PREDICTING WELL-BEING WITH WEARABLE SENSOR DATA

Bachelor's Project Thesis

Desmond Drijfhout, s2392895, d.s.k.drijfhout@student.rug.nl

Supervisor: Dr M.K. van Vugt

Abstract: Introduction The global population is ageing rapidly. The medical industry faces challenges such as overburdened health-care delivery systems and growth in physician demand. One solution that could potentially help handle the ageing population more efficiently is the automatic prediction of well-being on the basis of different well-being predictors. This could allow for more targeted care for patients. **Method** To examine whether it is possible to develop automatic prediction of well-being, we gathered data on sleep, activity, and time away from home for 22 elderly participants living in The Netherlands for a period of 10 days, using a medical sensor watch. In addition, we gathered self-reported assessments of well-being from the participants as well as nurse evaluations on the participants' well-being. We used and compared different types of classifiers to try to automatically predict well-being from the sensor data. We tried feature selection and different re-sampling techniques to improve the models. **Results** We show that, using the Random Forest classifier, both the self-reported assessments (after feature selection) and nurse evaluations (after re-sampling) can be predicted with respectively 94% accuracy (92.75% sensitivity) and 90.64% accuracy (90.14% sensitivity). Furthermore, we show that manually assessing the well-being of a patient solely based on looking at the sensor data is not a good predictor of the self-reported assessments nor the nurse evaluations. Last, we show potential problems, like overfitting, that we encountered collecting and evaluating the results.

1. Introduction

In 2015, the World Health Organization (WHO) reported that the global population is ageing rapidly (WHO, 2015). This brings several challenges on social and personal perspectives. Overburdened health-care delivery systems (Shrivastava et al., 2013), rising public expenditures for medical care (WHO, 2002a), and growth in demand for health workers (Lui et al., 2017). On a personal level, there is as well the desire of people to experience these extra years in good health (WHO, 2006).

One solution to the challenges that come with an ageing population falls within the automation of the recognition of well-being, such that care can be targeted better and more efficiently. If a computer can predict whether a person is well or not well, physicians will be able to plan their time more efficiently. For example, there will be no need for routine check-ups of people who are well anyway. Medical conditions can possibly be detected easier and earlier, providing better care to those who need it. Before looking into previous research and setting up a research question, we need to know how well-being is defined.

1.1. Well-being defined

Well-being is defined in the Constitution of the World Health Organization as "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" (WHO, 2006).

One of the ways to determine someone's well-being, is to ask a person how they would rate their own subjective well-being in terms of physical, mental and social health. One application that is used for measuring well-being using the three pillars physical, mental and social health, is the Patient Reported Outcomes Measured Information System (PROMIS). PROMIS was developed by the United States' National Institutes of Health to "standardize a common measurement system for patient-reported outcomes" (Cella et al., 2007). PROMIS was created by six different university institutes in the United States, who collaborated, each focussing on their own particular topic of well-being, to create a universally used system to measure well-being. We will see more of PROMIS later in this thesis.

Another way to determine someone's well-being would be to ask a professional (physician or

nurse) to examine a person objectively and rate the well-being based on the professional's methods.

A physician would arguably be more interested in predicting the objective evaluation of a person's well-being, although it is argued that subjective measurements of well-being should not be excluded in determining well-being of a person (Andrews, 1974 and Campbell, 1976).

Both the subjective well-being and objective well-being will be used in this paper respectively in the form of a questionnaire and in the form of a nurse's evaluation.

1.2. Predictors of well-being

The next question to ask would be: what are the predictors of a person's well-being? Many factors are known to be highly correlative with physical, mental and social well-being. Next are a few of those examples.

- **Sleep**. Having a regular sleep pattern is an important factor known to benefit a person's overall health and quality of life (Buysse, 2014).
- **Circadian rhythm.** Van Someren (2000) describes the circadian rhythm as "rhythms of about 24 hours that occur in most physiological processes and overt behaviour". Van Someren found the alterations in the circadian rhythm to strongly contribute to disturbed sleep patterns in elderly people. In another study of different case reports, Paavilainen et al. (2005) found changes in the circadian rhythm to be associated with changes in the actual clinically observed physical well-being of a patient.
- **Physical activity**. Anokye et al. (2012) reported a positive relationship between physical activity and the health-related quality of life.
- **Nutrition**. Keller (2004) reported a significant correlation between poor nutrition and the differences in health-related quality of life in frail adults.
- **Social interaction**. Results of the study by George et al. (1989) show that social support affects the outcome of depressive illness. Another pair of studies among depressed patients shows that depressive illness negatively impacts well-being (Hays et al., 1995; Wells et al., 1998).
- **Heart rate**. An elevated resting heart rate was found to be a risk factor for incident cardiovascular disease (Cooney et al., 2010), which can play a big role in physical well-being.

To collect the predictors and make a prediction about well-being without intruding too much in people's everyday life, passive sensors are required. In the next section, we will show examples of studies that have been able to predict physical, mental or social well-being using mobile or wearable sensors.

1.3. Predicting well-being with sensors

Wang et al. (2018) tried to predict whether or not a student was depressed on a week-by-week basis. They made use of smartphone and wearable sensors to collect the depression symptom features, which included sleep changes, diminished ability to concentrate (measured by phone usage) changes in activity and social engagement patterns, and heart rate. They used lasso regularized logistic regression (Tibshirani, 1996) to create the prediction model. They were able to predict whether or not a student was depressed with 81.5% recall (81.5% of the depressed cases were correctly identified), and 69.1% precision (69.1% of the inferred depressed cases were correct).

Being able to predict what movements a person makes can help provide healthcare and assistance when abnormal situations, such as falls, occur. Attal et al. (2015) showed a comparison between different classifier algorithms predicting particular movements in people's activity. Simple movements like sitting, standing and walking, or more complex intermediate movements like from sitting down to standing up and vice versa, or stair ascent and descent, were successfully classified with 99.2% accuracy. Attal et al. made a comparison between the supervised classifiers k-Nearest Neighbour, Support Vector Machine, Gaussian Mixture Models and Random Forest, where the k-Nearest Neighbour classifier showed the best results.

Lane et al. (2011) built a smartphone application that monitors the well-being of a person carrying the smartphone. In this study,

Lane et al. used existing healthcare professionals' guidelines to rate the well-being scores of the participants. The guidelines Lane et al. used include an ideal of seven hours of sleep and 150 to 300 minutes of moderate physical activity. The passive sensors on the smartphone were used to track sleep, physical activity and social interactions.

