described earlier do not make sense in this setting. Because outlier CAT scores are removed, the standard deviations of all patients are much smaller, and considering the CAT score averages, the distinction between typical averages is much less clear. With the removal of outlier scores, the ways in which patients choose CAT scores look much more similar.

Taking the almost uncleaned data, the baseline model has an MAE of 7.3, RMSE of 8.6 and an accuracy of 38%. In comparison with the baseline model performance on the cleaned data (see Section 3.4.1) the values are worse because the uncleaned data has much more variance. For each cluster of the almost uncleaned data, a gradient boosting model was trained. Similarly to the feature subsets described in Section 3.4.2, the models were trained and evaluated using varying subsets of the features. Sorting the features by the absolute values of their correlation coefficients with the CAT score and picking the best n_i features for each cluster i according to the accuracy performs best. For CAT score association, an MAE of 3.3, RMSE of 5.2 and accuracy of 77% was reached. For CAT score prediction, the MAE is the same while the RMSE and accuracy become slightly worse with values of 5.4 and 76% respectively. The RMSE values are worse than those achieved for the cleaned data but the MAE is slightly and the accuracy is much better. However, some of the estimated CAT scores are just random zero scores or 20-25 units higher than the corresponding true CAT score which does not happen

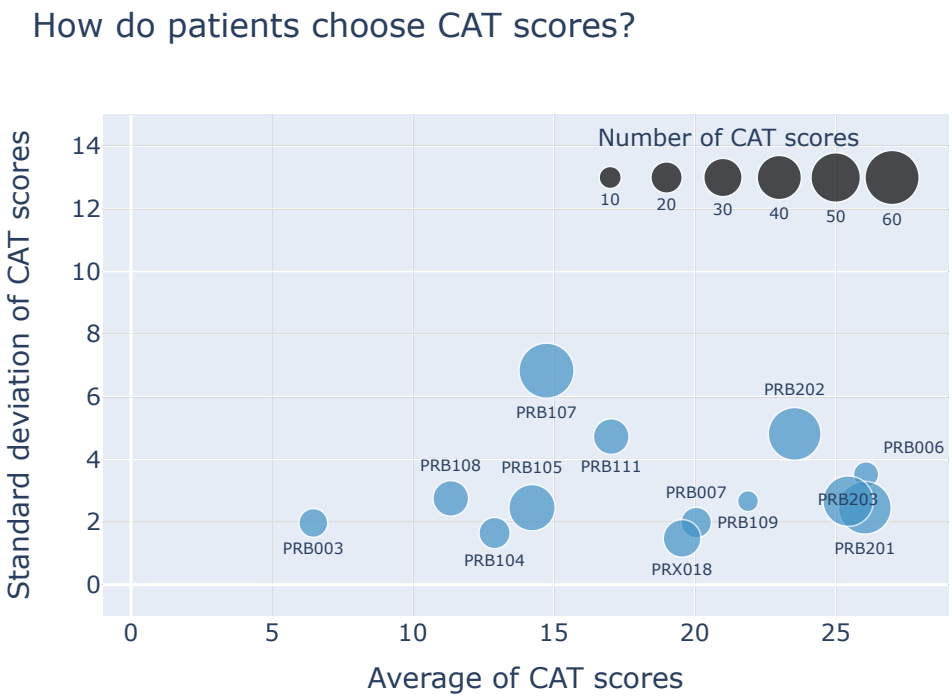


Figure 3.12: Patient profiles with properly cleaned data

for the models trained on the cleaned data as shown in Figures 3.10a and 3.10b. In addition, investigating the model estimates more closely reveals that most estimates are just the CAT score averages for each cluster. This is confirmed by building a second baseline model that always outputs the CAT score average of the corresponding cluster which achieves an MAE of 3.5, RMSE of 5.2 and accuracy of 73%. As a conclusion, the average and standard deviation of chosen CAT scores might be useful features in estimating future CAT scores.

Chapter 4

QIP dataset

The QIP datasets contains breathing rate data recorded by the RESpeck and nurses. The aim is to investigate the reliability and information content of each in comparison of the other.

4.1 Data exploration

The QIP dataset consists of overall 89 post-operative patients whose respiratory rate was measured during their stay at the Western General Hospital in Edinburgh. The respiratory rate was measured automatically by the RESpeck device and manually by nurses of the hospital. Only the data for 43 patients is fully digitalised. For them, data was collected from August 2019 until March 2020.

4.1.1 RESpeck data

There is only data for 36 patients wearing the RESpeck. They wore the device for 1.4 to 12.9 hours with an average of 8.3 hours and measurements taken at 12.5Hz. The collected RESpeck data contains various information connected to a patient's breathing as described in Section 3.1.1.

4.1.2 Nurse data

The data collected by nurses consists of

- the timestamp,
- the extend to which the patient is able to move their eyes (scale from 4 to 1), speak (scale from 5 to 1), and move their body (scale from 6 to 1), following the Glasgow Coma Scale (GCA) [33] to assess a patient's consciousness,
- the respiration rate in breathing rates per minute (bpm),
- the SpO2 scale and value to estimate the amount of oxygen in the blood [34],
- information on whether the patient breathes using air or oxygen supply,

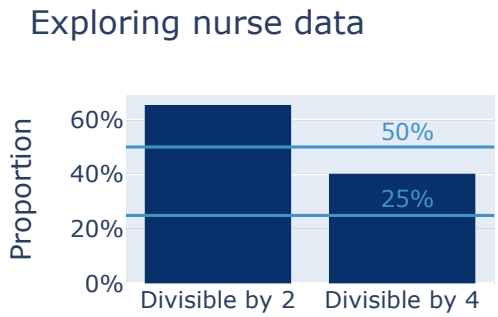


Figure 4.1: Divisibility of breathing rates as measured by nurses

- the systolic blood pressure in mmHg,
- the pulse per minute,
- the temperature in °C.

This report only considers the breathing rate and the associated timestamps. As the aim is to compare nurse and RESpeck measurements, only data for patients for whom there is also RESpeck data will be considered.

There are 386 nurse measurements with an average of 10.7 values per patient. The minimum amount of measurements for a patient is two and the maximum amount of measurements for a patient is 21. These measurements cover on average 9.2 hours where 0.8 hours is the shortest covered time period and 12.5 hours is the maximum time period.

As nurses work under a lot of time pressure, they often only count breaths for 15s or 30s and then multiply the results by 4 or 2 respectively to obtain the number of breaths per minute [1]. A plot of the proportion of breathing rates that are divisible by 2 and 4 for this dataset can be seen in Figure 4.1. Clearly, the proportion of even breathing rates, 66%, is much higher than 50%, which would usually be expected. The same applies to the proportion of breathing rates that are divisible by 4 which is 15% higher than what would usually have been expected. Even taking 66% as the base proportion of even breathing rates, we would only have expected 33% of breathing rates being divisible by 4 and not 40%. Hence, the data indicates that it is likely that at least some of the measurements are based on 15s and 30s monitoring periods.

4.2 Identifying reliable data

A model for identifying reliable data is introduced in Sections 2.2.2 and 3.2.2. By default, the data is considered in 20s windows that overlap by 50%, so each datapoint's reliability is evaluated in the context of two windows. The graph in Figure 4.2 shows the model's results for patient QIXX09. 'Not applicable' are datapoints that were always in windows in which some of the feature values were no numbers, so the model was unable to classify them. 'Reliable' are datapoints for which the model classified all

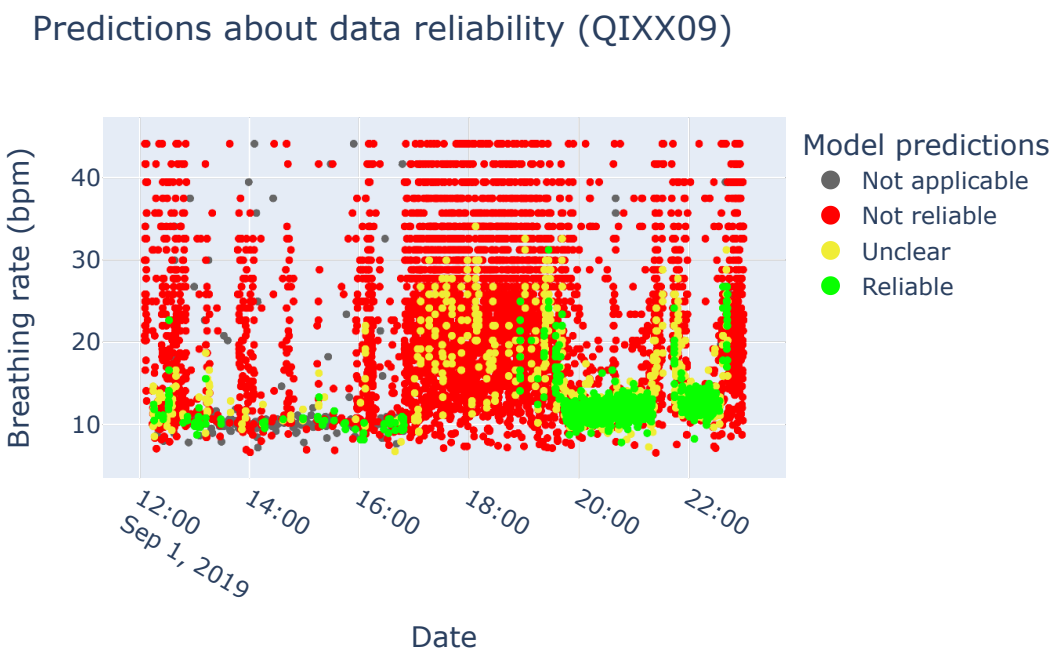


Figure 4.2: Data reliability for QIXX09 considering 20s windows with 50% overlap

corresponding windows as reliable. 'Not reliable' are datapoints for which the model classified all corresponding windows as unreliable. Datapoints labelled as 'unclear' are datapoints for which the model classified one window as reliable and the other as unreliable.

