university of groningen

faculty of mathematics and natural sciences

artificial intelligence

PREDICTING WELL-BEING WITH WEARABLE SENSOR DATA

Bachelor's Project Thesis

Ivana D.M. Akrum (i.d.m.tjong.a.hung@student.rug.nl)

Supervisors: Dr M.K. van Vugt

Abstract: The world's population is ageing rapidly but not more healthily. The healthcare industry needs some way to accommodate for the increasing healthcare needs of the elderly. This research proposes the idea of using classifiers in combination with wearable sensor data to monitor the well-being of the elderly. This will allow healthcare professionals to focus on those whom the algorithm has already identified as being unwell, thereby sparing them the time it would have taken to manually determine which patients need medical attention and which currently do not. After looking at literature on which factors influence health and cross-referencing that with which sensor data are feasible to measure within the scope of this research, the research aims to answer the question: Can classifiers be used to make predictions of well-being on the basis of sleep, physical activity, time away from home, and the circadian rhythm? To answer this question, an observational study of ten days was done for twenty participants in which a wearable measured the necessary health data. The participant well-being was assessed in three ways: through diagnoses from health professionals, through participant self-report, and through the sensor data itself. A K-Nearest Neighbour (KNN) and Random Forest (RF) classifier were trained that used the sensor data to make predictions of well-being as assessed through the data. The RF classifier gave the best results and showed an accuracy of 89%. However, the well-being assessed through sensor data was ultimately found to be unrepresentative of well-being, and the RF classifier performance was considerably worse when compared against the well-being assessed through health professionals and through self-report. As such, even the optimal classifier designed in this research cannot be used to make reliable predictions of well-being on the basis of sleep, physical activity, time away from home, and the circadian rhythm.

1. Introduction

The world's population is ageing rapidly, which means larger and larger percentages of the world's population are made up of elderly. In general, life expectancy has increased rapidly since the 20th century, and it is likely to continue to do so in the foreseeable future (United Nations Department of Social and Economic Affairs, 2007).

Although people are generally living longer, they are not healthier in their old age than the generations before them (Crimmins & Beltrán-Sánchez, 2011). It is thus necessary for the health industry to adapt to this change in the population, so that every elderly person can be provided with good and targeted healthcare.

In this paper the idea of predicting well-being with wearable sensor data is introduced as one method that can aid the healthcare industry with adapting to the ageing population. By automating the process of knowing when someone is well or not, time and costs are freed up for health professionals to focus on curing the people who are unwell. Furthermore, an increase of patient-generated health data through wearables (such as Apple Watch) in recent years (HIMSS Europe GmbH, 2018) has made it possible to use such data together with methods from machine learning to make these predictions of well-being.

Before focusing more specifically on how machine learning can be used, it is necessary to define well-being. This paper looks at general well-being, which is defined as a health state with two values: someone can either be well or unwell. The two states are considered in terms of their impact on the healthcare industry. It is assumed that when someone is well, they do not need medical attention, whereas when someone is unwell, they are no longer capable of adapting and self-managing their current health status by themselves and thus they require medical attention (Huber, et al., 2011).

Currently, well-being can be determined by health professionals or through patient self-report. While health professionals follow various guidelines that ensure their verdicts are valid and reproducible, self-report is still a very subjective way of measuring well-being.

To help standardize patient self-reports, the PROMIS Health Organisation created the Patient Reported Outcomes Measurement Information System (PROMIS), which contains various instruments for testing patient well-being across multiple domains (such as fatigue, anxiety, or social isolation). Particularly interesting for this study is the PROMIS Global Health instrument, as this survey most closely mirrors the idea of general well-being described in the previous paragraphs.

The Global Health survey can be divided into two categories: Global physical health and global mental health. The validity of the Global Health survey was tested along these two categories in a research which analysed 10 self-reported global health items. It was found that the outcomes of the survey showed consistency over these 10 items for the scores of both the physical and mental health category. Furthermore, the scores also correlated with a different instrument universally used for self-report known as the EuroQol-5D (Balestroni & Bertolotti, 2015). Evidence thus suggests the Global Health survey can be used efficiently to summarize general well-being (Hays, Bjorner, Revicki, Spritzer, & Cella, 2009).

In this research, both health professionals and self-report (on the basis of PROMIS) will be used to determine binary well-being labels. In machine learning, such labels can be used to teach an algorithm (computer program) how certain data correlate to the labels. Once the algorithm knows this, it can then make predictions from the data that, ideally, match the subjective well-being labels of the health professionals and the self-report with a hundred percent accuracy. This process is known as supervised learning, and the algorithms used for this are called classifiers.

The well-being labels the classifiers need are gathered from health professionals and self-report, but the classifiers also need data that are indicative of well-being. To determine which sensor data to measure in this research, it is thus first necessary to determine which health data are indicative of well-being.

First and foremost, there are the measurements that constitute the vital signs, which indicate the state of a patient's essential body functions. In Goldman's Cecil Medicine (one of the most influential internal medicine references), temperature, pulse, blood pressure, respiratory rate, and oxygen saturation are identified as five key vital signs. However, the book also states that these vital signs have little predictive value on health on their own: Abnormal vital signs do not necessarily translate to an unhealthy individual, and vice versa, normal vital signs do not necessarily mean someone is healthy (Cecil, Goldman, & Schafer, 2012). Although vital signs do play a factor in determining a patient's health status, they would thus be of little help to classifiers that have no information on a patient's larger medical background.

There are factors, however, that show a more direct influence on well-being. Two of these factors, as identified by The World Health Organisation, are physical activity and nutrition, which play an important role in healthy ageing (World Health Organization, 2015). High to moderate levels of physical activity are shown to reduce the risk of both the development of functional limitations (limitations that inhibit the ability to perform daily tasks) (Paterson & Warburton, 2010) and cognitive decline (Blondell, Hammersley-Mather, & Veerman, 2014) in old age. Malnutrition leads to reduced muscle and bone mass, thereby increasing frailty.

Other important health factors across all age groups are sleep and the circadian rhythm (specifically the sleep/wake cycle). Sleep and the circadian rhythm are closely related to each other, and evidence suggests disruptions in the circadian rhythm as a result of sleep deprivation may increase the severity of age-related chronic disorders (Spiegel, Leproult, & Van Cauter, 1999), as well as exacerbate depressive states (Salgado-Delgado, Tapia Osorio, Saderi, & Escobar, 2011). Irregular sleep schedules and short sleep cycles also both show a direct relationship to depression (Wang, et al., 2018). Finally, disruptions to the circadian rhythm as a result of exposure to light at night lead to increased odd ratios for obesity and heart attacks (Brainard, Gobel, Scott, Koeppen, & Eckle, 2015), and even promote tumorigenesis in rats (Vinogradova, et al., 2010).

A final health factor identified is location and mobility. Similarly to with sleep, evidence found that an increase in being stationary and a decrease in visiting places outside the home leads to higher scores for depression on the PHQ-8 and PHQ-4 tests, which are measurements for depression (Wang, et al., 2018).

In conclusion, physical activity, nutrition, sleep, and the circadian rhythm are identified as factors that influence physical well-being, while physical activity, sleep, the circadian rhythm, and location and mobility show an influence on mental well-being. Knowing which health factors influence well-being, the only matter left to discern is which of these factors can be measured as sensor data from wearables. This depends directly on the wearable used. When choosing a wearable to use, it is important to look at both which health data it can measure and the wearable's ease of use for elderly.