1.4. Research question

All the above studies demonstrate that it is possible to successfully predict the general or a particular type of well-being of a person using passive sensors. However, none of the studies seem to focus on the elderly population, in our eyes the main focus group, due to the ageing population and the challenges that come with it. In addition to that, none of the studies seem to include subjective measures, which are argued to measure well-being as well as objective measurements, and which should not be excluded in measuring well-being (Andrews, 1974; Campbell, 1976).

The research question will be: "Is it possible to predict the well-being of an elderly person by collecting sleep, physical activity, and time away from home data, collected through a passive sensor?". To answer this question, we collaborated with a company from the Netherlands, called MobileCare. MobileCare delivers an ecosystem to governments, consumers, care professionals and companies, in which they connect the social network of patients and clients in one application, relieving the stress from formal and informal caregivers. One product the company uses to collect data for the application is a wearable sensor called Vivago CARE watch, which allows caregivers to track a patient's health data on a distance. The Vivago CARE watch tracks sleep, physical activity, and time away from home data, which is one of the reasons why we use this data as our well-being predictors. The other reason is because the sleep, physical activity, and time away from home data are known to be predictors of well-being, as seen in section **1.2**. More information on the predictors will follow in section **2.3**.

The research question will be split up into three sub-questions. The first sub-question is: How well can subjective assessments of well-being be predicted? The second question is: How well can objective evaluations on well-being be predicted? The third question is: How well does well-being predicted by manually making an assessment on well-being by only looking at well-being predictors represent subjective and objective well-being?

The first two questions will give a direct answer to the research question. The third question will test the validity of manually looking at a person's data and making assumptions on the person's well-being using only that data.

2. Method

To determine whether we could predict well-being on the basis of objective data, we first needed to decide in what way we would determine well-being. Our way of determining a person's well-being in this study is threefold. First, a subjective well-being assessment is made by the person him or herself. Second, an objective well-being evaluation is made by qualified nurses. Last, a binary well-being label is produced by nurses based on sensor data. How we determined subjective and objective well-being is explained first. After that, we explain how we collected the predictors and the third well-being label.

2.1. Subjective well-being assessment

In section **1.1**, PROMIS is introduced as a way of measuring well-being in terms of physical, mental and social well-being. PROMIS consists of different banks of questions where each bank can measure a particular construct. To be sure we would measure only our settled construct (general well-being) and to have our own control over the evaluation function, we created a questionnaire with questions derived from the existing PROMIS system. We hand-picked questions from the item banks we found to be most representative for our construct. We altered and added some questions to be questioning the full physical, mental and social well-being range. In the end, the questionnaire existed of eleven questions that captured the well-being of a person. Five questions were linked to a person's physical well-being, four questions to their mental well-being and one question to their social well-being. The first ten questions had a range from zero to five, representing a "best to worse" scenario, thus scoring lower on the test meant a

better well-being. The last question was a pain-scale question ranging from zero (no pain) to ten (a lot of pain). After we were decided on the questions, we translated them from English to Dutch. The full questionnaire can be found in Appendix A.

2.2. Objective well-being evaluation

The objective well-being evaluation was provided by certified nurses, who gave a professional evaluation on a person's well-being. The nurses were contracted by the MobileCare company.

2.3. The predictors

To collect the predictors, we needed a passive sensor. The company was able to provide us with the Vivago CARE watch, which is a sensor watch designed for the medical industry. The watch measures four features: sleep, activity, whether the watch wearer is at home or not, and whether the person is wearing the watch or not. As seen in section **1.2**, sleep and activity are proven to correlate with well-being. Whether a person is at home or not can be seen as a form of social interaction, also proven to correlate with well-being, as seen in section **1.2**, since going outside the house or going to another place is often with the meaning of meeting other people. The activity feature ranges from zero to hundred and the other three features are measured in minutes.

The watch also calculates a circadian rhythm value for every day, ranging from zero to one, where closer to zero means a lower circadian alteration. As seen in section **1.2**, it is calculated through sleep- and activity patterns and functions as a measurement for well-being. The precise algorithm calculating the activity and the circadian rhythm is protected by the Vivago company.

2.4. Well-being label based on data

Besides the subjective self-reported assessments of well-being and the objective nurse evaluations on well-being, there is one more well-being label used in this thesis to discuss. Besides the sensor data obtained from the watch, the company was able to provide us with another binary well-being label created by nurses, based on the watch data. Nurses from the company looked at the sleep, activity and circadian rhythm and made an hourly estimate on the patient's well-being. This well-being label will from now on be referred to as the watch label, so as not to confuse it with the actual nurse evaluations.

2.5. Data collection

For ten consecutive days, we collected data from 22 MobileCare patients residing in the Netherlands. The average age of the participants is 74.95 (std=14.89), and the sample included 5 male and 17 female participants. The participants did not have any serious illnesses, but were being checked for general health daily by the nurses from the company, already before starting participating in our research. We will therefore refer to the participants as patients throughout the rest of the thesis. All patients wore the Vivago CARE watch during the ten days of collecting data. Besides the Vivago CARE watch data, we collected subjective well-being assessments and objective well-being evaluations of the patients.

Twice every day, we let the patients fill in the questionnaire discussed in section **2.1**, which we had uploaded to an online Google Form. The goal was to complete one questionnaire in the morning (10:00h) to evaluate the patient's past night's sleep and beginning of the day, and one questionnaire in the evening (20:00h) to evaluate how they felt the rest of the day. Most questionnaires were completed around those times, but due to other times harmonizing better with the patient's schedules, some questionnaires were completed at earlier (maximum one hour earlier) or later times (maximum one day later). When a patient could not fill in the questionnaire for one timeframe, we asked them to fill in the questionnaire at a later time, but linking the answers to how they felt at the time the questionnaire was supposed to be filled in. Some patients were able to fill in the questionnaire themselves, but considering their age, some of the patients were assisted by a nurse.

To collect the objective well-being labels, we asked the MobileCare nurses to evaluate the well-being of their patients twice every day, in the same timeframe the patients filled in the questionnaire. The patients were visited or video-called by their nurse, who made a binary judgment on the well-being of their patient (well or not well).

2.6. Data pre-processing of watch data

The Vivago CARE watch measures all features every minute. However, because we believe well-being does not change minute by minute, we converted the features to hourly data. The activity was averaged to hourly data. The sleep and whether a patient is at home or not were summed to hourly data to be in the interval of zero to sixty minutes, because they were binary values (per minute asleep or not asleep and per minute at home or not at home). The three features were also summed (sleep and at home or not) and averaged (activity) to a daily total value. We removed the rows where the patient was not wearing the watch.