Periods in which the breathing rate is jumping around are correctly classified as unreliable. See the time period from 17:00 - 19:00 in Figure 4.2 for example. These values most likely come from the patient moving around. Where the breathing rates are more steady the model classifies the data as reliable as expected. Normal breathing rates for adults at rest lie between 12 and 20 breaths per minute [35]. Breathing rates in the reliable data lie mostly within this range. However, there are many 'unclear' datapoints. To reduce the amount of these, alternative window sizes and window overlap proportions were considered.

The same model to detect reliable data was applied on features based on 10s, 20s, 30s, 46s and 60s windows overlapping by 50%, 67% and 80%. Figure 4.3 shows the proportion of reliable data for different configurations. For a 50% overlap, most datapoints are in two windows while for 67% and 80% overlaps, datapoints are in three or five windows respectively. A datapoint is only classified as 'unclear' if it is in as many windows classified 'unreliable' as windows classified 'reliable'. When each datapoint is in an odd number of windows, much less datapoints are classified as 'unclear' (a draw is still possible because some windows can not be assessed by the model). This is reflected in the graph: the proportion of reliable data is similar for 67% and 80% overlaps and distinctly higher than the proportion of reliable data for a 50% overlap. There are even fewer 'unclear' and 'not applicable' datapoints for a 80% overlap compared to a 67% overlap, so we will continue working with an overlap of 80%.

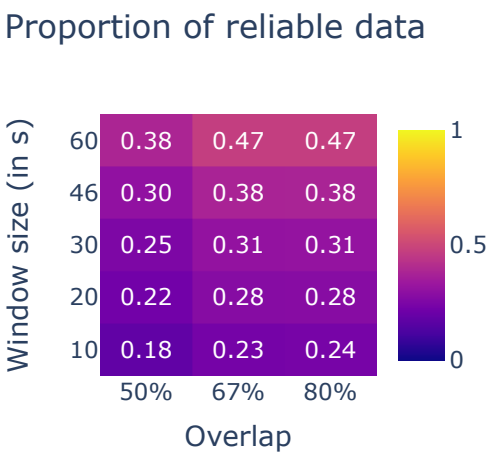


Figure 4.3: Proportion of reliable data

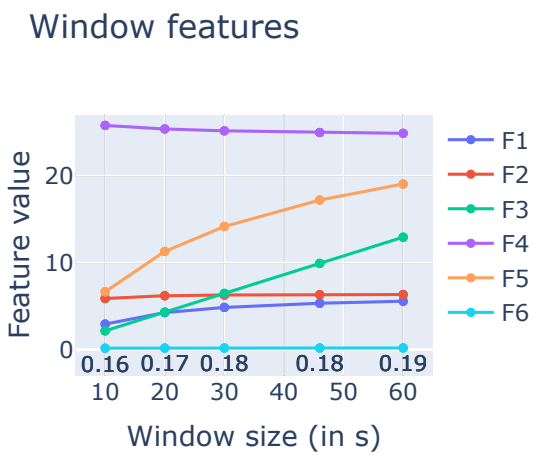


Figure 4.4: Effect of varying window sizes

As the window size increases so does the proportion of reliable data. However, Figure 4.5 gives an indication that changing the windows size might not be a good idea. The plot displays how frequent certain breathing rates are in the data classified as reliable for varying window sizes. Interesting to note is the "upstairs" pattern for breathing rates from 5 to 10 bpm and above 20 bpm which implies that these breathing rates are more common for larger window sizes. On the other hand there is a "downstairs" pattern for breathing rates between 15 and 20 bpm, and the breathing rates between 10 and 15 bpm take up the biggest proportion of reliable data when using 20s windows. Considering that normal breathing rates lie between 12 and 20 bpm [35], the frequency distribution of breathing rates seems most sensible for the 20s windows. Even considering that the patients are in a hospital and unusual breathing rates are to be expected, they are not expected to come up as frequent as for the 60s windows for example.

Changing the window size affects the features that are given as input to the model. Figure 4.4 shows how the average value for the features of each window varies as the window size varies. The standard deviation of the breathing rate (F1) and breathing signal (F6) increase as the window size increases as well as the number of non-NaN breathing rates (F3) and the time between the maximum and minimum breathing rates (F5). It is reasonable that these features depend on the size of the time interval considered. The mean difference between consecutive breathing rates (F2) and the mean time between peaks of the breathing signal (F4) do not vary much with the window size. Both features consider the means of periodically occurring patterns (consecutive breathing rates or consecutive peaks in the breathing signal). Varying the window size implies only that more values are considered to calculate the means but the time between values that are compared does not increase. Hence, it is reasonable that these features do not vary much for different window sizes.

The model was trained on 20s windows so it assesses the reliability of a window in comparison to what features look like in reliable 20s windows. For example, the average number of non-NaN values of the breathing rate is 6.6 in 20s windows, so values above

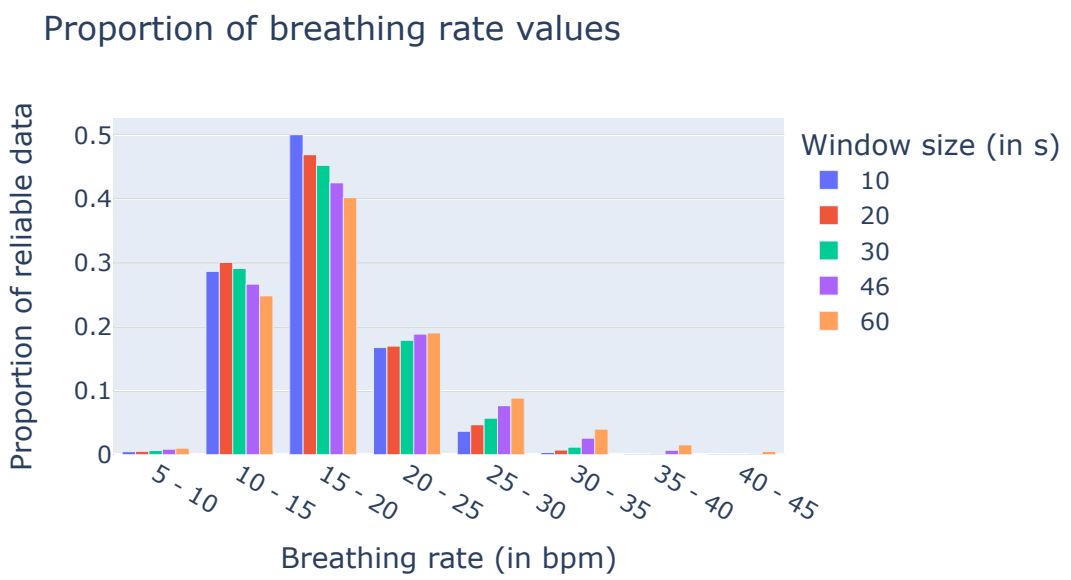


Figure 4.5: Proportion of breathing rate values in reliable data for varying window sizes with 80% overlap

this can be considered high. However, by considering larger windows, naturally the number of non-NaN values will increase as Figure 4.4 confirms. So if a higher number of non-NaN values in the breathing rate implies higher reliability then larger windows will always be more reliable. The results of applying the model on 60s windows can be seen in 4.6. The model does not scale well with varying windows sizes. To adjust the model to other window sizes it would need to be retrained. It is not possible to retrain the model in this context because this dataset does not include measurements that could be considered to be the true breathing rate. A possibility would be to use the dataset with which the original model was trained but this would go beyond the scope of this project. Hence, in this analysis the original model will be used with features based on 20s windows.

In contrast to the window size, the overlap does not matter to the model. The model assesses each window individually and changing the overlap of windows will not change the input features for the model. Hence, it will not affect the quality of classification of the model, only the resolution with which the data is considered. During the following analysis, only RESpeck data classified as 'reliable' based on 20s windows with 80% overlap will be used. This corresponds to 47% of the original RESpeck data.

4.3 Comparing RESpeck and nurse data

4.3.1 Data overlap

Data collected by the RESpeck and nurses does not necessarily overlap. Figure 4.7a shows an example where the nurses and RESpeck measured the breathing rate over approximately the same period of time. In contrast, Figure 4.7b shows a case in which