Considering these factors, this research will use the Vivago smart watch to gather sensor data. The Vivago watch is meant specifically to remotely monitor the well-being of elderly (Särelä, Korhonen, Lötjönen, Sola, & Myllymäki, 2003). Evidence shows it can accurately monitor sleep/wake patterns (Lötjönen, et al., 2003) as well as physical activity levels (Vanhelst, et al., 2012). It can thus be used to gather information on physical activity, sleep, and the circadian rhythm. It also has a measure for determining when someone is away from home, which corresponds to the location and mobility health factor identified.

By analysing known health factors in literature and describing two methods to determine subjective well-being labels (health professionals and self-report), all the data necessary for the creation of classifiers have been identified. However, the aim of this research is to minimise human involvement in determining well-being as much as possible, and yet the two methods that will be used to assess well-being still require either a health professional to examine patients or patients themselves to fill in a survey on their well-being.

On a long-term basis, neither of these methods are ideal ways of determining well-being, and thus a faster, more automated way of determining well-being is proposed: determining well-being through the sensor data directly. Since literature shows the sensor data have effects on well-being, it is feasible to think that certain values will correspond to being well and certain values will correspond to being unwell. Rather than using the well-being determined by health professionals and self-report, this research will thus use well-being labels determined through the sensor data to create the classifiers.

Bringing it all together, the research will try to answer the research question: **Can classifiers be used to make predictions of well-being on the basis of sleep, physical activity, time away from home, and the circadian rhythm?**

To answer this question, an observational study will be done to gather the necessary health data of elderly (aged sixty or older) that are generally doing well, that is to say, elderly who are generally capable of adapting and self-managing their health status independently, without the need for medical attention. After the observational study, classifiers will be created that use the sensor data to make predictions on well-being. The well-being labels that teach the algorithm how the data correlates to well-being will also be determined through the sensor data. Consequently, in the final part of this research, the validity of using a well-being determined through sensor data will be checked by comparing it against the well-being labels determined through health professionals and self-report. The methods are discussed more in-depth in the following section.

2. Methods

2.1. Observational study design

The observational study in this research is performed in collaboration with the Dutch company MobileCare. MobileCare provides additional support and care for people that subscribe to their services by using smart devices to monitor their well-being. The Vivago smart watch is one of the devices MobileCare uses for this monitoring of well-being.

The clients of MobileCare use their services in addition to regular healthcare, and it is assumed they are not in need of any specific medical interference, but rather have additional risk factors in their lives (such as old age) for which they desire more/closer care than our current health industry provides them with.

Of the MobileCare clients, twenty were recruited as participants for the observational study. The participants had a median age of 78 years. The youngest participant was 35 years old and the oldest was 95 years old. The participant sample consisted of 5 males and 15 females. All participants owned a Vivago smart watch and had an identification number as provided by MobileCare.

The experiment lasted 10 days during which hourly data from the Vivago smart watch was collected for each participant. The data the Vivago smart watch measured were:

- **Sleep**: The amount of sleep the participant had in minutes in the current hour.
- **Moderate activity**: A percentage of the maximum possible activity. Normal values for this variable for elderly are between 15 and 40 percent during the day and (close to) zero at night.
- **Time outside**: The amount of time spent outside of home in minutes.
- **Watch off**: The amount of time the watch is not worn in minutes.

For *sleep*, *time outside*, and *watch off*, the measurements in minutes were also provided in hours (given that the data are hourly, these values were always between 0 and 1 hours). For all variables, a daily total was given in addition to the hourly measurements.

Next to these variables, the Vivago watch also provided **the circadian**, a daily variable. The *circadian* is determined internally by the Vivago watch through the other data. To interpret the *circadian*: values below 0.4 are considered good, values between 0.4 and 0.5 are satisfactory, and any value above 0.5 is bad. A good circadian rhythm means a participant exhibited a stable sleep/wake cycle (long enough night time sleep and high activity during the day).

These sensor data were given hourly binary well-being labels (where 1 represents being well and 0 represents being unwell) retroactively. The well-being was determined by the data values, with the circadian being the largest deciding factor. This is one of three sets of well-being labels used in this research in total, and because it comes from the sensor data measured by the Vivago watch, it will be referred to as the *Vivago label* throughout this paper.

The Vivago label corresponds to the well-being determined by the sensor data. Next to this, as was stated in the *Introduction*, two more sets of well-being labels are gathered: one that represents well-being determined by health professionals, and one that represents well-being determined through self-report.

For the well-being determined by health professionals, personal coaches from MobileCare examined the participants twice daily via video call for the duration of the study. In these calls, they examined the participants to determine their well-being. They formed a professional opinion on the basis of the larger picture, rather than looking solely at data or through a strict set of questions and answers. They then gave each participant a binary well-being label like the Vivago label for the timeframe in which the participant was examined (e.g. Wednesday afternoon). This label will be referred to as the *coach label* throughout the paper.

The third and final well-being labels are those determined through self-report. This well-being was determined through a survey that participants took twice daily for the duration of the observational study: once in the morning, and once in the evening. As with the coach label, the well-being labels again correspond to the respective timeframe for which the survey was taken. The final binary labels from the self-report were determined through some data manipulation that will be discussed in more detail in section 2.3 *Data preparation*, but these well-being labels will be referred to as the *self-report label* throughout the paper. As the survey was designed specifically for this research, the next section will elaborate on its design in more detail.

2.2. Survey design

The survey is a list of eleven straight-forward questions that measure the participant's general well-being. As a starting point, the survey used the Global Health PROMIS survey as introduced in the *Introduction*. It was next tailored to the specific needs of the observational study as will be described further in this section.

Firstly, the original PROMIS survey was translated into Dutch, because this is the native language of the participant group. Next, the formulation of all questions was adjusted to reflect the last few hours prior to taking the survey, so that the survey was better suited to be taken twice daily. This was necessary because the original PROMIS was intended to be taken a lot less often with questions referring to health *in general* or for *the last seven days* (both of which would not have changed within a day). For this same reason, question 2 of the original PROMIS survey was removed, as it asked about general quality of life. Finally, question 4 was split into two separate questions to get a more representative outcome. Originally it asked for both *mood* and *ability to think* in one question, but in the revised survey these are each their own question.

The survey was made with Google Forms and, just as the original PROMIS survey, it gives a final score that gives information about a participant's global health (which corresponds to *the general well-being* referred to throughout the thesis). The survey consists of 11 questions in total.

Each question results in its own score, and the scores are ordinal data. Questions 1 to 10 are on a five-point discrete scale from well to unwell (where each point is presented in the survey by a descriptor such as Fair or Poor), while question 11 is a ten-point discrete scale (a numeric value indication of pain from worst imaginable to non-existent).

The final score for global health is then determined by summing the scores of each question. The minimum score a participant can get is 10 and the maximum is 60. Here, higher scores indicate doing worse on your global health.

In addition to the answers to the questions, the survey records the date and time it was taken, the date and time the participants say they are reporting on (twice per day), and their participant identification number. See section B *Survey* for the full survey.

2.3. Data preparation

The final scores from the survey cannot be compared to the Vivago and coach well-being labels as they are because they are not in the same format. It is not possible to do any statistical comparison about the difference between a well-being score of 60 and a binary well-being label of 1. The PROMIS answers thus need to be transformed into binary labels of well-being, so that all well-being labels are of the same format.