In the end, a datafile of eight columns (the hour of the day, the circadian, the three features sleep, activity and at home or not and the three summed or averaged daily attributes) and 4736 rows (22 patients times 24 hours times 10 days, minus the hours the watch was off) remained.

2.7. Data pre-processing of questionnaire data

Because we wanted to use a binary classifier, we needed the well-being labels to be binary (well or not well). The nurses already evaluated in binary. The results of the questionnaire for each patient were calculated into a binary value. First, for each filled in questionnaire, the results from all questions were summed and normalized using feature scaling:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Where X is the total summed score, X_{min} is the minimum score one could score in the test, X_{max} is the maximum score one could score in the test and X' is the new normalized score in the range of $[0, 1]$. Second, the median of all normalized scores from each questionnaire from each patient was calculated. Third, a binary well-being value was calculated, where a patient was marked as being not well if his or her normalized score was higher than the median and a patient was marked as being well if his or her normalized score was lower or equal to the median.

2.8. Combining the data

Every patient had a unique ID-number, which was collected the same in the questionnaire, the nurse evaluations and the watch sensor data. Using this ID-number, the data from the Vivago CARE watch, the data from the questionnaire and the data from the nurses could be matched. The morning self-reported assessments and nurse evaluations were mapped hourly with the data from 22:00 in the evening until 10:00 in the morning. The evening self-reported assessments and nurse evaluations were mapped hourly from 10:00 in the morning until 22:00 in the evening.

2.9. Classifiers

For the predicting of well-being, we used different classifiers trained and tested on the three different well-being labels. The classifiers used in this paper to create the prediction models are the three supervised classifiers k-Nearest Neighbour (KNN), Random Forest (RF) and Support Vector Machine (SVM). What follows are brief summaries of how each classifier works:

- KNN is a statistical-based classifier. It calculates the distances between the observed sample and the K-closest samples in a training set, where K is a tuneable variable. A majority vote between the K-closest training samples to the observed sample decides whether the observed sample will be classified as class A or class B.
- RF is a tree-based classifier. The classifier creates a number of different decision trees. The decision trees all classify a given observation as class A or class B. A majority vote between all the decision trees decides whether the final predicted class of the observation will be classified as class A or class B.
- SVM is another statistical classifier. It tries to divide the training set by creating a hyperplane formula between the two classes. An observed sample will be classified as class A or class B based on the created formula.

We used the KNN, RF and SVM classifiers, because they are widely used in the field of machine learning to analyse and predict binary labelled data.

Each model was created using 10-fold cross-validation, which means that the dataset was split up in two complementary subsets ten times, where each time one part (90%) of the subset was

used as a training set, and the other part (10%) as a test set. The models were implemented in the programming language R, using the caret package.

The models were compared based on the classifier's accuracy (percentage of correctly classified observations), sensitivity (if a patient is not well, what is the percentage of the model predicting the patient as being not well) and specificity (if a patient is well, what is the percentage of the model predicting the patient as being well).

After the first models were created, we tried to improve them by using two different techniques: feature selection and re-sampling, which are explained next.

2.10. Feature selection

One technique we used to try to improve the models is feature selection. In feature selection, a subset of features (predictors) of the original feature set is used to evaluate which features are the most important in obtaining the highest accuracy score. In this paper, we used the RF classifier to select the most important predictors. The RF classifier has a built-in method of calculating the importance of each feature, called the Mean Decrease Impurity. In essence, it decides the importance of a feature by counting the times the feature is used to split a node in the decision trees. For more information, see Louppe et al. (2013).

2.11. Re-sampling

Figure 1 shows the distribution in percentages of the classes well and not well for all three well-being labels. As can be seen, the nurse evaluations and the watch labels are fairly imbalanced. Imbalanced data can lead to a classifier having a heavy bias for the majority class. To counter this imbalance in the classes, we used a technique called re-sampling. Re-sampling will balance the classes of the dataset in terms of well and not well. This technique has been used in previous research on a number of different topics, including data mining for direct marketing (Ling & Li, 1998) and the detection of oil spills in satellite radar images (Kubat et al., 1998).

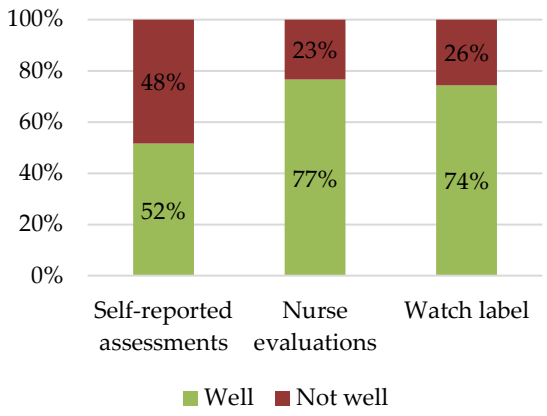

Figure 1: Well-being label distribution between the three different well-being labels.

In this paper, three different types of resampling techniques are applied to the nurse evaluation labels:

- **Oversampling.** In oversampling, the classes are balanced by randomly duplicating observations from the minority class, until the minority class has as much data samples as the majority class.
- **Undersampling.** Undersampling is the opposite of oversampling, but works in a similar way. In undersampling, the classes are balanced by randomly removing observations from the majority class, until the majority class has as much data samples as the minority class.
- **SMOTE.** SMOTE is a Synthetic Minority Oversampling Technique, meaning it will oversample the minority class, but does so by creating "new" observations on the basis of the statistics of the data. SMOTE makes the oversampled observations slightly different from the existing observations, by looking at two neighbouring data samples and calculating a new data sample somewhere randomly in between the neighbouring data samples. For more information on this oversampling technique, see Chawla et al. (2002).

2.12. Models

As seen in section **1.4**, the research question will be answered by answering three sub-questions.

To answer the first question of how well the self-reported assessments of well-being can be predicted, we created a model by training and testing the data on the self-reported assessments labels, after which we tried to improve the model using feature selection. The Random Forest feature selection suggested to remove three features: the hour of the day, the hourly activity and the hourly at home or not at home. An overview for the self-reported assessments models can be seen in **Figure 2**.