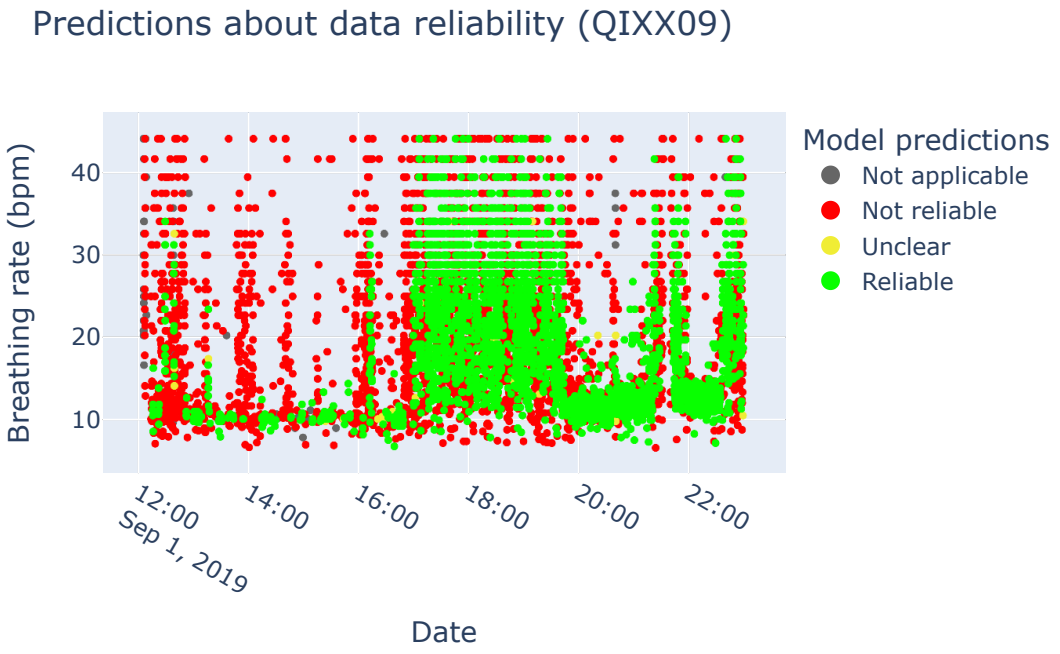


Figure 4.6: Data reliability for QIXX09 considering 60s windows with 80% overlap

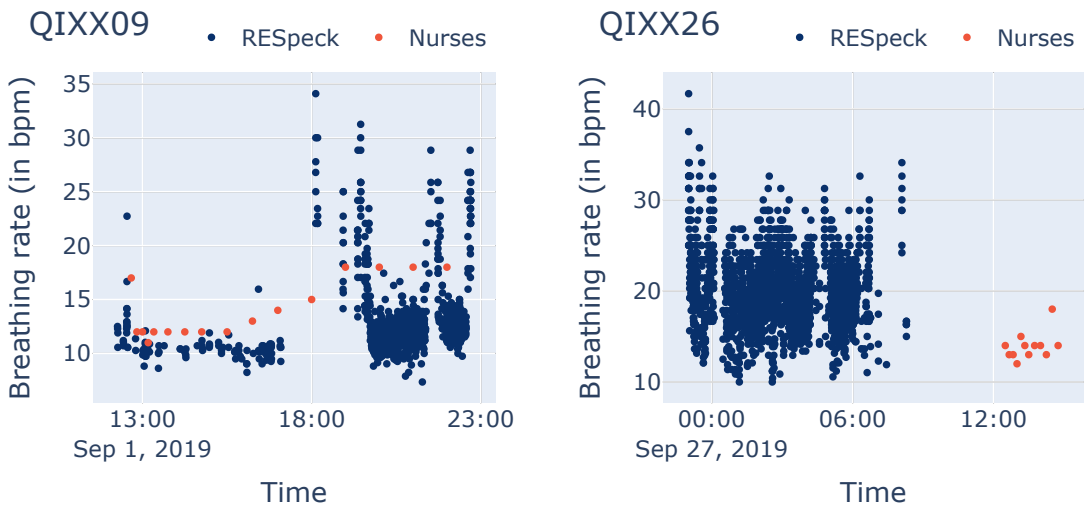
the nurse and RESpeck data do not overlap at all. Figure 4.8 gives an overview of how much the two types of data overlap. For each patient, the number of hours in which only RESpeck data, only nurse data and both were collected are shown. For 6 patients, there is no hour for which both RESpeck and nurse data was collected. For 13 patients, there are more than 5 hours of shared data, and overall there are 118 hours of shared data. To properly compare RESpeck and nurse data, values should be compared that were recorded at similar times for the same patient. Hence, in the next sections only data from the 118 hours for which there exists both nurse and RESpeck data will be used if not stated otherwise.

4.3.2 Relationship between RESpeck and nurse data

Taking the hourly average of nurse and RESpeck data for the hours and patients for which there is data from both sources, results in 118 breathing rate pairs. The average of the signed difference of the nurse breathing rate minus the RESpeck breathing rate over all pairs is -0.9 bpm. The RESpeck breathing rates are on average 0.9 bpm higher. A hypothesis test is used to assess the significance of this difference. The null and alternative hypothesis are chosen to be

$$\begin{aligned} \textbf{Null hypothesis}: &\quad P(\text{average signed difference} > 0) = 0.5 \\ \textbf{Alternative hypothesis}: &\quad P(\text{average signed difference} > 0) < 0.5 \end{aligned}$$

The hypothesis test is one-sided. The null hypothesis will be rejected if the p-value is smaller than $\alpha = 0.05$. Using bootstrapping, we resample $10,000$ times with replacement from the set of breathing rate pairs. This results in an p-value of 0.002 which



(a) QIXX09

(b) QIXX26

Figure 4.7: Qualitative description of overlap of RESpeck and nurse data

is smaller than $\alpha = 0.05$, so less than 5% of the sampling distribution has an average signed difference of more than 0. Hence, the null hypothesis can be rejected. The RESpeck breathing rates are likely to be higher than the nurse breathing rates in general.

This result is supported by similar findings analysing the breathing rates as measured by nurses. As mentioned in Section 2.2, the measured breathing rates are dependent on the monitoring period. More specifically, [23, 1] found that short monitoring periods lead to lower breathing rate estimates in manual measurements. Switching from 60s to 30s counting periods lead to a reduction of 0.46 or 0.95 bpm and a change from 60s to 15s lead to a reduction of 1.22 or 2.19 bpm in [1] and [23] respectively. The average difference between nurse and RESpeck breathing rates is close to the change in breathing rates obtained by moving from 60s to 30s count periods. This fits in with the observation in 4.1.2 that the nurse data contains a disproportionally high amount of breathing rates that are divisible by 2.

4.3.3 Information content

The RESpeck and nurses sample data at very different frequencies. To compare them, only hours for which there is data from both sources will be considered for now. If the RESpeck were to take measurements consistently throughout one hour, it would collect 45.000 values because it works at 12.5Hz. In practice, the RESpeck does not sample without interruption and during the preprocessing, less reliable data was removed (see Section 4.2), so there is less data in each hour. This results in about 10.300 values per hour on average. In contrast, a majority (55 of 118) of shared hours only contains one measurement taken by nurses. During some hours, more than 5 measurements were collected with a maximum of 8 in one hour. The average is 2.3 measurements per hour.

Does the amount of RESpeck data imply that it contains more information than the less frequently sampled nurse data? In information theory, information content, or Shannon

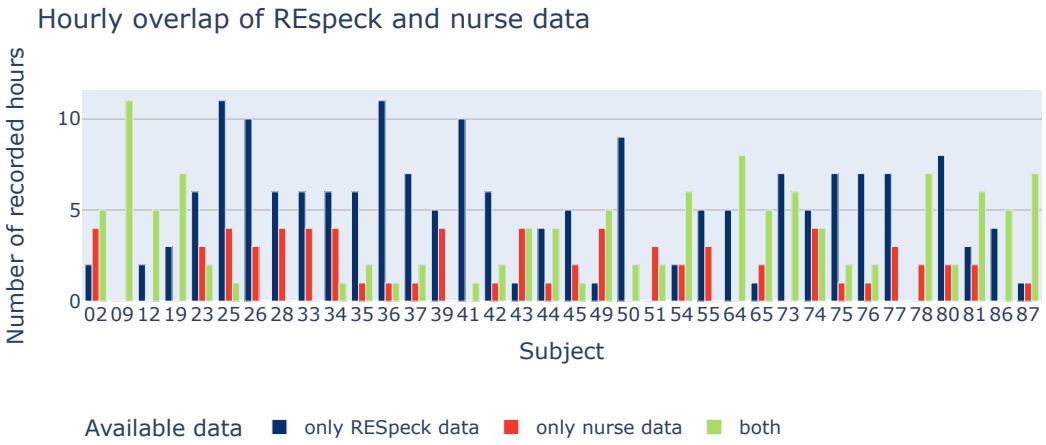


Figure 4.8: Quantitative description of overlap of RESpeck and nurse data

entropy, describes how difficult it is to predict the value of a variable. In other words, it measures the level of surprise or uncertainty of this variable. Shannon entropy is calculated using the frequency of values. Because RESpeck breathing rates usually have many decimal places, the RESpeck breathing rates were rounded to integers to measure the frequency of values. Using the natural logarithm, the entropy of the RESpeck breathing rates is 2.75 nats and the entropy of the nurse breathing rates is 2.34 nats. A hypothesis test with the following hypotheses,

$$\textbf{Null hypothesis:} \quad P(\text{RESpeck entropy} < \text{nurse entropy}) = 0.5,$$
$$\textbf{Alternative hypothesis:} \quad P(\text{RESpeck entropy} < \text{nurse entropy}) < 0.5,$$

using bootstrapping with 10,000 samples returns a p-value of 0.017. Hence, the null hypothesis can be rejected with $\alpha = 0.05$, so in general the RESpeck data entropy is likely to be higher than the nurse data entropy. This implies that the RESpeck data contains more information. However, nat is a difficult unit to interpret.

To compare how much more information the RESpeck breathing rates contain, mutual information and the uncertainty coefficient are calculated. The mutual information quantifies how much the knowledge of the value of one variable reveals about the value of the other variable. If the mutual information is close to 0, the variables are mostly independent. The larger the mutual information, the more one variable reveals about the other. Mutual information values can go up to infinity and are not easy to interpret. However, the uncertainty coefficient, which combines the mutual information and one variable's entropy, explains what fraction of this variable can be predicted given the other variable.

For calculating the mutual information and uncertainty coefficient, the RESpeck and nurse breathing rates were paired together. To avoid using averaged RESpeck values, they were paired by choosing the closest nurse datapoint for each RESpeck datapoint with a maximum distance of 1s, 30s, 1min, 10min, 30min, 45min or 60min. Table 4.1 shows the number of pairs, the mutual information (MI), and the uncertainty coefficient (UC) given the RESpeck or nurse breathing rates for each maximum distance. For a

Max. distance	Number of pairs	MI	UC	
			Given RESpeck data	Given nurse data
1s	27	1.32	62%	50%
30s	706	0.30	14%	11%
1min	1361	0.24	11%	9%
10min	11047	0.14	7%	5%
30min	22791	0.12	6%	4%
45min	26497	0.12	5%	4%
60min	29481	0.11	5%	4%

Table 4.1: Mutual information and uncertainty coefficient for RESpeck and nurse breathing rates considering pairs of different maximum time distances

distance of maximal 1s, there are only 27 pairs, so this will not be considered further. For the remaining distances there are enough value pairs to continue the analysis. As would be expected, the mutual information decreases as more distance is allowed between samples. This translates into decreasing uncertainty coefficients for increasing time distances. Surprisingly, the uncertainty coefficients are indicating that knowing about the breathing rate as measured by the RESpeck or nurses does not reveal much about the value measured by the other. At first glance, this seems to cause a contradiction with the relatively low average signed distance between RESpeck and nurse breathing rates of -0.9 bpm found in Section 4.3.2. However, it is important to note that hourly averages of the RESpeck and nurse breathing rates were used in that section. Here, the original breathing rates were used. The slightly higher uncertainty coefficients for predicting the nurse data given the RESpeck data in comparison to the uncertainty coefficients for predicting the RESpeck data given the nurse data indicate that the RESpeck breathing rates reveal more about the nurse breathing rate than the other way around.