To binarize the PROMIS answers, firstly the final scores (*fs*) are taken for each participant (*p*). Next, it is necessary to determine when such a score translates to being well (1), and when it translates to being unwell (0). It is said that a participant is unwell when they feel worse than the average of the participant group, and vice versa, they are well when they feel better than the average participant.

In the actual calculation, the average used to determine the binary label is the median. First, the final scores are scaled to a value between 0 and 1 to take into account that the lowest possible scores are not zero with formula 2.1, and then a binary label is given using the median scaled final score as a cut-off value. If the participant's scaled score is lower than the median, the score is interpreted as unwell (0). If the score is better than or at the median, it is interpreted as well (1).

$$scaled_p = \frac{fs_p - min(fs_p)}{max(fs_p) - min(fs_p)} \tag{2.1}$$

The median split is used (as opposed to the mean or mode) because there was no natural split in the data distribution, which was unimodal, as can be seen in Figure 2.1. Here, the x-axis shows the possible final scores, while the y-axis show their frequency. It is visible that the median, 32, is at the half point for all final scores.

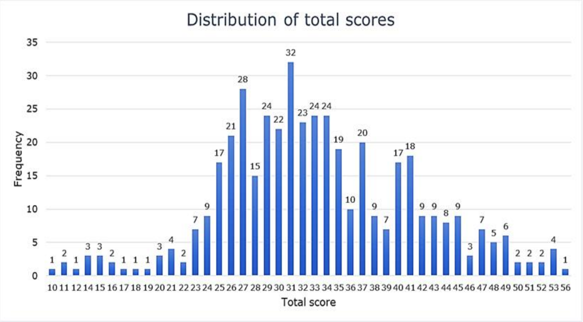


Figure 2.1: The distribution of the total scores of global health for all participants.

With all three of the sets of well-being labels generated, there remains only one problem: The Vivago labels are hourly, while the coach and self-report labels are twice daily. To remedy this, the twice daily datapoints of the coach and self-report labels are replicated to the hours they reflect. These hours are divided into two daily time periods. Looking at the survey, the time periods identified are from 10 P.M. to 9 A.M. ("morning") and from 10 A.M. to 9 P.M. ("afternoon"). Thus, a person who was feeling unwell (0) in the morning period of Wednesday is assumed to have been feeling unwell (gets a 0 label) for every hour from 10 P.M. to 9 A.M. on that date. The same logic is applied to an afternoon period, but then with the hours 10 A.M. to 9 P.M.

Essentially, after these manipulations the sensor data and labels can be used to create the classifiers. However, there is some final data preparation still necessary for the sensor data. The data can mostly be used as is, but the hourly data for which the watch is off for more than 0 minutes need to be removed, as these data become an unreliable hourly measure. It is also necessary to remove any NAN values from the data, which are values where something went wrong with the data gathering process. Once this is done, the classifiers can be created.

2.4. Classifiers

Having clarified how all the data is gathered and the well-being labels are established, it is possible to go more in-depth on the classifiers. While there are multiple common classifiers in the field of Artificial Intelligence that could be used to predict well-being from sensor data, this research focuses on the algorithms K-Nearest Neighbour (*KNN*) and Random Forest (*RF*). These show the best performance in initial exploratory research. In addition, literature also shows these classifiers have good performance for low dimensional data (Amancio, et al., 2014). Data are considered low dimensional when fewer features (which are data used by the classifier to make predictions) are gathered than that there are datapoints, and this is true for this research as will become clear when the features are described in more detail later. Firstly, a short explanation of how KNN and RF work.

The idea behind K-Nearest Neighbour is that the features of an inputted datapoint are likely to be similar to those of its neighbours. The datapoint is compared against *k* of its neighbours for similarity, which results in *k* neighbourhoods that contain similar data. This similarity is computed by the Euclidian distance D(x_1, x_2) between datapoint x_1 and datapoint x_2.

While K-Nearest Neighbour is a learning algorithm based on statistics, Random Forest uses decision trees for classification. A decision tree is a graph type where each node represents a decision, and each child is a consequence of said decision. Decision trees are a graphic method of describing a problem. When decision trees are used for classification, a decision is made at each node until the correct class for a datapoint is determined (Russel & Norvig, 2003).

Although the two algorithms work differently, both are forms of supervised learning. The processes described in the previous two paragraphs are how the classifiers train themselves to learn the correct data-to-label mapping. Of course, to know if it has succeeded in this, the classifier needs to be tested. In the testing phase, the classifier predicts labels based on the data and then compares it against the actual labels that represent well-being. For this research, the well-being is represented by the Vivago label during training and testing. The classifier predictions will also be tested against the coach and self-report labels, however. This is done to check the validity of using the Vivago labels to represent well-being.

Next to the well-being labels, the most important part of the classifiers is the inputted data. In section 2.1 *Observational study design*, all the data gathered is given. From this data, not all are chosen as final features for the prediction of well-being by the classifiers. The primary data chosen as predictors are the hourly values of *sleep* and *time outside* (in minutes), *activity* (as a percentage of maximum activity), and the *circadian* (a daily variable). Additionally, the daily totals of *sleep*, *time outside*, and *activity* are included as features to add dimensionality to the data. This is done because classifiers generally need a lot of data, and exploratory research showed higher dimensionality led to better performance.

The data not included as features in the classifier design are the sensor data for *sleep* and *time outside* given in hours, and the watch-off measure. The sensor data in hours are not included as they are the same data as the health data in minutes, but less precise as values such as 1/3rd are rounded to the 16th decimal. The watch-off data are not included because they have no reflection on well-being.

With the data and well-being labels established, there is only one part left to the design of the classifier, which are methods to improve generalisability. For classification it is the norm to hold back a part of the training data so that the classifier performance can be checked against data it has never seen before. This is done to prevent the classifiers from overfitting, working only for the specific data they are provided with, instead of predicting well on all possible future data. To implement this training check, KNN will use 10-fold cross-validation, and Random Forest will use Out-of-Bag (OOB) estimation. Both these methods will generate multiple models of the same classifier, eventually finding the model with the best performance. For KNN, this means finding the optimal *k* neighbours to compare each datapoint against. For RF, it means finding the optimal amount of features to use at each tree split.

The testing of the optimal models for each classifier will produce accuracy scores that show how many datapoints are categorised correctly, as well as sensitivity and specificity scores that show how well the classifier can predict zeroes (being unwell) and ones (being well) respectively. It is important to know this distinction, because a classifier that always correctly predicts when someone is well but never correctly predicts when someone is unwell is of no use, and vice versa. The sensitivity and specificity scores will also show if the classifier is overpredicting either being well or being unwell.

Additionally, the testing results in a set of predicted labels, which can be compared against the already established well-being labels. Both classifiers will be tested against the Vivago labels, but only the classifier with the best performance (either KNN or RF) will be compared against the coach and self-report labels.

To check whether the difference between the predicted labels and the correct labels is significant, a McNemar's test is used. This test compares the consistency of the two label sets, that is to say, it looks at how many zeroes became ones (actually unwell predicted as well) and vice versa (actually well predicted as unwell). The null hypothesis of such a test is then that there is no significant difference between the two label sets (or in other words: the amounts of zeroes and ones stayed consistent).