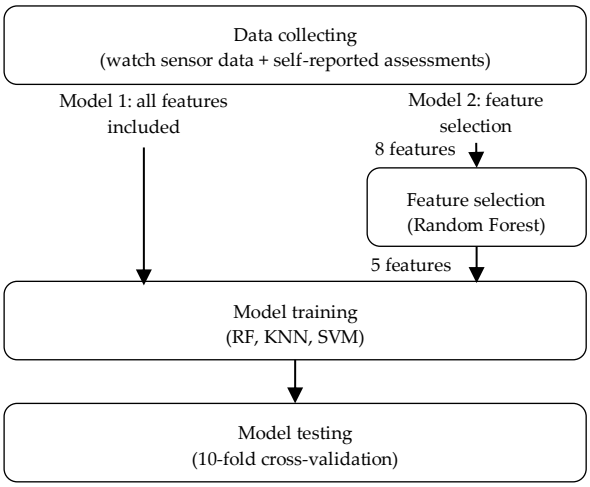

Figure 2: Overview of the models created using the self-reported assessments labels. Model 1 includes all features, model 2 uses feature selection.

Here is an overview of the classifier parameters used for the self-reported assessments models. The parameter values that achieved the highest accuracy for the classifiers were used.

- The parameters for RF were the number of trees (ntree) and the number of features to consider randomly as candidates at each split (mtry). Both models 1 and 2 performed best with 100 trees. All values possible for mtry were tested and the best mtry for model 1 and 2 were respectively 8 and 3.
- For the KNN classifier, the K parameter represents the number of neighbours a single data sample will look at to classify itself. Varying K from 1 to 30 found that the optimal value of K for model 1 and 2 was K = 3.
- The SVM classifier has the cost parameter C to tune, which after trying out different values (1 to 5) turned out to be best left on the default C = 1 for both models.

To answer the second question of how well the nurse evaluations on well-being can be predicted, we created a model by training and testing the data on the nurse evaluation labels, after which we tried to improve the model by using the different re-sample techniques. An overview for the nurse evaluation models can be seen in **Figure 3**.

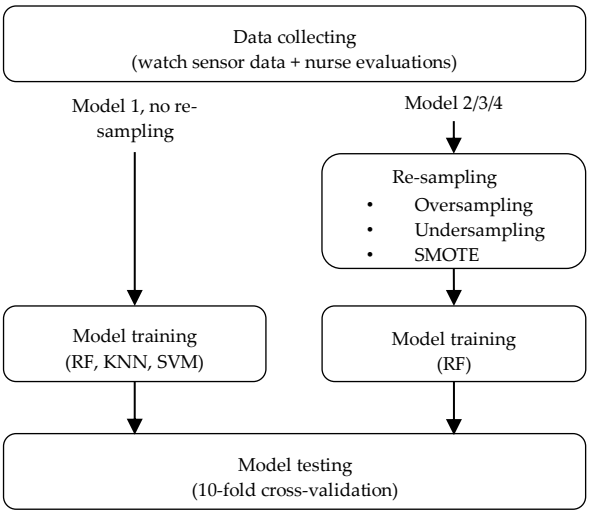

Figure 3: Overview of the models created using the nurse evaluation labels. Model 1 uses all features and no re-sampling, model 2 uses all features and different re-sampling techniques.

An overview of the classifier parameters used for the nurse evaluation models:

- For RF, the number of trees was again 100 for all models. For model 1, the classifier scored the best accuracy with mtry = 3. For the other models, the accuracy was highest with mtry = 11.
- In KNN, the number of neighbours K had an optimal value of K = 2.
- For the SVM classifier, the cost was set to C = 1.

To answer the last question, to what extent do the watch labels represent the nurse evaluations and self-reported assessments, we tested how well the watch labels could be predicted by the best self-reported assessments model and by the best nurse evaluations model. An overview can be seen in **Figure 4**. We also looked at the similarities between the different well-being labels by how well they matched at each measuring point. **Figure 5** shows that the labels are rather different from each other. We will see in the results how

this reflects back to the answer to sub-question three.

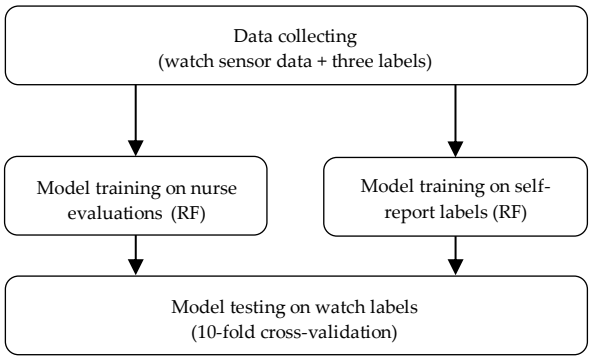

Figure 4: Overview of the models created by training on nurse and self-report labels and testing on watch labels.

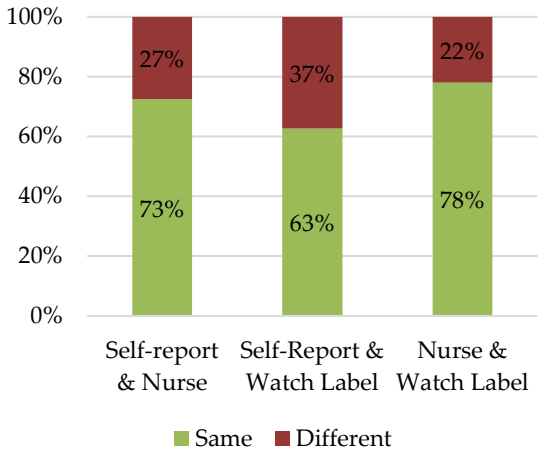

Figure 5: Similarity between the different labels

3. Results

The results will be divided up into three parts. First, we will take a look at the classifier performances on the self-reported assessments of well-being before and after feature selection. Second, we will take a look at the classifier performances on the nurse evaluations on well-being before and after applying the different re-sampling techniques. Third, we will see to what extent the watch labels can be predicted using the best models found in the first two sections.

3.1. Self-reported assessments results

The results obtained for the patient's self-reported assessments of well-being before and after feature selection can be found respectively in **Figure 6** and **Figure 7**. The standard deviation of the 10-fold cross-validation, explained in section **2.9,** is displayed for the accuracy in both graphs. **Figure 8** and **Figure 9** show the ROC curves for respectively the models before feature selection and the models after feature selection. The ROC curve plots the sensitivity and specificity at different threshold values. The RF classifier shows a line close to the upper left corner, meaning it has a good trade-off between a good sensitivity and a good specificity.