Chapter 5

Conclusions

Based on the data of 13 patients, the NHS Borders dataset was investigated. During data exploration, it became apparent that the exercises were grouped into sessions inconsistently and that the CAT scores for some patients contained outliers or were too few. As a consequence, exercises were considered individually, outlier CAT scores were removed and patients with only a few scores were excluded from the analysis. The correlations of various features with the CAT score were calculated and possible explanations for these were discussed. The strongest correlation with the CAT score was identified with the `PR_STEP_UPS_avgBreathingRate` feature ($\rho = 0.52$, $p = 2 \times 10^6$ for the same day and $\rho = 0.38$, $p = 0.02$ for the previous day). The PR_SIT_TO_STAND, PR_SHOULDER_PRESS and PR_LEG_SLIDE exercises were generally very indicative of the CAT score on the same and the following day. Filtering the RESpeck data for reliable data was found to result in features that are more strongly correlated with the CAT score. For CAT score association and CAT score prediction, the best model has an MAE of 3.0 and 3.3, an RMSE of 4.1 and 4.5 and an accuracy of 75% and 73% respectively. The CAT score prediction model's performance is similar to what was achieved in [29] and better than what was achieved in [30]. This is the first time that CAT score association was tried for RESpeck data.

The initial hypothesis that the data collected by the RESpeck device and by recording Rehab exercises during a specific day can be used to estimate a patient's CAT score for that day and to predict the score for the following day is confirmed by this analysis. The CAT score estimation is not perfect but it is significantly better than the baseline model and relatively close to the true score. The improved performance of the final model that was trained on the entire training set suggests that the model's performance will be even better for even more data. Furthermore, it was discovered how important it is to optimise models according to an error metric that makes sense with the model architecture. If the aim is for model estimates to be as close as possible to the true scores, a regression model optimised according to RMSE is reasonable. If the aim is to get high accuracy with respect to the impact level, a classifier optimised according to accuracy is more reasonable.

The QIP data was analysed for 36 patients. Before comparing the breathing rates measured by the RESpeck with those measured by the nurse, the RESpeck data was

filtered for reliable data. To increase the amount of reliable data without loosening the restriction of what concerns reliable data, various parameter settings of the classification model were investigated. Considering the data with 20s sliding windows that overlap by 80% was shown to be the best approach. For most of the analysis, the breathing rates measured by the RESpeck and those measured by nurses were compared on an hourly basis for hours for which there is data from both sources. It was shown that the average RESpeck breathing rates are on average 0.9 bpm higher than the average nurse breathing rates. This coincides with the finding that nurses were likely to have measured some breathing rates during 15s and 30s monitoring periods and studies showing that shorter monitoring periods lead to lower estimates [1, 23]. This seems to indicate that the RESpeck's breathing rates are closer to the true breathing rates and more reliable. Using methods from information theory, it was shown that the information content of the RESpeck breathing rates is higher than the information content of the nurse breathing rates. However, the uncertainty coefficient of unaveraged breathing rates indicates that the breathing rates measured by the RESpeck and those measured by the nurses do not give much information about each other. This puts the validity of unaveraged RESpeck measurements into question.

The hypothesis that breathing rates collected by the RESpeck are more reliable and informative than those measured by nurses could neither really be confirmed nor rejected. Overall, it was shown that the hourly averages of RESpeck breathing rates are close to the hourly averages of nurse breathing rates so the averaged RESpeck measurements are at least as good as the nurse breathing rates. Considering the time pressure and workload under which nurses usually work, a device that can automatically collect reliable hourly averaged breathing rates is an improvement.

5.1 Future works

For the NHS Borders analysis, more features could be considered. From the RESpeck data, only the breathing rate and activity level were considered in this project. A closer look at the breathing signal might provide useful features. For the histogram-based gradient boosting regression tree, only the default parameters were used. Having a closer look at the parameters and adjusting them as necessary might improve the model significantly. Furthermore, including more data would improve the model's performance as well.

For the QIP analysis, future projects might want to further investigate the data from an information theory point of view. In particular, the differences in using averaged and raw RESpeck measurements would be worthwhile to investigate and what averaging intervals work best. For the classification model that analyses the reliability of RESpeck data, the use of varying window sizes was discussed. Because it is important to use the same window size during training and later usages, the original window size of 20s was kept here. However, the law of large numbers would suggest that it is better to consider larger windows for assessing the reliability of the data. Hence, it might be of value to train the reliability model on the original data for varying window sizes, in particular because [28] does not explain why a window size of 20s was used.

Bibliography

[1] M Rimbi et al. *Respiratory rates observed over 15 and 30 s compared with rates measured over 60 s: practice-based evidence from an observational study of acutely ill adult medical patients during hospital admission*. QJM: An International Journal of Medicine. 2019. URL: https://academic.oup.com/qjmed/article/112/7/513/5393268 (visited on 04/03/2023).

[2] Michelle A. Cretikos et al. *Respiratory rate: the neglected vital sign*. 2008. URL: https://www.mja.com.au/system/files/issues/188_11_020608/cre11027_fm.pdf (visited on 09/04/2023).

[3] Global Initiative for Chronic Obstructive Lunge Disease. *2023 report*. 2023. URL: https://goldcopd.org/2023-gold-report-2/ (visited on 10/02/2023).

[4] NHS. *Overview - Chronic obstructive pulmonary disease (COPD)*. 2019. URL: https://www.nhs.uk/conditions/chronic-obstructive-pulmonary-disease-copd/ (visited on 10/02/2023).

[5] Avani R. Patel et al. *Global Initiative for Chronic Obstructive Lung Disease: The Changes Made*. Cureus. 2019.

[6] World Health Organization. *World health statistics 2008*. ISBN 9789240682740. 2008.

[7] NHS. *Spirometry*. 2021. URL: https://www.nhs.uk/conditions/spirometry/ (visited on 21/01/2023).

[8] Ioanna G Tsiligianni et al. *Assessing health status in COPD. A head-to-head comparison between the COPD assessment test (CAT) and the clinical COPD questionnaire (CCQ)*. BMC Pulmonary Medicine. 2012. URL: https://bmcpulmmed.biomedcentral.com/articles/10.1186/1471-2466-12-20 (visited on 10/02/2023).

[9] Andrew Cave and Ioanna Tsiligianni. *IPCRG Users' Guide to COPD "Wellness" Tools*. IPCRG. 2010. URL: https://www.ipcrg.org/sites/ipcrg/files/content/attachments/2019-10-23/ipcrg_users_guide_to_copd_wellness_tools.pdf (visited on 10/02/2023).

[10] *How is your COPD? Take the COPD Assessment Test™ (CAT)*. URL: https://www.catestonline.org/patient-site-test-page-english.html%20Last%20accessed:%2025th%20Jan%202023 (visited on 25/01/2023).

[11] Mike Polkey, Claus Vogelmeier and Mark Dransfield. *COPD Assessment Test. User Guide*. 2022. URL: https://www.catestonline.org/hcp-homepage/research.html (visited on 24/01/2023).

[12] Paul W Jones, Margaret Tabberer and Wen-Hung Chen. *Creating scenarios of the impact of COPD and their relationship to COPD assessment test (CAT™) scores*. BMC Pulm Med. 2011.

[13] Nisha Gupta et al. *The COPD assessment test: a systematic review*. European Respiratory Journal. 2014. URL: https://erj.ersjournals.com/content/44/4/873.short (visited on 10/02/2023).

[14] Roman Jaeschke, Joel Singer and Gordon H. Guyatt. *Measurement of health status: Ascertaining the minimal clinically important difference*. Controlled clinical trials. 1989. URL: https://www.sciencedirect.com/science/article/pii/0197245689900056 (visited on 10/02/2023).

[15] Harma Alma et al. *Clinically relevant differences in COPD health status: systematic review and triangulation*. European Respiratory Journal. 2018. URL: https://erj.ersjournals.com/content/52/3/1800412.short (visited on 10/02/2023).

[16] Spruit MA et al. *An official American Thoracic Society/European Respiratory Society statement: key concepts and advances in pulmonary rehabilitation*. Am J Respir Crit Care Med. 2013. URL: https://pubmed.ncbi.nlm.nih.gov/24127811/ (visited on 18/02/2023).

[17] National Institute for Health and Care Excellence. *Chronic obstructive pulmonary disease in over 16s: diagnosis and management*. 2018. URL: https://www.nice.org.uk/guidance/ng115/ (visited on 18/02/2023).

[18] British Thoracic Society Pulmonary Rehabilitation Guideline Group. *BTS Guideline on Pulmonary Rehabilitation in Adults*. Thorax. 2013. URL: https://pubmed.ncbi.nlm.nih.gov/23880483/ (visited on 18/02/2023).

[19] Alison JA et al. *Australian and New Zealand Pulmonary Rehabilitation Guidelines*. Respirology. 2017. URL: https://pubmed.ncbi.nlm.nih.gov/28339144/ (visited on 18/02/2023).

[20] Aroub Lahham and Anne E. Holland. *The Need for Expanding Pulmonary Rehabilitation Services*. 2021. URL: https://www.mdpi.com/2075-1729/11/11/1236 (visited on 18/02/2023).

[21] Andrew Keating, Annemarie Lee and Anne E Holland. *What prevents people with chronic obstructive pulmonary disease from attending pulmonary rehabilitation? A systematic review*. Chronic respiratory disease. 2011. URL: https://journals.sagepub.com/doi/pdf/10.1177/1479972310393756 (visited on 18/02/2023).