3. Results

3.1. Predicting well-being

For the distribution of the well-being labels, the Vivago labels show 75% of the datapoints correspond to being well and 25% correspond to being unwell. The coach labels are similar, showing 77% datapoints as being well and 23% as unwell. The self-report labels show a completely different distribution of 53% datapoints corresponding to being well to 47% corresponding to being unwell. In other words, the Vivago and coach labels are both biased towards being well, while the self-report labels are reasonably balanced.

The testing of the classifiers using the data described in section 2.4 *Classifiers* and the Vivago labels results in the values visible in Figure 3.1. These are the results for a K-Nearest Neighbour (KNN) classifier with *k = 3* and a Random Forest (RF) classifier that uses *3* features at each tree split.

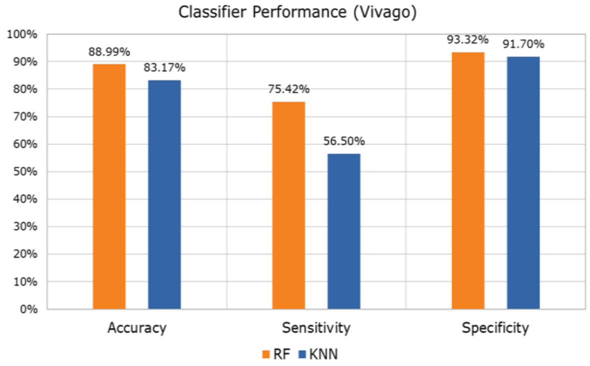


Figure 3.1: Performance of K-Nearest-Neighbour (KNN, in blue) and Random Forest (RF, in orange) with Vivago labels.

The figure shows that RF (in orange) performs better than KNN (in blue) on all metrics, but especially on sensitivity. In simpler terms, this means RF is much better at predicting when someone is unwell than the KNN classifier.

Although RF has a higher performance than KNN, both classifiers exhibit a much higher specificity than sensitivity. What this translates to is that both classifiers are better at predicting when someone is well than when someone is unwell, according to the Vivago well-being labels.

Since neither classifiers show a hundred percent accuracy, the statistical difference between the predicted labels and true labels

(Vivago labels) is tested through a McNemar's test, whose results are shown in Table 3.1.

Table 3.1: Statistical comparison of consistency along predicted and Vivago labels.

RF	**Vivago labels**		**McNemar's test p-value**
Predicted labels	0 *(unwell)*	1 *(well)*	0.3443
0 *(unwell)*	267	74	
1 *(well)*	87	1034	

KNN	**Vivago labels**		**McNemar's test p-value**
Predicted labels	0 *(unwell)*	1 *(well)*	1.006e-04
0 *(unwell)*	200	92	
1 *(well)*	154	1016	

If we take a standard α of 0.05, the results for the statistical comparison show that the null hypothesis can only be rejected for the KNN classifier. In other words, the predicted labels of the KNN classifier are statistically different from the Vivago labels. This means that despite of its 83.17% accuracy, its predictions ultimately do not match well-being assessed through sensor data.

For the Random Forest classifier, however, the p-value for the McNemar's test is greater than 0.05, and therefore the null hypothesis cannot be rejected. This means that, for the Random Forest classifier, there is no evidence to suggest that the predicted labels are significantly different from the well-being labels assessed through sensor data.

3.2. Comparing Vivago labels against the coach and self-report labels

The results in the previous section show how well the classifiers can make predictions of well-being with the assumption that the Vivago labels accurately represent well-being. To check the validity of this assumption, the Random Forest classifier is also tested against the coach (in yellow) and self-report (in blue) labels. The results of this test can be seen in Figure 3.2.

The results show that the performance of the RF classifier decreases across all metrics for both the coach and self-report labels. Despite of this decrease, the values for specificity remain high, indicating an RF classifier trained with Vivago labels is good at predicting when someone is well. The same cannot be said for predicting when someone is unwell, however, as the decrease in sensitivity from testing against the Vivago labels to testing against the coach and self-report labels is quite drastic.

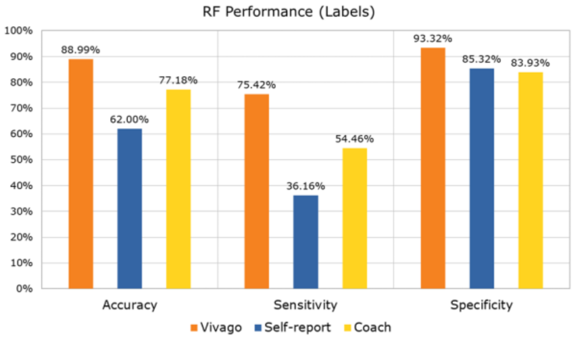


Figure 3.2: Comparison of Random Forest (RF) performance between Vivago (in orange), self-report (in yellow), and coach (in blue) labels.

To draw any other conclusion from these results, it is useful to have an idea of how much the coach and self-report labels already differed from the Vivago label initially (before any classification). The initial distributions of the coach and self-report labels when compared against the Vivago label can be seen in Table 3.2. The table also includes a McNemar's test for the significance of any differences with the Vivago labels.

The table shows that both the coach and self-report labels return p-values less than 0.05 for their McNemar's test. This means that both the coach and self-report labels reject the null hypothesis of the McNemar's test. It can thus be said that there is already a significant difference between both the coach and self-report label when compared against the Vivago label before any classification has even taken place.

Table 3.2: Statistical comparison of consistency between the Vivago labels and coach and self-report labels respectively.

	Coach labels		**McNemar's test p-value**
Vivago labels	0 *(unwell)*	1 *(well)*	2.33e-03
0 *(unwell)*	543	503	
1 *(well)*	410	2702	

	Self-report labels		**McNemar's test p-value**
Vivago labels	0 *(unwell)*	1 *(well)*	< 2.20e-16
0 *(unwell)*	735	311	
1 *(well)*	1237	1875	

Although it might be intuitive that the predicted labels of the classifier will also be statistically different from the Vivago labels, it is important to still check whether this is really the case. In table 3.3, the same statistical comparison is depicted as in table 3.2, but here the coach and self-report labels are compared to the predicted labels of the classifier (rather than the Vivago labels themselves).

Table 3.3 shows that, indeed, for both the coach and self-report labels, the null hypothesis of the McNemar's test is still rejected, as both labels return a p-value less than 0.05. The predictions that the classifier makes are thus significantly different from the well-being assessed through health professionals and self-report. The table shows the predicted labels differ from the coach and self-report labels even more than the Vivago labels did, as there are more predicted zeroes that are actually supposed to be ones according to the respective well-being labels and more ones that are actually supposed to be zeroes.

Table 3.3: Statistical comparison of consistency between the predicted labels and nurse and self-report labels respectively.

	Coach labels		**McNemar's test p-value**
Predicted labels	0 *(unwell)*	1 *(well)*	9.41e-03
0 *(unwell)*	519	515	
1 *(well)*	434	2690	

	Self-report labels		**McNemar's test p-value**
Predicted labels	0 *(unwell)*	1 *(well)*	< 2.20e-16
0 *(unwell)*	713	321	
1 *(well)*	1259	1865	

4. Discussion and Conclusion

The results show that a Random Forest classifier that uses 3 features per tree split and Out-of-Bag estimation can predict well-being on the basis of sleep, physical activity, time away from home, and the circadian rhythm with an 89% accuracy. It can predict when someone is well correctly 93% of the time, and whether someone is unwell 75% of the time. Ideally, one would rather see a higher sensitivity than specificity in the health industry because interfering medically when it is not necessary is largely preferable over not interfering when it is, but the overall accuracy and specificity are quite high.