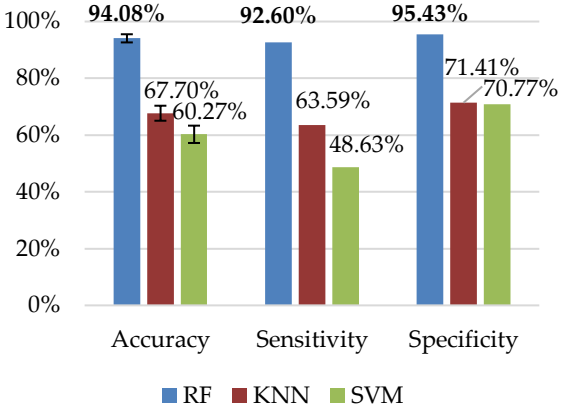

Figure 6: Performance results for the RF, KNN and SVM classifiers on the patient's self-reported assessments of well-being.

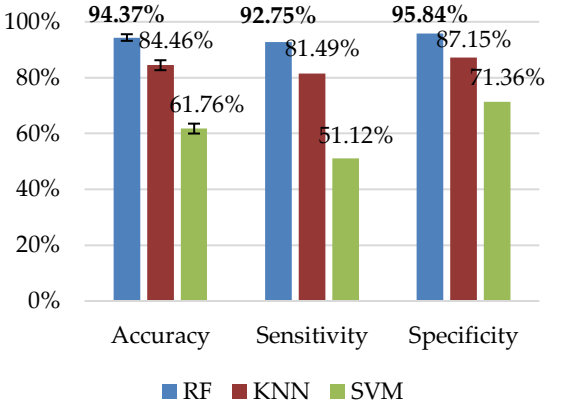

Figure 7: Performance results for the RF, KNN and SVM classifiers on the patient's self-reported assessments of well-being after feature selection.

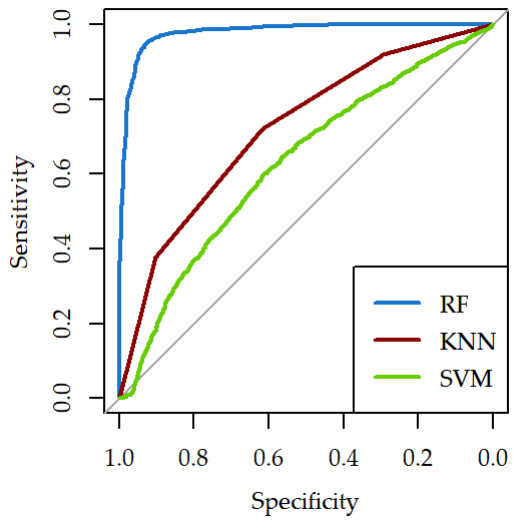

Figure 8: ROC curve of the three models before feature selection. The AUCs for RF, KNN and SVM are respectively 0.9784, 0.7219 and 0.6281.

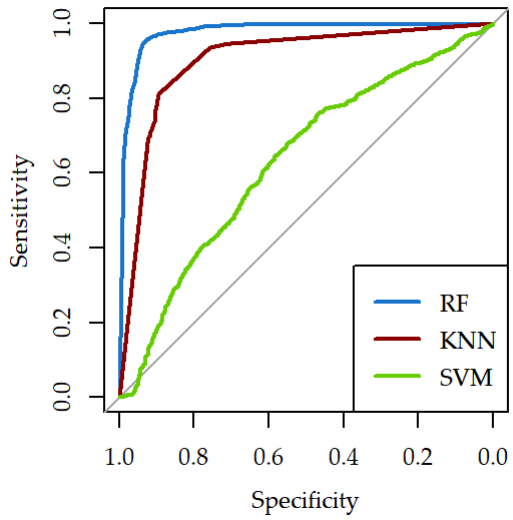

Figure 9: ROC curves of the three models after feature selection. The AUCs for RF, KNN and SVM are respectively 0.9765, 0.9025 and 0.6329.

Table 1 shows the confusion matrix from the RF classifier after feature selection, to show an exact number of correctly and incorrectly classified instances.

Table 1: Confusion matrix showing correctly and incorrectly classified instances predicted by the best RF model for the patient's self-reported assessments of well-being.

		Reference	
		Not well	**Well**
Predicted	**Not well**	1829	91
	Well	143	2095

Figure 10 shows the course of the feature selection process. With five, six, seven or eight features, the accuracy does not change much. Using less than five features drops the accuracy considerably. Note that the y-axis does not start at zero percent.

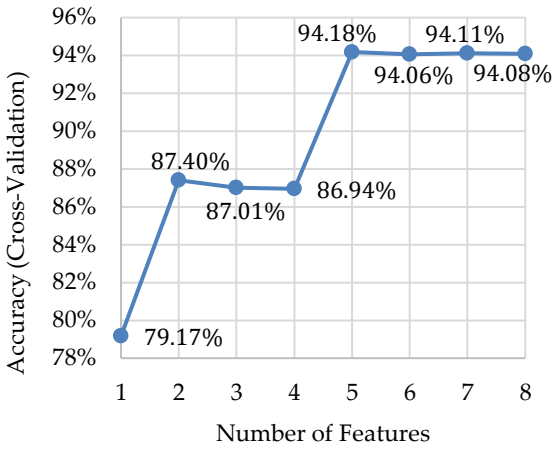

Figure 10: The course of the feature selection process on the patient's self-reported assessments of well-being.

3.2. Nurse evaluations results

The results from predicting the patient's self-reported assessments of well-being showed a reasonably high accuracy. We will now see if the results obtained for predicting the nurse evaluations on well-being can achieve a similar accuracy. The results before and after re-sampling can be found respectively in **Figure 11** and **Figure 12**. The standard deviation of the 10-fold cross-validation is again displayed for the accuracy. **Figure 13** shows the ROC curves of the models before re-sampling. An ROC curve for the models after re-sampling is not shown, because the sensitivity and specificity scores were too much alike for all three re-sampling methods, making the ROC lines overlap too much. **Table 2** shows the RF confusion matrix without resampling.

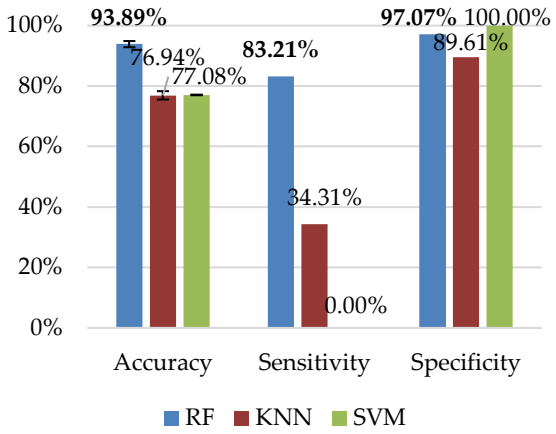

Figure 11: Performance results for the RF, KNN and SVM classifiers on the nurse evaluations.