[22] Md. Nazim Uzzaman et al. *Effectiveness of home-based pulmonary rehabilitation: systematic review and meta-analysis*. European Respiratory Review. 2022. URL: https://err.ersjournals.com/content/31/165/220076 (visited on 18/02/2023).

[23] Andrew Hill et al. *The effects of awareness and count duration on adult respiratory rate measurements: An experimental study*. 2017. URL: https://doi.org/10.1111/jocn.13861 (visited on 07/03/2023).

[24] Haipeng Liu et al. *Recent development of respiratory rate measurement technologies*. Physiological Measurement. 2019. URL: https://iopscience.iop.org/article/10.1088/1361-6579/ab299e/meta (visited on 11/02/2023).

[25] Desen Cao et al. *Application of a Wearable Physiological Monitoring System in Pulmonary Respiratory Rehabilitation Research*. 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). 2018. URL: https://ieeexplore.ieee.org/abstract/document/8633113 (visited on 11/02/2023).

[26] Christian Peter Subbe and Sean Kinsella. *Continuous Monitoring of Respiratory Rate in Emergency Admissions: Evaluation of the RespiraSense™ Sensor in Acute Care Compared to the Industry Standard and Gold Standard*. MDPI. 2018. URL: https://www.mdpi.com/1424-8220/18/8/2700 (visited on 11/02/2023).

[27] D. K. Arvind et al. *Home-Based Pulmonary Rehabilitation of COPD Individuals Using the Wearable Respeck Monitor*. EAI International Conference on Body Area Networks. 2022. URL: https://link.springer.com/chapter/10.1007/978-3-030-95593-9_15 (visited on 11/02/2023).

[28] Gordon B. Drummond et al. *Classifying signals from a wearable accelerometer device to measure respiratory rate*. 2021. URL: https://openres.ersjournals.com/content/7/2/00681-2020.short (visited on 09/04/2023).

[29] Darius Fischer. *Predicting the well-being of COPD patients with respiratory data from pulmonary rehabilitation*. University of Edinburgh. 2016. URL: https://project-archive.inf.ed.ac.uk/msc/20161889/msc_proj.pdf (visited on 11/02/2023).

[30] Ioana Mihailescu. *Rehab3 App Augmented with Breathing Exercises and the Analysis of Remote Pulmonary Rehabilitation Data from NHS Borders Patients*. University of Edinburgh. 2022.

[31] Andrew Bates et al. *Respiratory Rate and Flow Waveform Estimation from Tri-axial Accelerometer Data*. 2010 International Conference on Body Sensor Networks. 2010. URL: https://ieeexplore.ieee.org/abstract/document/5504743?casa_token=ezHux42L6Q0AAAAA:Ps9Junf8PtYcrT12N-mI-lWCxJgWCaBrynQHPK38m-Bd_07fabVf_v7-bTDH6QADD6UIS6OetQ (visited on 10/02/2023).

[32] Jason Brownlee. *Gradient Boosting with Scikit-Learn, XGBoost, LightGBM, and CatBoost*. 2020. URL: https://machinelearningmastery.com/gradient-boosting-with-scikit-learn-xgboost-lightgbm-and-catboost/ (visited on 31/03/2023).

[33] Wikipedia. *Glasgow Coma Scale*. 2023. URL: https://en.wikipedia.org/wiki/Glasgow_Coma_Scale (visited on 04/03/2023).

[34] Wikipedia. *Oxygen saturation (medicine)*. 2023. URL: https://en.wikipedia.org/wiki/Oxygen_saturation_(medicine) (visited on 04/03/2023).

[35] Royal College of Physicians. *National Early Warning Score (NEWS) 2. Standardising the assessment of acute-illness severity in the NHS. Executive summary and recommendations*. 2017. URL: https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2 (visited on 04/03/2023).

Appendix A

Study approval letter for NHS Borders dataset

NHS Borders

Research, Development & Innovation
Clinical Governance & Quality

Clinical Governance & Quality
Borders General Hospital
Melrose
Roxburghshire TD6 9BS

Telephone 01896 826719
Fax 01896 826040
www.nhsborders.org.uk

Professor D. K. Arvind
Centre of Speckled Computing
School of Informatics
University of Edinburgh
10 Crichton Street
Edinburgh
EH8 9AB

Date 13 April 2021

Our Ref 21/BORD/IN01

Enquiries to Joy Dawson
Extension 01896 826717
Email research.governance@borders.scot.nhs.uk

Dear Professor Arvind

Study Reference Number: 21/BORD/IN01

Study Title: Remote Monitoring and Pulmonary Rehabilitation of COPD (and COVID-19 recovered) Patients in the NHS Borders Region

Thank you for submitting your innovation project to NHS Borders. Your application has been reviewed and I am happy to confirm that NHS Borders has agreed to participate as a test bed site for the study.

Conditions of Approval

- NHS Borders will evaluate the use of the Respeck device in the home and agrees to approach patients to help to test the device and app.
- The University of Edinburgh will provide the necessary devices to test the innovation.
- Patients approached to participate will provide consent to providing data for the project; however no identifiable data will be shared with University of Edinburgh.
- NHS Borders will support the project by providing clinical advice as to how the app can help support the clinical care pathway for the management of COPD
- NHS Borders pulmonary rehab team have confirmed that they have capacity to support this project and that this can be undertaken without existing funding support
- A summary report should be submitted by email to research.governance@borders.scot.nhs.uk within 2 months of the project ending
- The support of NHS Borders Pulmonary Rehab team will be acknowledged in any presentations and publications.

May I wish you success with your project. Should you have any queries in relation to this letter please contact Joy Dawson on 01896 826717 or research.governance@borders.scot.nhs.uk.

Yours Sincerely

Joy Dawson
Research Governance Manager & NHS Borders Innovation Champion

Appendix B

Exercises

Exercise ID	Description
0	PR_SIT_TO_STAND
1	PR_KNEE_EXTENSION
2	PR_SQUATS
3	PR_HEEL_RAISES
4	PR_BICEP_CURL
5	PR_SHOULDER_PRESS
6	PR_WALL_PUSH
7	PR_LEG_SLIDE
8	PR_STEP_UPS
9	PR_WALKING

Table B.1: Exercise ID interpretation

Appendix C

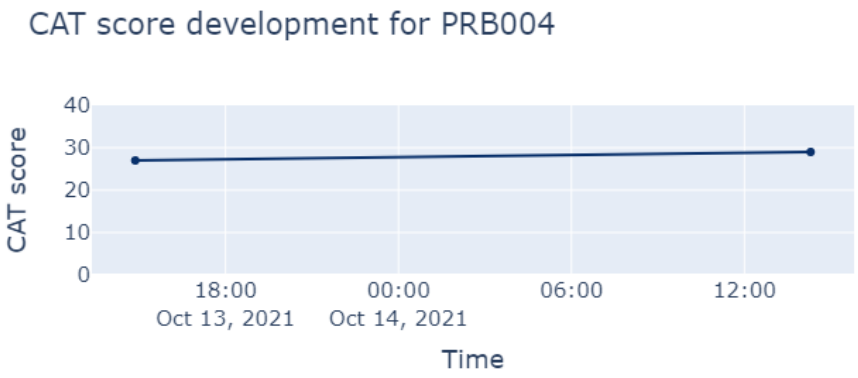
NHS Borders data: available data

Patient id	How much data is available?			Included
	RESpeck	Rehab	CAT	in Analysis
PRB001	17M datapoints (15.8 days)	194 exercises	0 scores	No
PRB002	0 datapoints	0 exercises	0 scores	No
PRB003	45M datapoints (41.9 days)	125 exercises	17 scores	Yes
PRB004	2M datapoints (2.0 days)	1 exercise	2 scores	No
PRB005	35M datapoints (32.0 days)	256 exercises	2 scores	No
PRB006	23M datapoints (21.1 days)	40 exercises	13 scores	Yes
PRB007	32M datapoints (29.6 days)	178 exercises	19 scores	Yes
PRB008	only minute-averaged	incomplete	incomplete	No
PRB102	60M datapoints (55.5 days)	99 exercises	3 scores	No
PRB103	39M datapoints (35.8 days)	253 exercises	33 scores	No
PRB104	57M datapoints (52.5 days)	43 exercises	20 scores	Yes
PRB105	60M datapoints (55.5 days)	362 exercises	44 scores	Yes
PRB106	0.2M datapoints (0.2 days)	14 exercises	8 scores	No
PRB107	21M datapoints (19.7 days)	368 exercises	62 scores	Yes
PRB108	39M datapoints (35.7 days)	120 exercises	26 scores	Yes
PRB109	13M datapoints (12.3 days)	29 exercises	9 scores	Yes
PRB111	15M datapoints (13.8 days)	135 exercises	27 scores	Yes
PRB201	26M datapoints (23.7 days)	330 exercises	57 scores	Yes
PRB202	29M datapoints (27.0 days)	355 exercises	66 scores	Yes
PRB203	66M datapoints (61.0 days)	314 exercises	52 scores	Yes
PRX018	62M datapoints (57.3 days)	105 exercises	30 scores	Yes
PRX900	51M datapoints (47.3 days)	0 exercises	44 scores	No

Table C.1: Summary of NHS Borders data exploration

Appendix D

NHS Borders data: CAT score development

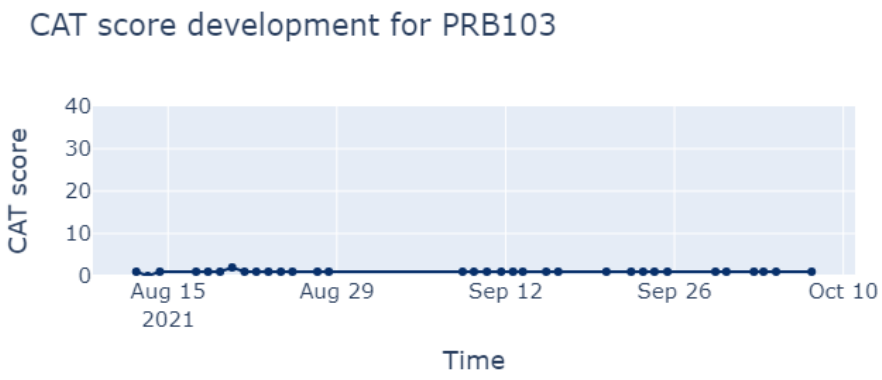



CAT score development for PRB005
CAT score
40
30
20
10
0
Apr 6
2022
Apr 9
Apr 12
Apr 15
Apr 18
Apr 21
Time


