Automatic Prediction of Health Status using Smartphone Derived Behaviour Profiles

Daniel Kelly[†], Kevin Curran[†], Brian Caulfield[‡]

***Abstract*—Objective: Current methods of assessing the affect a patients' health has on their daily life are extremely limited. The aim of this work is to develop a sensor based approach to health status measurement in order to objectively measure health status. Methods: Techniques to generate human behaviour profiles, derived from smartphone accelerometer and gyroscope sensors, are proposed. Experiments, using SVM regression models, are then conducted in order to evaluate the use of the proposed behaviour profiles as a predictor of health status. Results: Experiments were conducted on data from 171 participants, with an average of 114 hours of data per participant. Regression models were trained and tested on the 10 SF-36 self-ratings. Results showed that the 8 individual SF-36 scales and 2 component scores could be predicted with an average correlation of 0.683 and 0.698 respectively. General Health was predicted with an average correlation of 0.752. Conclusion: Research shows that the Clinically Important Difference for SF-36 self-ratings are approximately 10 points. Health status prediction errors in this work were 11.7 points on average. While the problem has not been fully solved, this work present a hugely promising direction for health status prediction. Significance: Using the proposed techniques, health status could be measured using unobtrusive, inexpensive and already available hardware. It could provide a means for clinicians to accurately and objectively assess the daily life benefits of treatments on an individual patient basis.**

I. Introduction

CHRONIC diseases are the most common causes of death and disability throughout the world [1]. In the UK, for example, 70% of all healthcare costs are chronic disease related [2]. Treatments for chronic diseases, such as medication and lifestyle change, can result in improved clinical health outcome measures such as physiological measurements, hospital re-admission rates or mortality rates. However, while these outcome measures are important, they do not fully represent the experiences of an individual patient and their response to particular treatments [3]. Health status measurements, such as Health Related QOL (HRQOL), are patient reported measures used as a means of quantifying the impact of disease on patients daily life [4]. These measures have become a central feature in many chronic disease studies [5]. Research has shown that poor health status scores were associated with mortality, hospital readmission and increased healthcare consumption [4].

Studies on health status questionnaire reliability indicate that, while most measures are reliable for group comparisons, the majority of measure are not reliable for individual comparisons. For example, 7 of the 8 measures from the Short Form 36 (SF-36) health status questionnaire were consistently shown to not be of sufficient reliability to assess patients on an individual basis [6]. There is therefore a need for more accurate and reliable methods of measuring health status such that clinicians can assess health status on individual basis. The overall aim of our work is to utilize novel sensor technology in the community to objectively measure a persons health status.

Modern smartphones, equipped with multiple sensors built within the common and non-invasive form factor of a mobile phone, have the potential of tracing human activities at scales that were previously unattainable. The aim of this work is to develop an unobtrusive smartphone sensing system which can objectively measure a persons' longitudinal behaviour and make accurate predictions about their health status based on their behaviour. In order to accurately model the mapping between mobile sensor data and health status, a participant set, with a broad spectrum of health measurements, is required. Recording patient data alone would represent a small window in the health status spectrum. Thus, before patient specific investigations are carried out, general methods of mapping sensor data to health status must be investigated. Therefore, the aim of this work is to analyse daily life measurements, and health status information, from adults in the general population with a diverse set of health measurements. We propose using motion sensors, built within a smart-phone, to measure a persons behaviour by recording longitudinal motion patterns. Our hypothesis is that motion sensor data, through a procedure of feature extraction and machine learning, can be used to predict the health status of a person. To test our hypothesis, we conduct a study in which participants record their motion patterns using a smart-phone and record their health status using a self-reported questionnaire. Experiments are then performed on the data to discover if, and to what extent, behaviour based features, extracted from recorded motion data, can be used to predict health status.