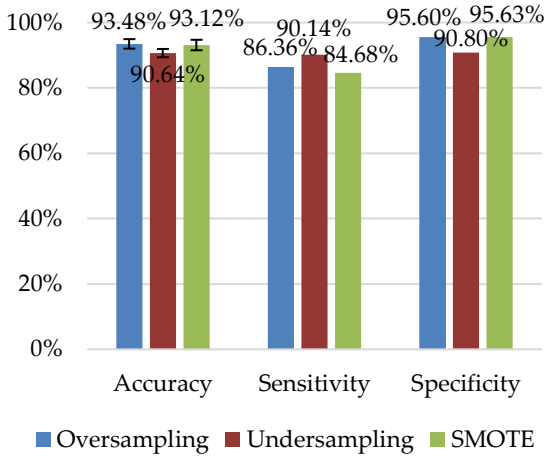

Figure 12: Performance results for the RF, KNN and SVM classifiers on the nurse evaluations after resampling.

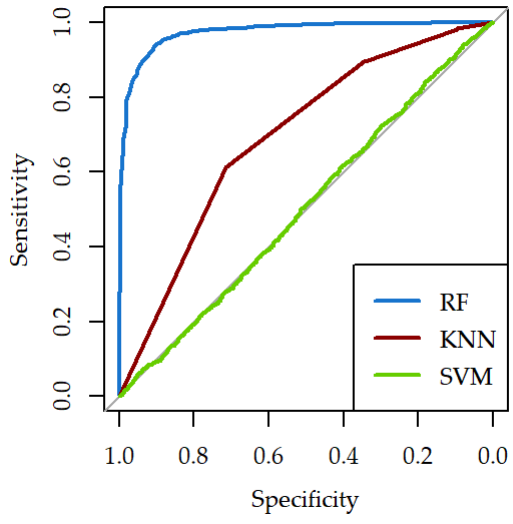

Figure 13: ROC curves of the three models without re-sampling. The AUCs for RF, KNN and SVM are respectively 0.9733, 0.6956 and 0.5045.

Table 2: Confusion matrix showing correctly and incorrectly classified instances precited by the best RF model for the nurse evaluations on well-being.

		Reference	
		Not well	**Well**
Predicted	**Not well**	793	94
	Well	160	3111

3.3. Watch label testing results

Below in **Figure 14** are the results for testing to what extent the watch labels can be predicted using the two best models from section **3.1** and section **3.2**.

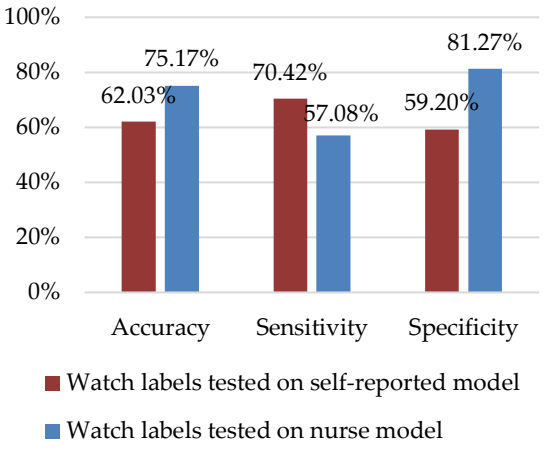

Figure 14: Watch labels tested on the best self-reported model and the best nurse evaluations model.

4. Discussion

To discuss the results we found, we need a way to compare the different classifier performances. This is not only done by comparing the accuracy score of the models, but more importantly the sensitivity: the ability of the models to correctly classify a patient as not well. In this research, sensitivity is a more important measurement than the specificity, because classifying a "not well" patient as "well", will hurt the patient more than classifying a "well" patient as "not well". For example, when a not well patient is classified as being well, nurses are less likely to check up on the patient, which can lead to more serious well-being problems for the patient. When a well patient is classified as not being well, nurses will likely unnecessarily check up on the patient, only costing the nurse's work and time, but the patient's well-being is not in danger. One model may show a higher accuracy score than another, but if the second model has a higher sensitivity

score, the second model may be better for this research purpose.

4.1. Predicting self-reported well-being

We set out to determine whether it is possible to predict well-being on the basis of wearable sensor data. The RF classifier showed promising results predicting the patient's self-reported assessments of well-being with 94.37% accuracy, 92.75% sensitivity and 95.84% specificity after feature selection. Looking at the ROC graph from **Figure 9**, KNN shows good results after feature selection as well with an AUC of 0.9025. All classifiers improved in accuracy, sensitivity and specificity after feature selection. The most important features to predict the self-reported well-being for this dataset are the circadian, the hourly sleep time and the three summed or averaged daily features (sleep, activity, at home or not). In conclusion, the best model is able to predict the subjective well-being with an accuracy of 94.37% and a sensitivity of 92.75%.

4.2. Predicting nurse evaluations

The RF classifier showed again the best results for predicting the nurse evaluations with 93.89% accuracy, 83.21% sensitivity and 97.07% specificity, without re-sampling. Because nurse evaluation classes are imbalanced, the models show a heavy bias for the "well" labels. The SVM classifier even only predicts a patient as being well, because it cannot learn enough from the "not well" label. This can also be seen in the ROC graph of **Figure 13**, where the SVM ROC "curve" is almost a straight line through the 50/50 separation line. After re-sampling, the RF classifier showed a lower accuracy score for all re-sampling techniques, but also an increase in sensitivity (**Figure 12**). Undersampling shows the highest increase in sensitivity, with almost 7%. Thus, for the nurse evaluations, the RF classifier model after undersampling is chosen as best model. The conclusion here as well, is that the best model is able to predict the objective well-being with an accuracy of 90.64%.

4.3. Validity of the watch label

The watch label, as explained in section **2.4**, was created by nurses from the company, who looked at the sleep, activity and circadian rhythm, and made a binary estimate on the patient's well-being. The results of **Figure 14** show that the percentage of correctly predicted watch labels is 75.17% when tested on the nurse evaluations model. Although this seems all right, the sensitivity score is fairly low with 57.08%. When trained on the patient's self-reported assessments of well-being model, the sensitivity is a little higher with 70.42%, but the overall accuracy is low with 62.03%. With the patient's self-reported assessments of well-being as a ground truth, this means that the watch labels cannot be used as a predictor for the patient's self-reported assessments of well-being. The same is true for when the nurse evaluations are considered as a ground truth for well-being, which means that the watch labels cannot be used as a predictor for the nurse evaluations on well-being.