CAT score development for PRB006
CAT score
40
30
20
10
0
Apr 24
2022
May 1
May 8
May 15
May 22
Time


CAT score development for PRB007
CAT score
40
30
20
10
0
Jul 3
2022
Jul 10
Jul 17
Jul 24
Jul 31
Aug 7
Aug 14
Aug 21
Time


CAT score development for PRB102
CAT score
40
30
20
10
0
Jun 13
2021
Jun 20
Jun 27
Jul 4
Jul 11
Jul 18
Jul 25
Aug 1
Time

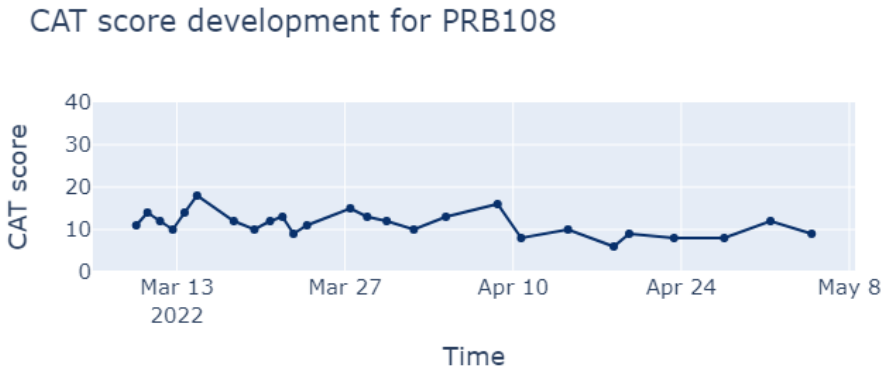

CAT score development for PRB103
CAT score
40
30
20
10
0
Aug 15
2021
Aug 29
Sep 12
Sep 26
Oct 10
Time


CAT score development for PRB104
CAT score
40
30
20
10
0
Aug 29
2021
Sep 12
Sep 26
Oct 10
Time


CAT score development for PRB105
CAT score
40
30
20
10
0
Sep 26
2021
Oct 10
Oct 24
Nov 7
Nov 21
Dec 5
Time


CAT score development for PRB106
CAT score
40
30
20
10
0
Oct 11
2021
Oct 14
Oct 17
Oct 20
Oct 23
Oct 26
Oct 29
Time


CAT score development for PRB107
CAT score
40
30
20
10
0
Mar 13
2022
Mar 27
Apr 10
Apr 24
May 8
May 22
Jun 5
Time


CAT score development for PRB108
CAT score
40
30
20
10
0
Mar 13
2022
Mar 27
Apr 10
Apr 24
May 8
Time

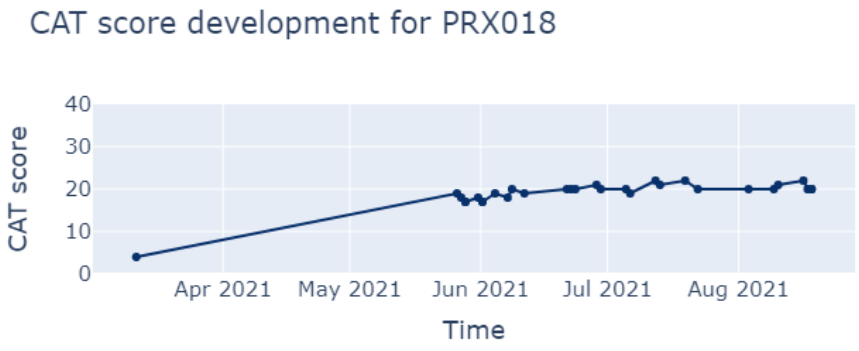

CAT score development for PRB109
CAT score
40
30
20
10
0
Sep 3
2022
Sep 6
Sep 9
Sep 12
Sep 15
Sep 18
Sep 21
Time


CAT score development for PRB111
CAT score
40
30
20
10
0
Oct 9
2022
Oct 23
Nov 6
Nov 20
Time


CAT score development for PRB201
CAT score
40
30
20
10
0
Jul 4
2021
Jul 18
Aug 1
Aug 15
Time


CAT score development for PRB202
CAT score
40
30
20
10
0
May 22
2022
Jun 5
Jun 19
Jul 3
Jul 17
Jul 31
Aug 14
Time


CAT score development for PRB203
CAT score
40
30
20
10
0
Jun 5
2022
Jun 19
Jul 3
Jul 17
Jul 31
Aug 14
Time

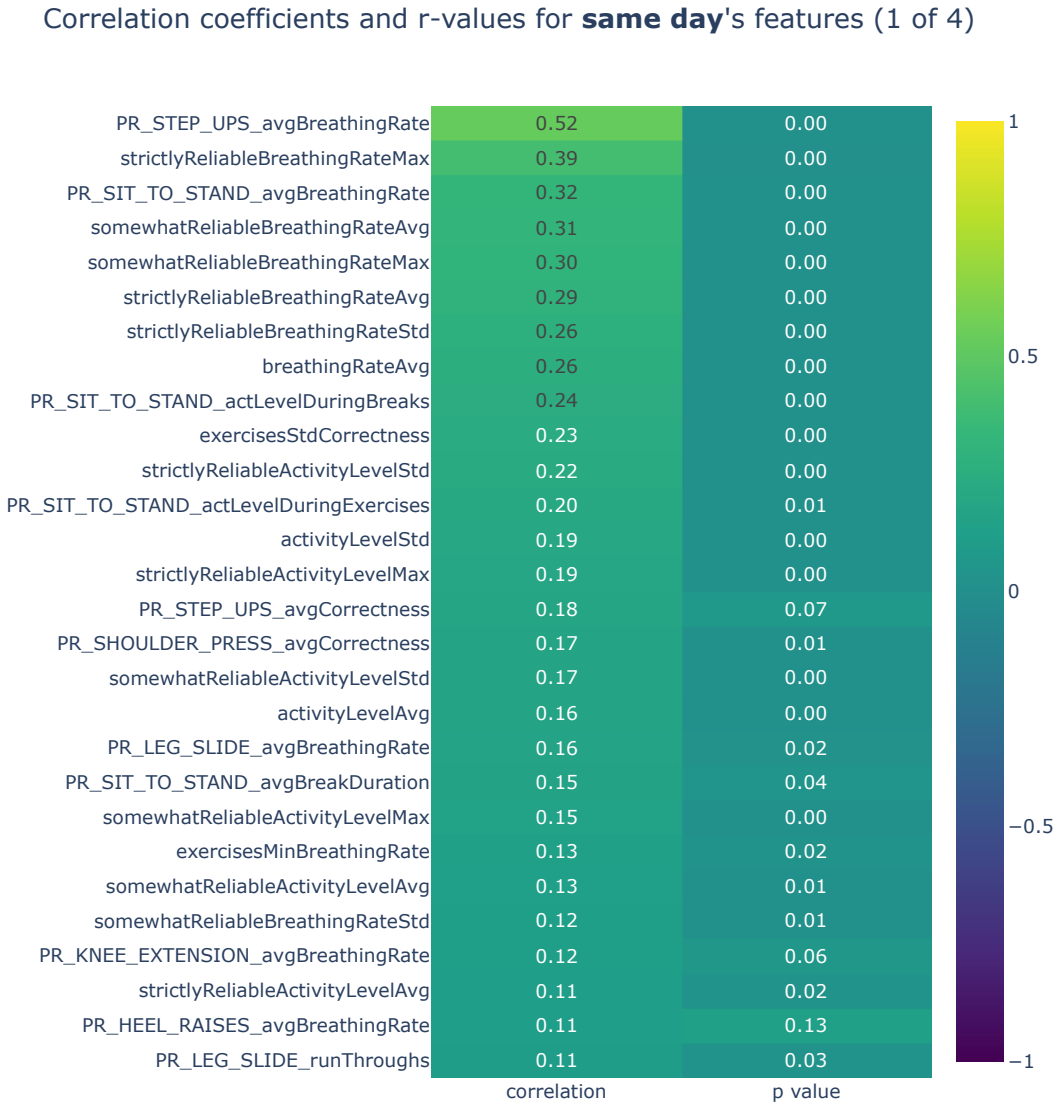

CAT score development for PRX018
CAT score
40
30
20
10
0
Apr 2021
May 2021
Jun 2021
Jul 2021
Aug 2021
Time


CAT score development for PRX900
CAT score
40
30
20
10
0
Oct 10
2021
Oct 24
Nov 7
Nov 21
Time