Despite of this quite good classifier performance, the results also show that both the initial Vivago labels and the predicted labels from the classifier are statistically different from the coach and self-report labels. Furthermore, a classifier trained on the Vivago labels will have a much lower overall performance when compared against the coach and self-report labels. Thus, the well-being labels based on the sensor data do not match the well-being assessed through health professionals and self-report.

Conclusively, despite of the good results of the RF classifier, the well-being label that it is trained on is considered unrepresentative of well-being. As such, in response to the research question "*Can classifiers be used to make predictions of well-being on the basis of sleep, physical activity, time away from home, and the circadian rhythm?*": **Even the optimal classifier designed in this research cannot be used to make reliable predictions of well-being on the basis of sleep, physical activity, time away from home, and the circadian rhythm.**

There are ways the classifiers can be further improved, however. One of the largest problems of even the optimal classifier is its sensitivity. To improve the sensitivity, it is necessary to look at possible causes for why it is low.

One possible cause is the imbalance in the Vivago labels between the amount of information on being well and the amount of information on being unwell. The KNN classifier deals with the imbalance through the cross-validation; it balances each fold to have approximately the same amount of information on both being well and being unwell. The better performing RF classifier, however, does not deal with the imbalance in this research. This results in it overpredicting being well, because it has the most information on that class.

There are essentially two ways to solve this problem for RF. The first solution is to adjust the weights of each class so that the class for being unwell is given more importance than the class for being well. This makes sense in this research, as predicting being unwell is indeed more important than predicting being well.

More ideal, however, would be for the classifier to predict both being well and unwell accurately, and a different solution is sampling the data differently. It is possible to oversample the minority class and undersample the majority class. Oversampling the minority class means fabricating more datapoints from the existing datapoints for being unwell through replication. These new datapoints will have exactly the same information as the already existing datapoints. It solves the imbalance, but increases the risk of overfitting the classifier.

The other way of sampling is undersampling the majority class, which means taking only a partial amount of datapoints that correspond to being well (usually the same amount as there are datapoints for being unwell). This is a good solution, but it reduces the total amount of data the classifier has to make predictions, and typically the more data classifiers have, the more accurately they can learn how the datapoints correspond to the labels.

This brings up the other solution to dealing with the imbalanced data: making the data more balanced. In the case of this research, that means getting more information for how sensor data correlates to being unwell. This might be difficult in another observational study considering the target group for the gathering of sensor data is elderly people who are generally doing well, but a longer observational study would certainly give a bit more data, which could go well in combination with sampling techniques.

Alternatively, instead of an observational study, it could be possible to take sensor data from existing databases of wearables. Most wearables have a history of sensor data saved for the users for various time periods. Using this as a way of gathering the necessary sensor data could help with gathering more balanced data.

Of course, if sensor data is gathered in such a way, it would still need to be labelled as either being well or being unwell. If the sensor data is not gathered real-time, it is not possible to assess the well-being for the same time points through health professionals or self-report. Even in the case of a longer observational study, as the amount of data increases, the burden on health professionals and patients performing self-report increases. As such, assessing well-being on the basis of data would be the ideal way of creating the necessary labels, but this research shows that well-being assessed through sensor data is not representative of well-being. It is the second reason for why the classifiers in this research cannot be used to make predictions of well-being.

As with the low sensitivity, there are some causes that could explain these results. One cause is the way the twice daily coach and self-report labels were replicated to match the hourly Vivago labels. This assumes that well-being can as easily be assessed at an hourly level as a twice daily level, but this might not be the case. After all, how much does well-being change at an hourly level? Is it typical to see large fluctuations in well-being between one hour and the next? If the well-being was defined in a more continuous format, perhaps differences between the hours would be relevant, but as a binary label, the hourly changes are likely to say less about general well-being than a twice daily or even daily label.

The idea of a daily time sampling being better for assessing well-being is supported by previous research as well. To determine how sleep affected depression, for example, Wang et al. looked at daily averages for amount of sleep, wake time, and bed time. As well as how much time in total was spent at home, and how much was spent away from home (Wang, et al., 2018). The same trend can be seen in the other literature that looks at the effect of health factors on well-being. Therefore, if this research is to be repeated, it would be better to assess well-being through sensor data on a daily basis. This well-being might then more closely match the well-beings determined through health professionals and self-report.

In fact, this research already uses daily sensor data, namely the daily totals for *sleep, physical activity,* and *time away from home.* The circadian is also a daily variable. The good thing about the circadian is that it considers the effect of time on the other sensor data. For example, the circadian will be better if the hourly sleep measurement showed more sleep during the night hours than during day hours. By using the circadian, hourly information about the other health data is not entirely lost.

If well-being is better assessed at a daily level, then research should show that the daily totals and the circadian have a larger effect on well-being than the hourly measurements. Unfortunately, this research fails to examine the effect of each of the various features used by the classifiers. During the cross-validation and OOB estimation, the tuning parameters of each model are tweaked so that after the cross-validation and OOB estimation, the optimal model is left. This means that both the models chosen for KNN and RF use the optimal amount of features, however, it is not known which features these are exactly. It would thus be good to do a follow-up research that looks at how each feature impacts well-being. Such a research would also add to the growing literature on how certain sensor data affect well-being.

In summary, from this research it cannot be concluded that classifiers can be used to predict well-being on the basis of sleep, physical activity, time away from home, and the circadian rhythm, as the classifiers in this research overpredict being well, while the ability to predict being unwell is arguably more important in the healthcare industry. Furthermore, this research shows that a well-being assessed through wearable sensor data is not representative of well-being as assessed by health professionals and self-report. This could indicate that perhaps the ideal time sampling of such sensor data is at a daily level.

Suggestions for improving the research in the future based on these results are: providing more data on being unwell to balance the dataset, assessing well-being through sensor data on a daily basis, and more thoroughly exploring the effect of each of the health features used.

If these improvements are implemented, the already high accuracy and specificity of the RF classifier gives hope that the idea of using classifiers to predict well-being using wearable sensor data might still lead to positive results. This would be very good, because the world population will continue to grow older and live longer, and the health industry thus very much needs a way to adapt to this phenomenon.