4.4. The best classifier

The tests suggest RF as the best classifier for this dataset. This might come as a surprise, because as seen in section **1.3**, the statistical classifiers KNN and Logistic Regression showed the best results in previous research. However, one of the reasons why the RF classifier performs better than the KNN and the SVM classifiers may be that RF takes into account the feature importance, whereas in KNN and SVM, all features are equally important. RF performs especially well in this dataset, because as seen in section **3.1** and **4.1**, some of the features have a more predictive value than others. In the next section, we will explain why this might also lead to the problem of overfitting.

4.5. Potential problems

Although the results seem to be heavily in favour of a positive answer to the research question, one always has to keep a critical eye on the results.

One potential problem is the size of the dataset. The dataset is not considerably large, meaning the classifiers might perform differently in a real-world scenario. In other words, the examples in this small dataset might not generalize perfectly to the entirety of all elderly people, or even all elderly people living in their own homes in the Netherlands.

Another problem that might be occurring is overfitting. As explained in section **2.6**, we used the daily summed and averaged totals (sleep, activity, at home or not) as features. This has as a

result that many rows in the dataset have the same value, making it hard for the classifiers to distinguish different data samples, thus making it obvious to classify them as a certain class.

Another cause for the possible overfitting is due to the self-reported assessment and nurse evaluation labels being distributed over hourly data. The self-reported assessments and the nurse evaluations happened twice every day, once in the morning and once in the evening. To take advantage of the hourly features we had available from the watch sensor, we distributed the twice a day labels over 24 hours. As seen in section **2.8,** the morning assessment and evaluation labels were mapped with the hours from 22:00 in the evening until 10:00 in the morning. The evening labels were mapped from 10:00 in the morning until 22:00 in the evening. In combination with the daily summed and averaged totals however, the classifier sees a lot of the same data samples having the same well-being label. This might also explain why the feature selection, described in section **2.10**, decided that the daily summed and averaged totals influenced the classifier's models the most. However, when only the three summed or averaged daily features (sleep, activity, and at home or not) are left as features, the classifier drops considerably from roughly 94% accuracy to 86% accuracy, as can be seen in section **3.1**, which demonstrates that the classifier does need the hourly changing features.

4.6. Using other predictor features

As seen in **1.2**, there are other features that may be able to predict well-being. The features used in this research (sleep, activity, time away from home) were mostly chosen because of the availability and ease of collecting through the Vivago CARE watch. It is possible other features are better suited for predicting well-being, or extra features engineered from the features we used. For example, in this thesis, we did not look into daily sleep or activity differences or moving averages, which could potentially predict a change in well-being.

5. Conclusion

First, we showed that subjective well-being can be predicted with an accuracy of 94.37% (92.75% sensitivity, 95.84% specificity). Second, we found that the objective well-being as determined by a nurse can be predicted with an accuracy of 90.64% (90.14% sensitivity, 90.80% specificity). Furthermore, we found that manually assessing well-being of a person solely based on looking at sleep, activity, at home or not and circadian data is not a good predictor of subjective nor objective well-being.

Those conclusions however, must still be viewed with a critical eye, since a chance of overfitting in the classifiers is not excluded. We think solving this potential problem needs to be the aim of future research that focusses on this research question.

One possible solution would be to assess the well-being of a person hourly, just like the predictors are collected hourly. This however is practically unfeasible, since patients would need to fill in questionnaires hourly and nurses would need to evaluate their patients hourly, during the day but also at night. This cannot be done without intruding too much in people's everyday life. Another more feasible solution would be to assess the well-being of a person daily instead of twice a day, and make daily predictions on the well-being of a person instead of hourly. This would still help nurses to plan their time more efficiently, for example where daily routine check-ups of people who are well can be skipped. However, for a classifier to be able to predict accurately, it needs a lot of data. As in our research, ten days would not be enough to accurately predict daily well-being. How much data is needed, depends on the classifier and the distribution between the "well" and the "unwell" class. Collecting daily well-being assessments for eight months would result in the same number of data points we were able to use for this research. Finding enough participants and nurses willing to make all those daily assessments might be the biggest problem future researchers have to face, when solving the automation of predicting well-being.

6. References

Andrews, F. M. (1974). Social indicators of perceived life quality. Social Indicators Research, 1(3), 279-299.

Anokye, N., Trueman, P., Green, C., Pavey, T., & Taylor, R. (2012). Physical activity and health related quality of life. Bmc Public Health, 12, 624-624.

Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., & Amirat, Y. (2015). Physical human activity recognition using wearable sensors. Sensors (basel, Switzerland), 15(12), 31314-38.

Buysse, D. J. (2014). Sleep health: can we define it? Does it matter?. Sleep, 37(1), 9-17.

Campbell, A. (1976). Subjective measures of well-being. American psychologist, 31(2), 117.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., ... & Rose, M. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Medical care, 45(5 Suppl 1), S3.

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357. doi:10.1613/jair.953

Cooney, M., Vartiainen, E., Laatikainen, T., Juolevi, A., Dudina, A., & Graham, I. (2010). Elevated resting heart rate is an independent risk factor for cardiovascular disease in healthy men and women. American Heart Journal, 159(4), 612-619.

George, L., Blazer, D., Hughes, D., & Fowler, N. (1989). Social support and the outcome of major depression. The British Journal of Psychiatry, 154(4), 478-485.

Hays, R., Wells, K., Sherbourne, C., Rogers, W., & Spritzer, K. (1995). Functioning and well-being outcomes of patients with depression compared with chronic general medical illnesses. Archives of General Psychiatry, 52(1), 11-9.

Keller, H. (2004). Nutrition and health-related quality of life in frail older adults. The Journal of Nutrition, Health & Aging, 8(4), 245-52.

Kubat, M., Holte, R., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. Machine Learning, 30(2-3), 195-215.

Lane, N. D., Mohammod, M., Lin, M., Yang, X., Lu, H., Ali, S., ... & Campbell, A. (2011). Bewell: A smartphone application to monitor, model and promote wellbeing. In 5th international ICST conference on pervasive computing technologies for healthcare (pp. 23-26).

Lane, N., Lin, M., Mohammod, M., Yang, X., Lu, H., Cardone, G., . . . Choudhury, T. (2014). Bewell: Sensing sleep, physical activities and social interactions to promote wellbeing. Mobile Networks and Applications : The Journal of Special Issues on Mobility of Systems, Users, Data and Computing, 19(3), 345-359.