A. Related Work

To the authors knowledge, there are no related works specifically investigating methods to automatically predict patient reported health outcomes, such as health status, using unobtrusive motion sensing. Motion sensing has however been utilized for different types of health based monitoring [7]. Studies have mainly been conducted in controlled conditions with patients wearing specialized sensors [8]. For example, a number of works have used multiple motion sensors to develop instrumented versions of the timed up-and-go test for identifying gait impairments related to Parkinsons Disease [9]

[†]D. Kelly and K. Curran are with the School of Computing and Intelligent Systems, Ulster University, Northern Ireland e-mail: d.kelly@ulster.ac.uk
[‡]B. Caulfield is with the INSIGHT Center, University College Dublin

and falls risks [10]. Cook et al. [11] conducted a study to analyse the impact health conditions, Parkinsons Disease in particular, have on daily behaviour. Subject behaviour was monitored using a combination of smart home environment sensors and motion sensors. Feature extraction and machine learning techniques were applied to the data. Results showed that statistically significant behaviour differences existed between two groups and that the two groups could be recognized automatically by a machine learning classifier. Recently, research has shown that smartphones can be utilized to infer health related information. For example, Juen et al. describe a smartphone based walking monitor for patients with Chronic Obstructive Pulmonary Disease (COPD) [12]. Kelly et al. [13] conducted a case series, performing a preliminary investigation on differences in movement patterns of COPD patients reporting problems versus COPD not reporting problems.

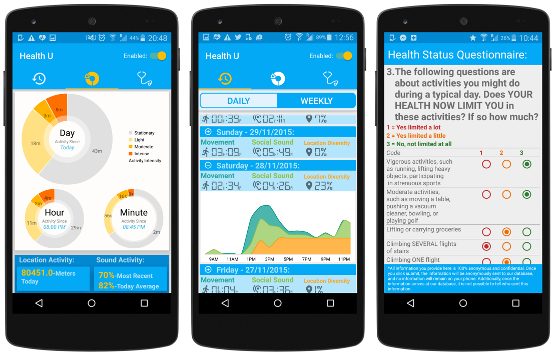
In terms of automatic methods to predict health status, there has been a number of works investigating the use of machine learning models for the prediction of health status. However, none of these works utilize automatically generated observations, such a motion sensing, in order to make health predictions. One such work is that of Yang et al. [14], where experiments are carried out to evaluate machine learning techniques as predictors of two health status metrics (Cornell Scale for Depression in Dementia (CSDD) and Physical Self-Maintenance Scale (PSMS)). Experiments were based on 15 Geriatric Patients and feature vectors, describing patient behaviour, where manually built using visual observation of each patient using video footage. The prediction problem was mapped to a binary problem, where the aim of the classifier was to classify sub-categories of the health status scores as good or bad. Average precision rates of 0.86 and 0.91 were reported for the classification of categories in the CSDD and PSMS health status scales respectively. Paskhomov et al. [15] use automated natural language techniques to extract health descriptors from manually recorded patient medical records. Machine learning techniques were utilized in order to make predictions about patient health status, as measured using SF-36 and EuroQol five dimensions (EQ5D) measurement tools. Experiments show 'moderate' agreement between patient reported health status and machine learning predicted health status. The best concordance between automatic classification and patient reported health status was achieved for the 'pain' component, with a positive and negative agreement of 0.76 and 0.78 respectively (using Cohens Kappa coefficient).

While there is a large body of research work in the general area of sensors and health and well-being, there exists few works dealing specifically with automatic prediction of health status without the need for costly, time consuming and invasive manual observations. In this work, we perform an investigation into the use of smartphones as a method of automatically generating behaviour observations for the purpose of health status prediction. In particular, we aim to investigate the use of non-invasive motion sensors to identify links between a persons' activity and a person' health status.

II. Methods

A. Data Collection

A custom Android App, named "Health-U", was developed in order to facilitate a crowdsourced approach to motion sensor and health status data collection. This App was published on Google Play, allowing anyone with an Android phone to participate in the study. The App was designed to record raw Accelerometer and Gyroscope data throughout the day for each participant. Raw sensor data is processed at the end of each day and summary measures, describing movement profiles for each hour, are generated and uploaded to a central server. To improve user retention within the experiment, functionality was added to the App to provide users with visual feedback on the duration and intensity of their activities over time using graphs and statistics (see Figure 1).



Health Status Questionnaire:

3.The following questions are about activities you might do during a typical day. Does YOUR HEALTH NOW LIMIT YOU in these activities? If so how much?

1 = Yes limited a lot
2 = Yes limited a little
3 = No, not limited at all

Code	1	2	3
Vigerous activities, such as running, lifting heavy objects, participating in strenuous sports	○	○	●
Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	○	○	●
Lifting or carrying groceries	○	●	○
Climbing SEVERAL flights of stairs	●	○	○
Climbing ONE flight	○	●	○

*All information you provide here is 100% anonymous and confidential. Once you click submit, the information will be anonymously sent to our database, and no information will remain on your phone. Additionally, once the information arrives at our database, it is not possible to tell who sent this information.