References

Amancio, D. R., Comin, C. H., Casanova, D., Travieso, G., Bruno, O. M., Rodrigues, F. A., & da Fontoura Costa, L. (2014). A Systematic Comparison of Supervised Classifiers. *PLoS ONE, 9*(4). doi:10.1371/journal.pone.0094137

Balestroni, G., & Bertolotti, G. (2015). EuroQold-5D (EQ-5D): an instrumenet for measuring quality of life. *Monaldi Archives for Chest Disease, 78*(3). doi:10.4081/monaldi.2012.121

Blondell, S. J., Hammersley-Mather, R., & Veerman, J. (2014, May). Does physical activity prevent cognitive decline and dementia?: A systematic review and meta-analysis of longitudinal studies. *BMC public health, 14*, 510. doi:10.1186/1471-2458-14-510

Brainard, J., Gobel, M., Scott, B., Koeppen, M., & Eckle, T. (2015, May). Health Implications of Disrupted Circadian Rhythms and the Potential for Daylight as Therapy. *Anesthesiology, 122*, 1170-1175. doi:10.1097/ALN.0000000000000596

Cecil, R. L., Goldman, L., & Schafer, A. I. (2012). *Goldman's Cecil Medicine* (24 ed.). Philadelphia: Elsevier/Saunders. doi:10.1016/C2009-0-42832-0

Crimmins, E. M., & Beltrán-Sánchez, H. (2011). Mortality and Morbidity Trends: Is there Compression of Morbidity? *The Journals of Gerontology: Series B, 66B*(1), 75-86. doi:10.1093/geronb/gbq088

Hays, R. D., Bjorner, J. B., Revicki, D. A., Spritzer, K. L., & Cella, D. (2009). Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (promis) global items. *Quality of life research : An international journal of quality of life, aspects of treatment, care and rehabilitation, 18*(7), 873-880. doi:10.1007/s11136-009-9496-9

HIMSS Europe GmbH. (2018). HIMSS Analytics Annual European eHealth Survey 2018. Europe. Retrieved on February 7, 2019, from HIMSS Europe website: https://www.himss.eu/himss-analytics-annual-european-ehealth-survey-2018

Huber, M., Knottnerus, J. A., Green, L. W., Jadad, A., van der Horst, H. E., Kromhout, D., . . . Smid, H. (2011, July 26). How should we define health? *BMJ*(343). doi:10.1136/bmj.d4163

Lötjönen, J., Korhonen, I., Hirvonen, K., Eskelinen, S., Myllymäki, M., & Partinen, M. (2003, January). Automatic Sleep-Wake and Nap Analysis with a New Wrist Worn Online

Activity Monitoring Device Vivago WristCare. *Sleep, 26*(1), 86-90. doi:10.1093/sleep/26.1.86

Paterson, D., & Warburton, D. (2010, May). Physical activity and the functional limitations in older adults: A systematic review related to Canada's Physical Activity Guidelines. *The international journal of behavioral nutrition and physical activity, 7*, 38. doi:10.1186/1479-5868-7-38

Russel, S. J., & Norvig, P. (2003). *Artificial Intelligence, A Morden Approach* (2nd ed.). Prentice Hall.

Salgado-Delgado, R., Tapia Osorio, A., Saderi, N., & Escobar, C. (2011). Disruption of Circadian Rhythms: A Crucial Factor in the Etiology of Depression. *Depression Research and Treatment*. doi:10.1155/2011/839743

Särelä, A., Korhonen, I., Lötjönen, J., Sola, M., & Myllymäki, M. (2003). IST Vivago (R) - an intelligent social and remote wellness monitoring system for the elderly. *4th International IEEE EMBS Special Topic Conference on Information Technology Applications in Biomedicine, 2003.*, (pp. 362-365). doi:10.1109/ITAB.2003.1222554

Spiegel, K., Leproult, R., & Van Cauter, E. (1999). Impact of sleep debt on metabolic and endocrine function. *The Lancet, 354*, 1435-1439. doi:https://doi.org/10.1016/S0140-6736(99)01376-8

United Nations Department of Social and Economic Affairs. (2007). *World Economic and Social Survey 2007: Development in an Ageing World.* New York: United Nations. Retrieved on February 6, 2018, from http://www.un.org/en/development/desa/policy/wess/wess_archive/2007wess.pdf

Vanhelst, J., Hurdiel, R., Mikulovic, J., Bui-Xuân, G., Fardy, P., Theunynck, D., & Béghin, L. (2012). Validation of the Vivago Wrist-Worn accelerometer in the assessment of physical activity. *BMC Public Health, 12*, 690. doi:10.1186/1471-2458-12-690

Vinogradova, I. A., Anisimov, V. N., Bukalev, A. V., Ilyukha, V. A., Khizhkin, E. A., Lotosh, T. A., . . . Zabezhinski, M. A. (2010). Circadian disruption induced by light-at-night accelerates aging and promotes tumorigenesis in young but not in old rats. *Aging (Albany NY), 2*(2), 82-92. doi:10.18632/aging.100120

Wang, R., Wang, W., daSilva, A., Huckins, J. F., Kelley, W. M., Heatherton, T. F., & Campbell, A. T. (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). doi:10.1145/3191775

World Health Organization. (2015). *World Report on Ageing and Health.* Retrieved from https://www.who.int/ageing/publications/world-report-2015/en/