Louppe, G., Wehenkel, L., Sutera, A., Geurts, P., & 27th Annual Conference on Neural Information Processing Systems NIPS 2013 27th Annual Conference on Neural Information Processing Systems, NIPS 2013 Lake Tahoe, NV, USA 2013 12 05 - 2013 12 10. (2013). Understanding variable importances in forests of randomized trees. Advances in Neural Information Processing Systems, (2013 01 01).

Ling, C. X., & Li, C. (1998, August). Data mining for direct marketing: Problems and solutions. In Kdd (Vol. 98, pp. 73-79).

Liu, J. X., Goryakin, Y., Maeda, A., Bruckner, T., & Scheffler, R. (2017). Global health workforce labor market projections for 2030. Human resources for health, 15(1), 11.

Paavilainen, P., Korhonen, I., & Partinen, M. (2005). Telemetric activity monitoring as an indicator of long-term changes in health and well-being of older people. Gerontechnology, 4(2).

Shrivastava, S. R. B. L., Shrivastava, P. S., & Ramasamy, J. (2013). Health-care of Elderly: Determinants, Needs and Services. International journal of preventive medicine, 1(1), 1224-5.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (methodological), 58(1), 267-288.

Van Someren, E. (2000). Circadian and sleep disturbances in the elderly. Experimental Gerontology, 35(9), 1229-1237.

Wang, R., Wang, W., Dasilva, A., Huckins, J., Kelley, W., Heatherton, T., & Campbell, A. (2018). Tracking depression dynamics in college students using mobile phone and wearable sensing. Proceedings of the Acm on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2(1), 43.

Wells, K., Stewart, A., Hays, R., Burnam, M., Rogers, W., Daniels, M., . . . Ware, J. (1989).

The functioning and well-being of depressed patients. results from the medical outcomes study. Jama, 262(7), 914-9.
World Health Organization. (2002a). Active ageing: A policy framework. Geneva, Switzerland: World Health Organization.
World Health Organization. (2002b) Towards policy for health and ageing. Geneva, Switzerland: World Health Organization.
World Health Organization. (2006) Constitution of the World Health Organization. Geneva, Switzerland: World Health Organization.
World Health Organization. (2015). World report on ageing and health. Geneva, Switzerland: World Health Organization.

7. Appendix A

Appendix A shows the questionnaire questions that the patients were asked to fill in to assess their general well-being. First in English, then translated to Dutch.

7.1. English questionnaire

Question 1:
- Q: How did you feel in general in the past hours?
- A: Excellent, Very good, good, fair, poor

Question 2:
- Q: How was your physical health in the past hours?
- A: Excellent, Very good, good, fair, poor

Question 3:
- Q: How was your mood in the past hours?
- A: Excellent, Very good, good, fair, poor

Question 4:
- Q: How was your ability to think clearly and concentrate in the past hours?
- A: Excellent, Very good, good, fair, poor

Question 5:
- Q: How satisfying were your social activities and relationships in the past hours?
- A: Excellent, Very good, good, fair, poor

Question 6:
- Q: How well were you able to carry out your usual social activities and roles (this includes activities at home, at work and in your community, and responsibilities as a parent, child, spouse, employee, friend, etc.) in the past hours?
- A: Excellent, Very good, good, fair, poor

Question 7:
- Q: How well were you able to carry out your everyday physical activities such as walking, climbing stairs, carrying groceries, or moving a chair in the past hours?
- A: Completely, Mostly, Moderately, A little, Not at all

Question 8:
- Q: How often did you feel bothered by emotional problems such as feeling anxious, depressed or irritable in the past hours?
- A: Never, Rarely, Sometimes, Often, Always

Question 9:
- Q: How much were you able to let go of unwanted thoughts in the past hours?
- A: Whenever I wanted, Most of the time, Sometimes, Rarely, Almost never

Question 10:
- Q: How would you rate your fatigue right now?
- A: None, Mild, Moderate, Severe, Very severe

Question 11:
- Q: How would you rate your pain in the past hours?
- A: Scale from 0 (no pain) to 10 (worst pain imaginable)

7.2. Translated Dutch questionnaire

Question 1:
- Q: Hoe voelde u zich in het algemeen in de afgelopen uren?
- A: Uitstekend, Zeer goed, Goed, Redelijk, Slecht

Question 2:
- Q: Hoe was uw fysieke gezondheid in de afgelopen uren?
- A: Uitstekend, Zeer goed, Goed, Redelijk, Slecht

Question 3:
- Q: Hoe was uw humeur in de afgelopen uren?
- A: Uitstekend, Zeer goed, Goed, Redelijk, Slecht

Question 4:

- Q: In hoeverre kon u helder nadenken en u concentreren in de afgelopen uren?
- A: Uitstekend, Zeer goed, Goed, Redelijk, Slecht

Question 5:

- Q: Hoe bevredigend waren uw sociale activiteiten en relaties in de afgelopen uren?
- A: Uitstekend, Zeer goed, Goed, Redelijk, Slecht

Question 6:

- Q: Hoe goed was u in staat om uw gebruikelijke sociale activiteiten en rollen (dit zijn activiteiten thuis, op het werk en in uw gemeenschap, en verantwoordelijkheden als ouder, kind, echtgenoot, werknemer, vriend, enz.) in de afgelopen uren uit te voeren?
- A: Uitstekend, Zeer goed, Goed, Redelijk, Slecht

Question 7:

- Q: Hoe goed was u in staat om uw dagelijkse fysieke activiteiten, zoals lopen, traplopen, boodschappen doen, of een stoel verplaatsen, in de afgelopen uren uit te voeren?
- A: Uitstekend, Zeer goed, Goed, Redelijk, Slecht

Question 8:

- Q: Hoe vaak voelde u zich de laatste uren lastig gevallen door emotionele problemen zoals angstigheid, depressiviteit of prikkelbaarheid?
- A: Nooit, Zelden, Soms, Vaak, Constant

Question 9:

- Q: In hoeverre kon u in de laatste uren ongewenste gedachten loslaten?
- A: Altijd, Meestal, Soms, Zelden, Bijna nooit

Question 10:

- Q: Hoe zou u uw vermoeidheid beoordelen op het moment?
- A: Geen, Mild, Matig, Ernstig, Zeer ernstig

Question 11:

- Q: Hoe zou u uw pijn in de afgelopen uren beoordelen?
- A: Op een schaal van 0 (geen pijn) tot 10 (ergst denkbare pijn)