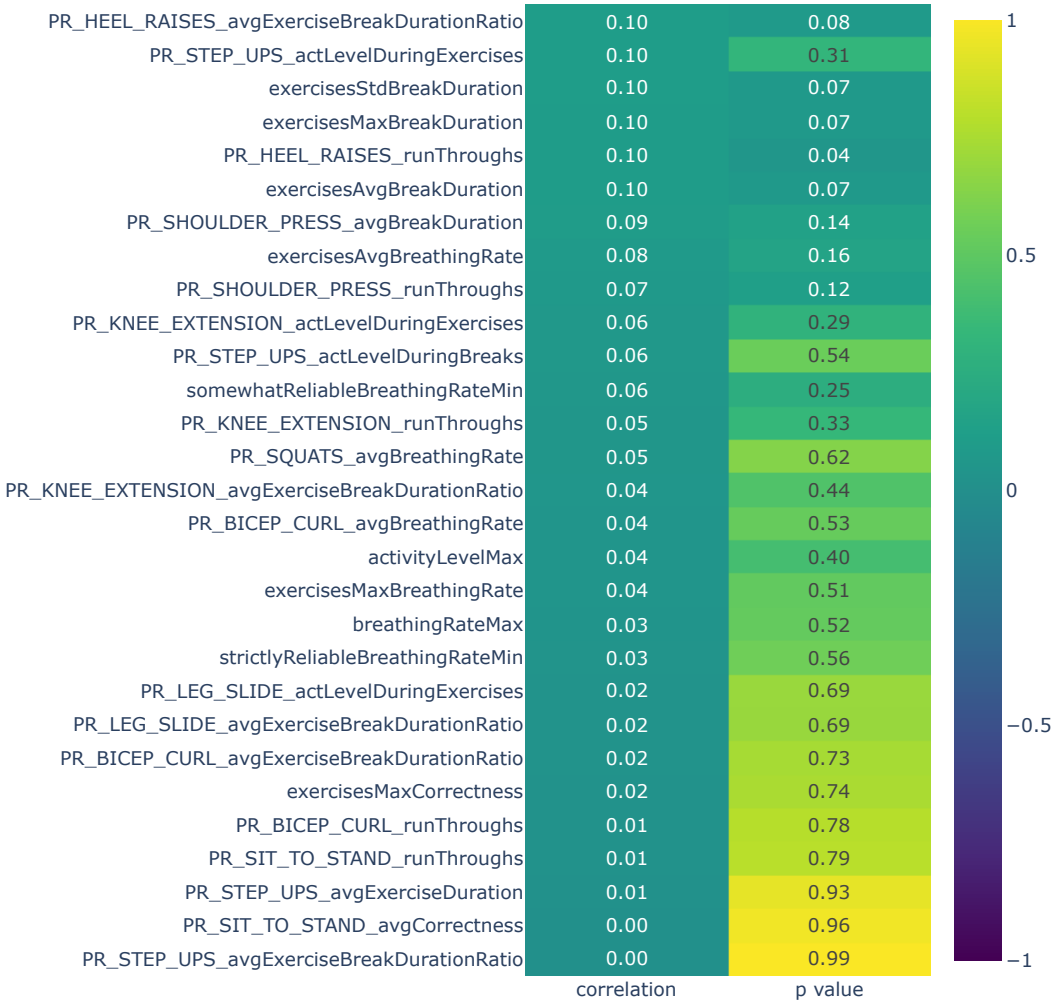
Appendix E

NHS Borders data: correlations

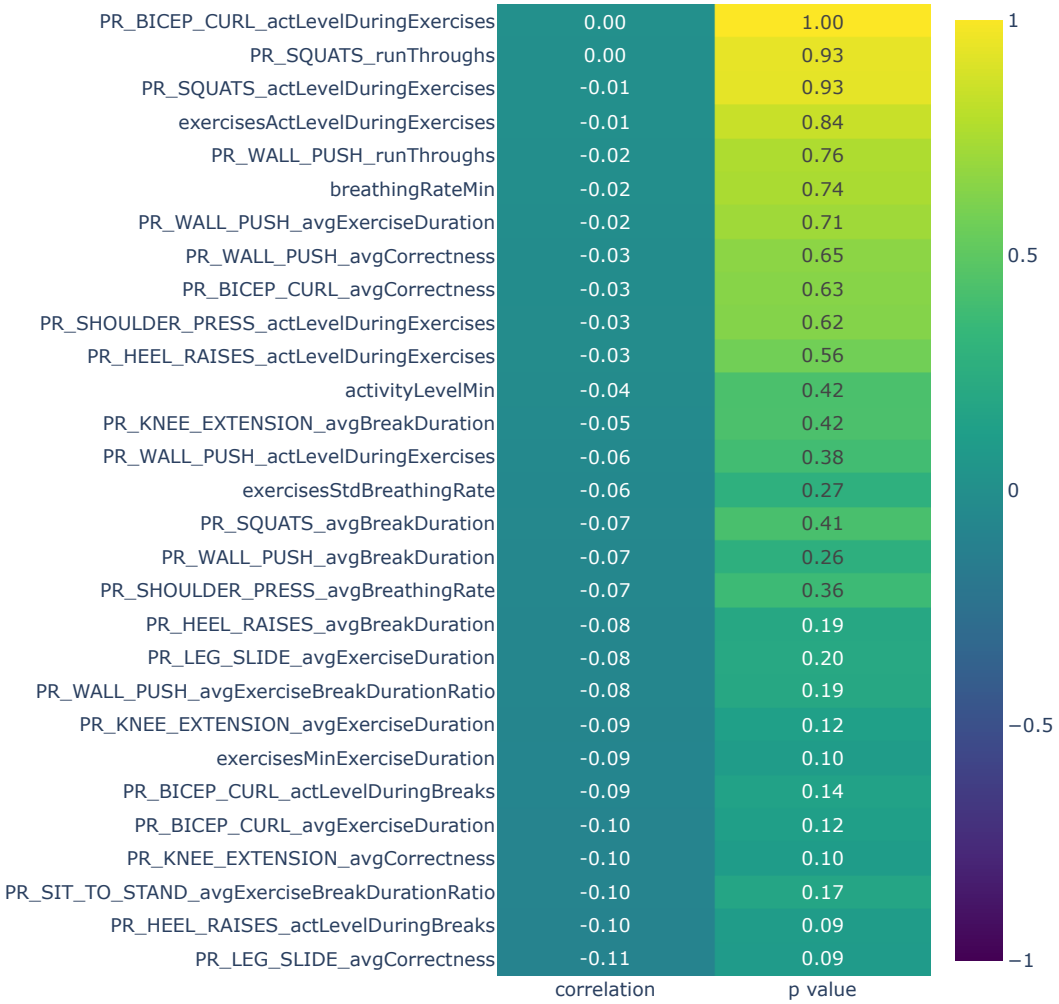
E.1 Correlations for features of the same day



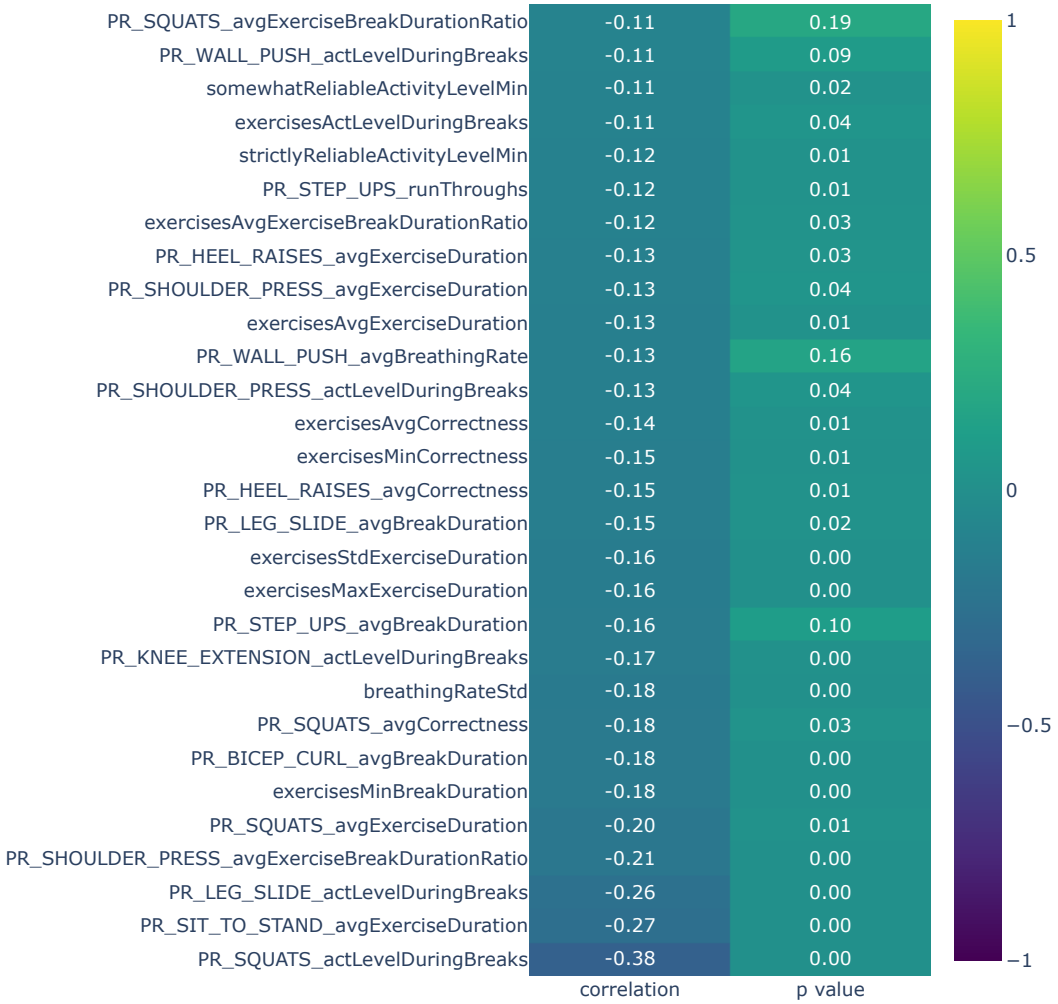
Correlation coefficients and r-values for **same day**'s features (2 of 4)