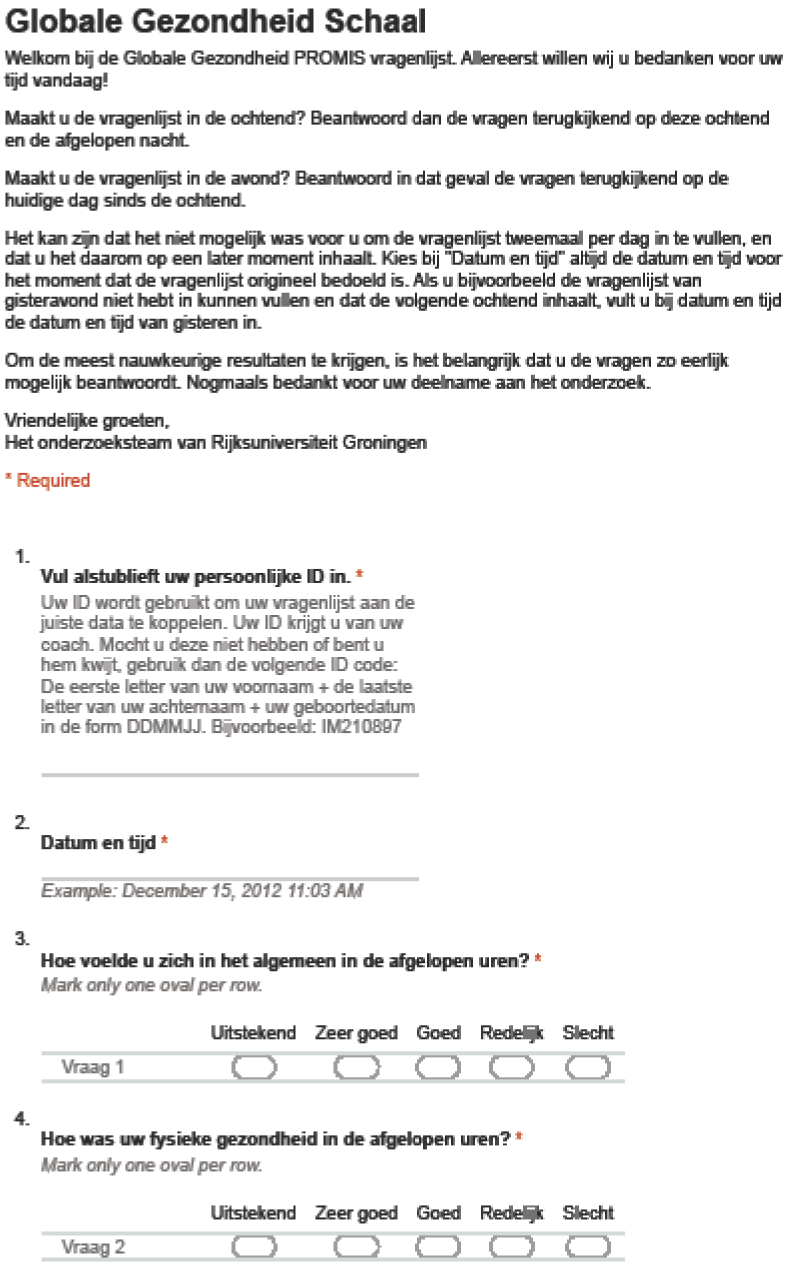
A. Classifier code in R

```
######### R Code for KNN and Random Forest on dataset ###########
######### Uses caret package, written by: I.D.M. Akrum ##########

# Install packages, necessary only once per computer
install.packages('caret')
install.packages('ISLR')

# Set working directory to where the data is
setwd("/Data Location")
library(ISLR)
library(caret)

vivago <- read.csv("FinalData.csv")

################## Data Processing ########################
vivago <- vivago[, -c(16:18)] # Remove patient information
vivago <- vivago[, -c(11:13)] # Remove Watchoff variable
vivago <- vivago[, -1] # Remove Day variable

# Remove hourly version of all variables
vivago <- vivago[, -3]
vivago <- vivago[, -7]
vivago <- na.omit(vivago)
head(vivago)

# Set train-test split
set.seed(400)
indxTrain <- createDataPartition(vivago$vivagoLabel, p = 2/3, list = FALSE)
training <- vivago[indxTrain,]
testing <- vivago[-indxTrain,]

# Checking the label distributions in the data sets
prop.table(table(training$vivagoLabel)) * 100
prop.table(table(testing$vivagoLabel)) * 100
prop.table(table(vivago$vivagoLabel)) * 100

# Train the KNN classifier
set.seed(400)
ctrl_knn <- trainControl(method="cv", number = 10)
knnFit <- train(as.factor(vivagoLabel) ~ ., data = training, method = "knn", trControl = ctrl_knn, preProcess = c("center","scale"), tuneLength = 20)
knnFit

plot(knnFit, type="b", main="KNN Accuracy", col="blue", xlab="#Neighbours", ylab = "Accuracy (10-Fold Cross-Validation)")
```

```
# Test the KNN classifier
set.seed(400)
knnPredict <- predict(knnFit, newdata = testing)
knn_cf <- confusionMatrix(knnPredict, as.factor(testing$vivagoLabel), positive = "0")
knn_cf

################## Random Forest ########################
# Train the Random Forest classifier
set.seed(400)
ctrl_rf <- trainControl(method="oob")
rfFit <- train(as.factor(vivagoLabel) ~ ., data = training, method = "rf", trControl = ctrl_rf, preProcess =
c("center","scale"), tuneLength = 20)
rfFit

plot(rfFit, type="b", main="RF Accuracy", xlab="#predictors", ylab = "Accuracy (Out-Of-Bag Estimate)")

# Test the Random Forest classifier
set.seed(400)
rfPredict <- predict(rfFit, newdata = testing)
rf_cf <- confusionMatrix(rfPredict, as.factor(testing$vivagoLabel), positive = "0")
rf_cf

############## Comparison With Other Labels #####################
# Set testing <- original vivago data + all labels
testing <- read.csv("FinalDataAllLabels.csv")

# Remove the rows where the watch is off
testing <- testing[testing$watchOffMinutes==0, ]
testing <- testing[,-c(20:22)] # remove patient information
testing <- testing[, -1] #remove the day variable

# Omit NAs from test data, then save the labels in their own vectors
testing <- na.omit(testing)
nurseLabels <- testing$nurseLabel
surveyLabels <- testing$surveyLabel

# Checking the label distributions in testing and label vectors (here we'll see a completely different
distribution)
prop.table(table(testing$vivagoLabel)) * 100
prop.table(table(surveyLabels)) * 100
prop.table(table(nurseLabels)) * 100

# Check the difference between the labels without any classification
confusionMatrix(as.factor(testing$vivagoLabel), as.factor(nurseLabels), dnn = c("Vivago", "Nurse"))
confusionMatrix(as.factor(testing$vivagoLabel), as.factor(surveyLabels), dnn = c("Vivago", "Survey"))
```

```
# Predict on other labels with the Random Forest classifier
set.seed(400)
rfPredict2 <- predict(rfFit, newdata = testing)
rf_cf_survey <- confusionMatrix(rfPredict2, as.factor(surveyLabels), positive = "0")
rf_cf_mental <- confusionMatrix(rfPredict2, as.factor(testing$mentalLabel), positive = "0")
rf_cf_physical <- confusionMatrix(rfPredict2, as.factor(testing$physicalLabel), positive = "0")
rf_cf_nurse <- confusionMatrix(rfPredict2, as.factor(nurseLabels), positive = "0")

accuracies <- c(rf_cf_survey$overall[1], rf_cf_nurse$overall[1], rf_cf_mental$overall[1],
rf_cf_physical$overall[1])
sensitivities <- c(rf_cf_survey$byClass[1], rf_cf_nurse$byClass[1], rf_cf_mental$byClass[1],
rf_cf_physical$byClass[1])
specificities <- c(rf_cf_survey$byClass[2], rf_cf_nurse$byClass[2], rf_cf_mental$byClass[2],
rf_cf_physical$byClass[2])
rf_labels_table <- matrix(c(accuracies,sensitivities,specificities),nrow = 4,ncol = 3,
                    dimnames = list(c("Survey", "Nurse", "Mental", "Physical"), c("Accuracy", "Sensitivity",
"Specificity")))
```

B. Survey



Globale Gezondheid Schaal

Welkom bij de Globale Gezondheid PROMIS vragenlijst. Allereerst willen wij u bedanken voor uw tijd vandaag!

Maakt u de vragenlijst in de ochtend? Beantwoord dan de vragen terugkijkend op deze ochtend en de afgelopen nacht.

Maakt u de vragenlijst in de avond? Beantwoord in dat geval de vragen terugkijkend op de huidige dag sinds de ochtend.

Het kan zijn dat het niet mogelijk was voor u om de vragenlijst tweemaal per dag in te vullen, en dat u het daarom op een later moment inhaalt. Kies bij "Datum en tijd" altijd de datum en tijd voor het moment dat de vragenlijst origineel bedoeld is. Als u bijvoorbeeld de vragenlijst van gisteravond niet hebt in kunnen vullen en dat de volgende ochtend inhaalt, vult u bij datum en tijd de datum en tijd van gisteren in.

Om de meest nauwkeurige resultaten te krijgen, is het belangrijk dat u de vragen zo eerlijk mogelijk beantwoordt. Nogmaals bedankt voor uw deelname aan het onderzoek.