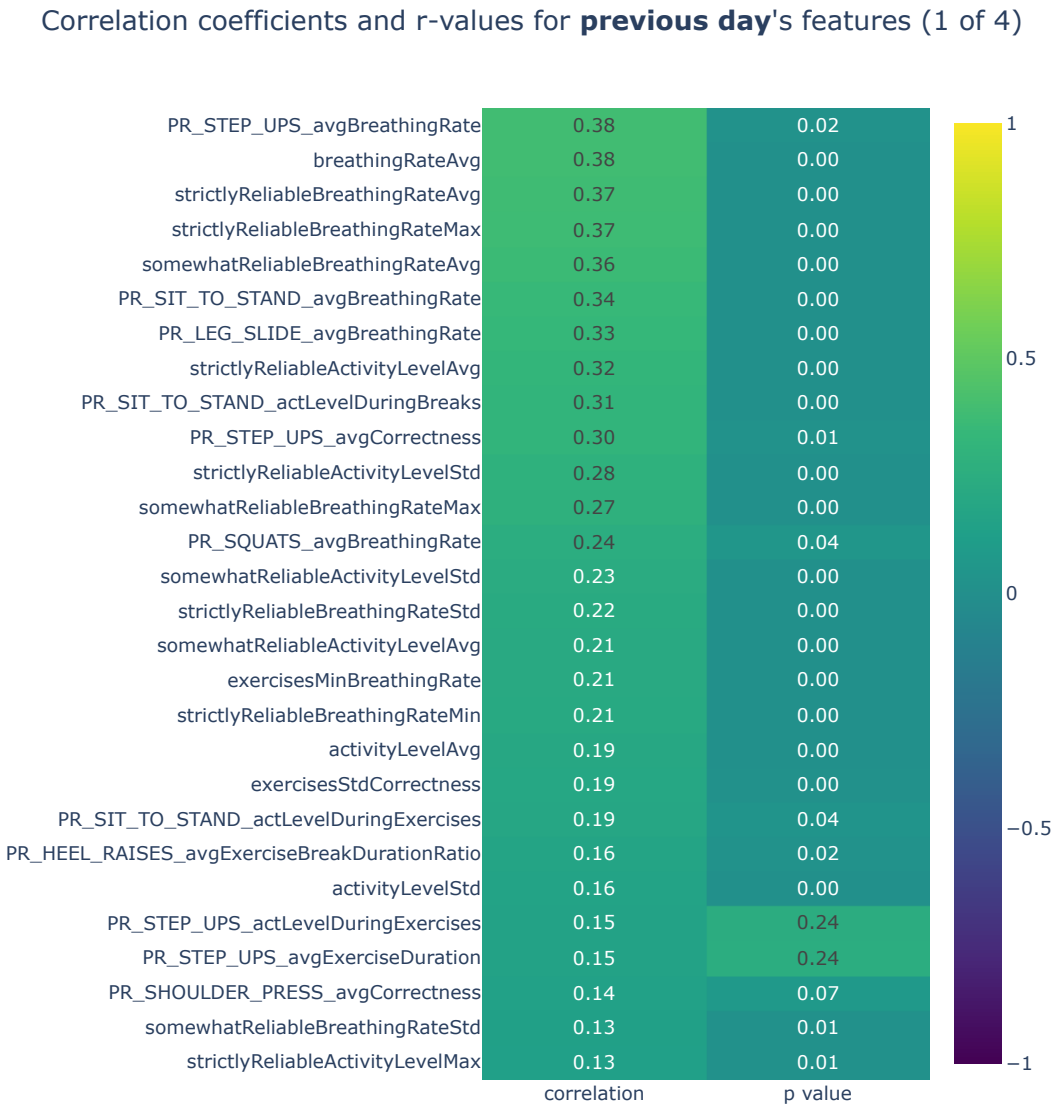
Correlation coefficients and r-values for **same day**'s features (3 of 4)



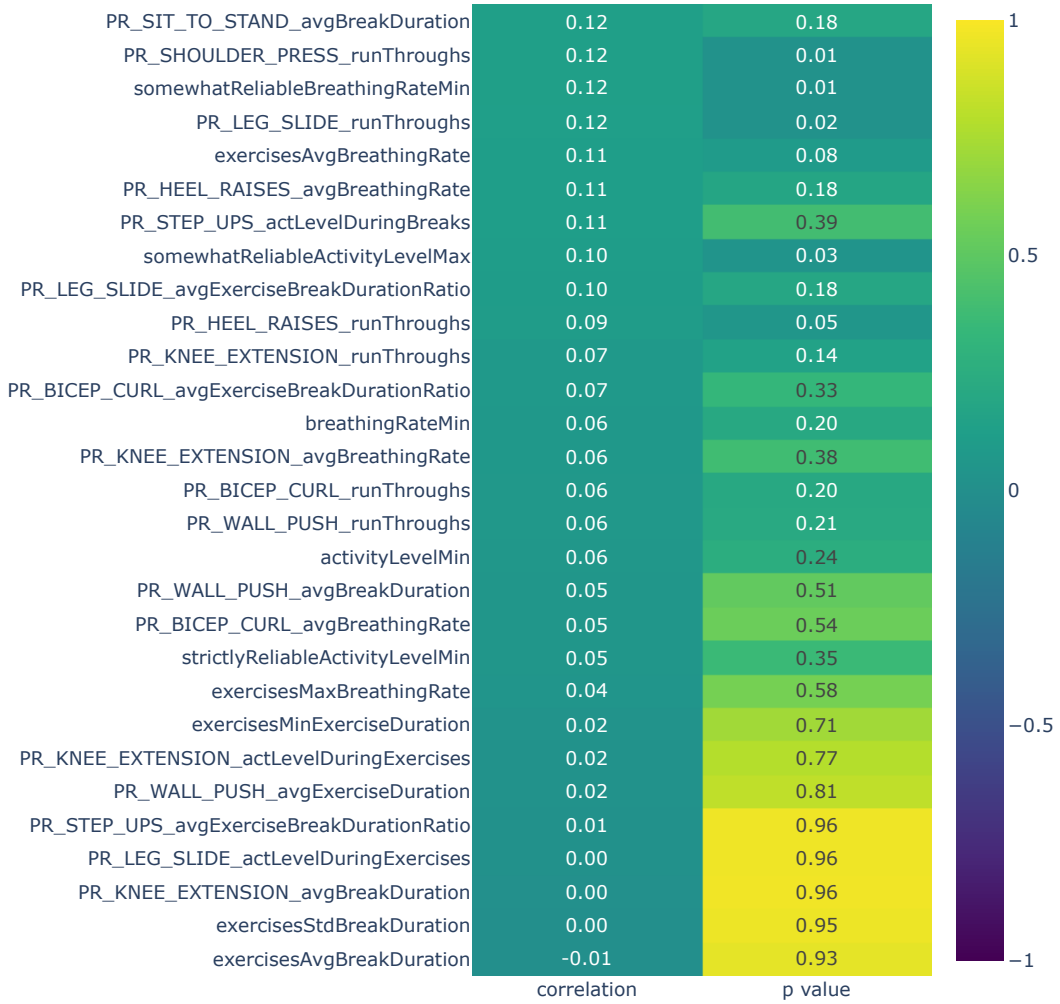
Correlation coefficients and r-values for **same day**'s features (4 of 4)



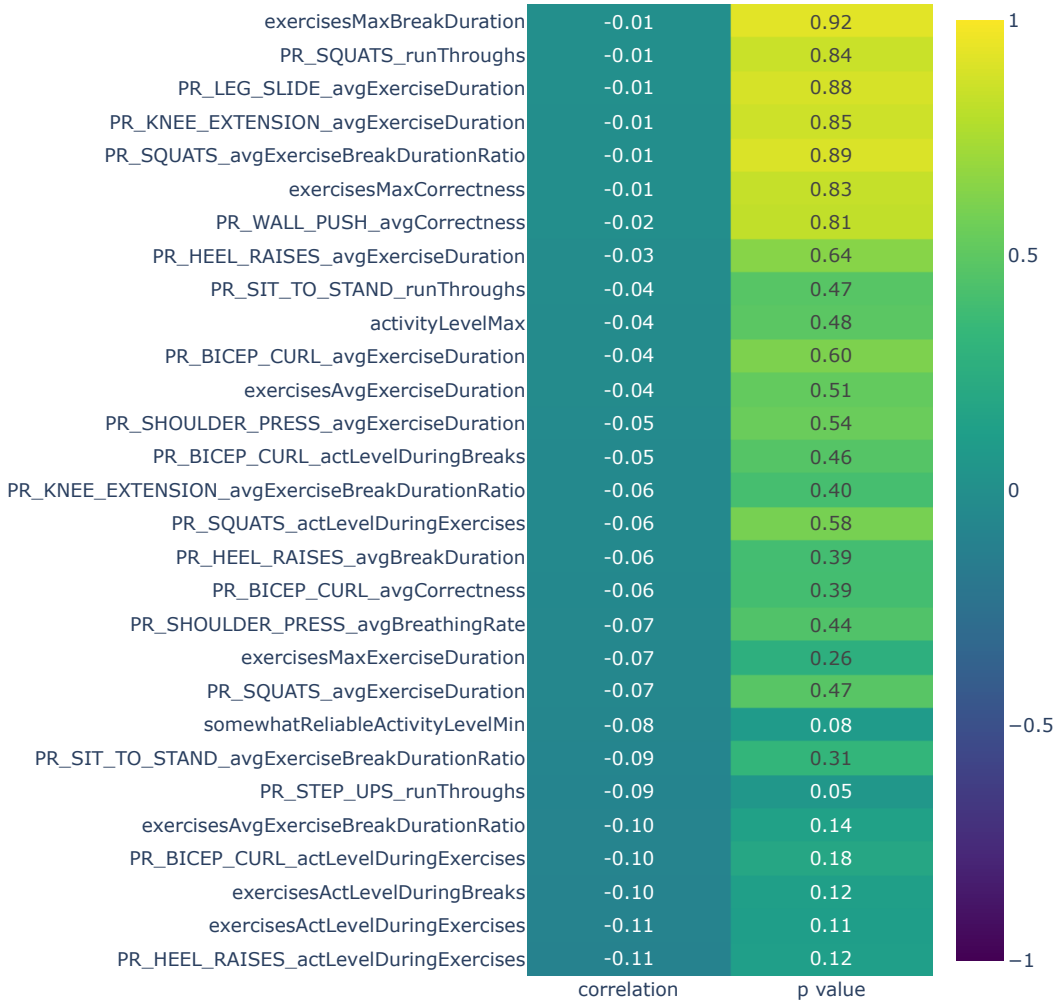
E.2 Correlations for features of the previous day



Correlation coefficients and r-values for **previous day**'s features (2 of 4)



Correlation coefficients and r-values for **previous day**'s features (3 of 4)



Correlation coefficients and r-values for **previous day**'s features (4 of 4)