Vriendelijke groeten,
Het onderzoeksteam van Rijksuniversiteit Groningen

* Required

1. **Vul alstublieft uw persoonlijke ID in.** *

Uw ID wordt gebruikt om uw vragenlijst aan de juiste data te koppelen. Uw ID krijgt u van uw coach. Mocht u deze niet hebben of bent u hem kwijt, gebruik dan de volgende ID code: De eerste letter van uw voornaam + de laatste letter van uw achternaam + uw geboortedatum in de form DDMMJJ. Bijvoorbeeld: IM210897

2. **Datum en tijd** *

Example: December 15, 2012 11:03 AM

3. **Hoe voelde u zich in het algemeen in de afgelopen uren?** *

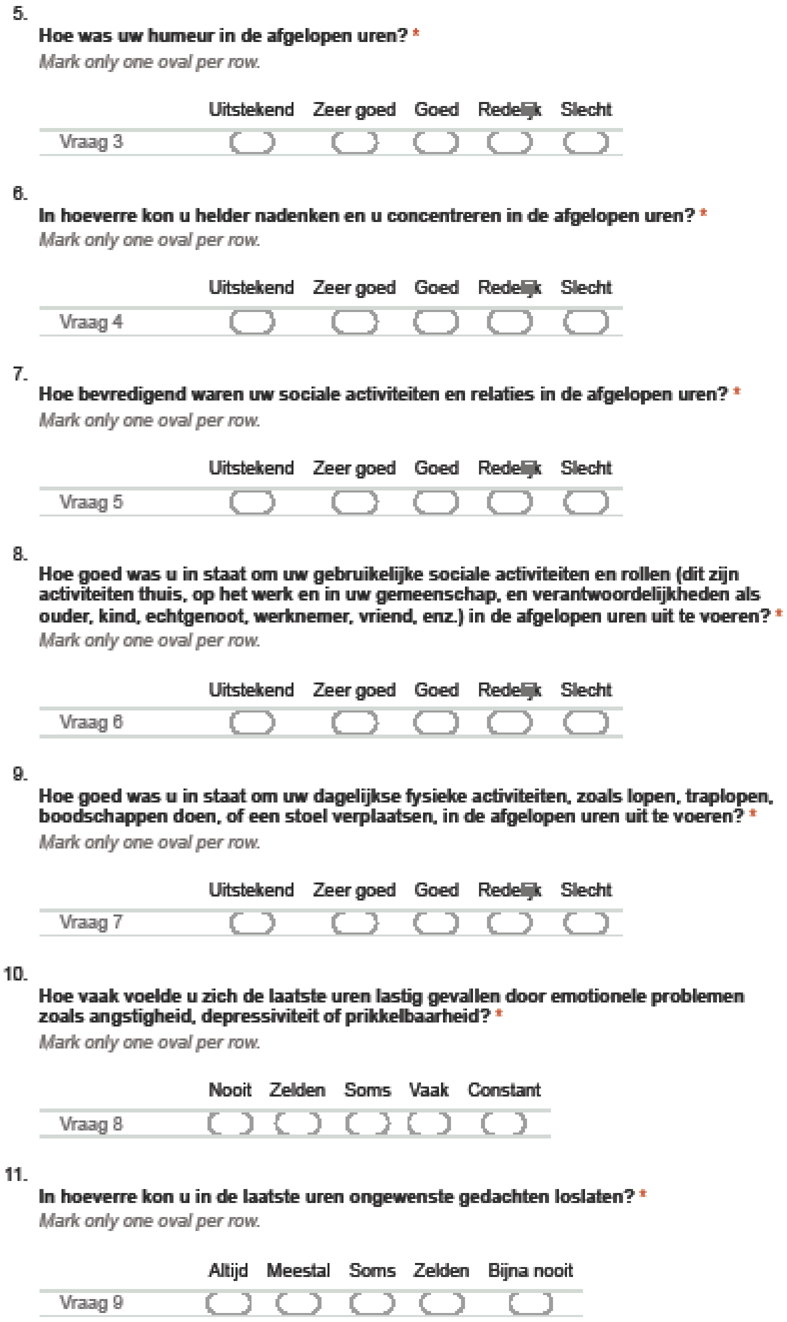
Mark only one oval per row.

	Uitstekend	Zeer goed	Goed	Redelijk	Slecht
Vraag 1	○	○	○	○	○

4. **Hoe was uw fysieke gezondheid in de afgelopen uren?** *

Mark only one oval per row.

	Uitstekend	Zeer goed	Goed	Redelijk	Slecht
Vraag 2	○	○	○	○	○



5.
Hoe was uw humeur in de afgelopen uren? *
Mark only one oval per row.

	Uitstekend	Zeer goed	Goed	Redelijk	Slecht
Vraag 3	○	○	○	○	○

6.
In hoeverre kon u helder nadenken en u concentreren in de afgelopen uren? *
Mark only one oval per row.

	Uitstekend	Zeer goed	Goed	Redelijk	Slecht
Vraag 4	○	○	○	○	○

7.
Hoe bevredigend waren uw sociale activiteiten en relaties in de afgelopen uren? *
Mark only one oval per row.

	Uitstekend	Zeer goed	Goed	Redelijk	Slecht
Vraag 5	○	○	○	○	○

8.
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? *
Mark only one oval per row.

	Uitstekend	Zeer goed	Goed	Redelijk	Slecht
Vraag 6	○	○	○	○	○

9.
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? *
Mark only one oval per row.

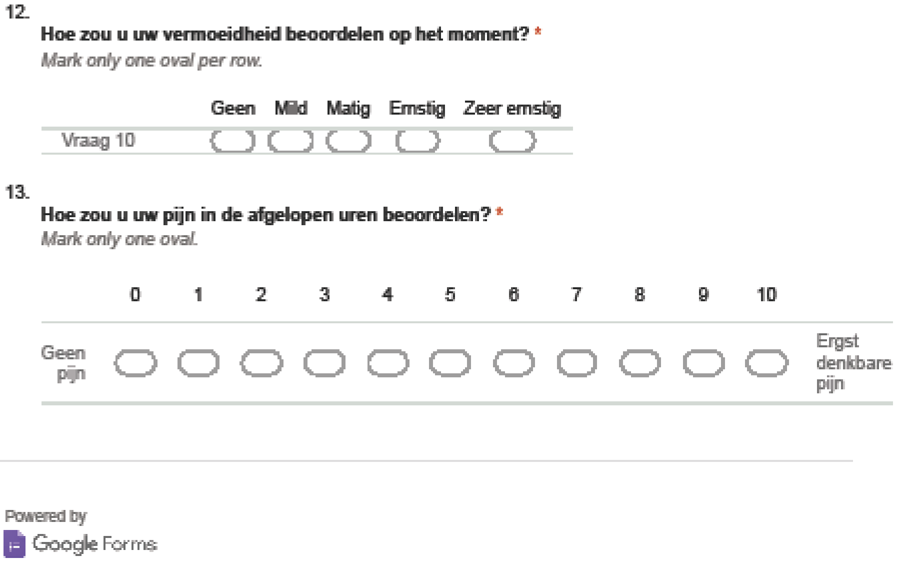
	Uitstekend	Zeer goed	Goed	Redelijk	Slecht
Vraag 7	○	○	○	○	○

10.
Hoe vaak voelde u zich de laatste uren lastig gevallen door emotionele problemen zoals angstigheid, depressiviteit of prikkelbaarheid? *
Mark only one oval per row.

	Nooit	Zelden	Soms	Vaak	Constant
Vraag 8	○	○	○	○	○

11.
In hoeverre kon u in de laatste uren ongewenste gedachten loslaten? *
Mark only one oval per row.

	Altijd	Meestal	Soms	Zelden	Bijna nooit
Vraag 9	○	○	○	○	○



12.

Hoe zou u uw vermoeidheid beoordelen op het moment? *

Mark only one oval per row.

	Geen	Mild	Matig	Ernstig	Zeer ernstig
Vraag 10	○	○	○	○	○

13.

Hoe zou u uw pijn in de afgelopen uren beoordelen? *

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Geen pijn	○	○	○	○	○	○	○	○	○	○	○	Ergst denkbare pijn

Powered by

Google Forms