Fig. 1. "Health-U" App - (Left) Visual feedback showing current activity, (Middle) Activity history showing daily activity, (Right) Health Status Questionnaire.

The App was designed to include a health status measurement tool in order to record participant health status. A requirement of the study is to record data for a set of participants with a broad spectrum of health measurements. However, while participants can include adults from the general population for this study, our overall aim in the future is to evaluate patients with chronic illnesses, such as COPD. Thus, in order for this study to lead onto any patient specific studies in the future, a health status measurement which is valid for healthy participants and patients with a chronic illness must be used. The measurement tool must therefore be a general purpose health questionnaire that is not illness specific but must be valid in the context of accessing patients with chronic illnesses [3]. The SF-36 questionnaire meets this criteria. SF-36 is a non-illness specific health status measure which has been validated in a general adult population [16] and in a chronic illness patient population [17], [18]. The SF-36 was therefore chosen as the measurement tool for this study.

The SF-36 is a general health instrument that measures eight health related concepts: physical functioning (PF-10 items), role limitations due to physical problems (RP-4 items), bodily pain (BP-2 items), general health perceptions (GH-5 items), vitality (VT-4 items), social functioning (SF-2 items), role limitations due to emotional problems (RE-3 items), and perceived mental health (MH-5 items). Each question has multiple choice answers, with each answer having a predefined numerical score between 0-100. Answers relating to positive health contribute to a higher score, while answers relating to negative health contribute to a lower score. Each of the eight component scores are then computed using an average of specific question scores related to that component. Z-scores are then computed for each of the eight component scores and combined using weighted averages to compute two summary component measures: the Physical (PCS) and Mental (MCS) Component Summary Scores [19]. Both summary scores, PCS and MCS, are computed such that the mean and standard deviation, of a set of scores in a population, are 50 and 10 respectively. A questionnaire UI screen was integrated into the App to allow users to answer the SF-36 questions via radio buttons (See Figure 1(Right)).

1) Participant Health Statistics: After downloading and launching the App for the first time, participants are shown a participant consent screen where details about the study, and data collected during the study, are explained. Participants are then given the choice to consent via a button labelled "I Consent" or to reject via a button labelled "Do not participate". Ethical approval for this study was granted by Ulster University Ethics committee and the contents of the participant consent screen were reviewed by the Ethics Committee.

The App was downloaded by a total of 1751 users, of which 760 completed the SF-36 questionnaire. Table I details the mean and standard deviation SF-36 self-ratings, for the 8 different concepts and the 2 summary measures, of participants based on categories of gender, age and country. It can be seen, for example, that PF is generally higher in younger participants. Conversely, MH is generally higher for older participants.

B. Data Processing

In this section methods used to process smartphone recorded motion data are described. Two processing stages are implemented. However, prior to describing the two processing stages, we describe some preliminary investigations aimed at improving our understanding of the data.

1) Stage 0 (Preliminary Data Analysis): It was initially postulated that features relating to activity duration could potentially be used as a health status indicator. In order to investigate this we investigated two duration based measures: 1) Total Movement Duration (TMD) and 2) Average Stationary Period (ASP). TMD specifies the total amount of time in which the phone was detected as moving during a given day. The phone was deemed to be moving if the variance of the accelerometer magnitude was greater than a predefined threshold. For each n second window where the phone was deemed to be moving, n seconds were added to the overall TMD measure for that day. ASP was calculated as the average of a set of stationary period durations for a given day. The set of stationary period durations store the set of times between when the phone stopped moving and when the phone started to move again (i.e. the amount of time the phone was stationary). ASP therefore stores the average period of time a participants phone was stationary for during a given day. During calculation of ASP, stationary periods longer than 4 hours were discarded in order to discount sleep time and times when the participant placed the phone on a flat surface.

	Υ *(TMD)* Mean (SD)	Λ *(ASP)* Mean (SD)
Gender		
Female	1h:27m (4m:28s)	22m (14m)
Male	1h:51m (5m:19s)	21m (14m)
Age		
18-21	1h:34m (5m:13s)	15m (15m)
22-25	1h:48m (3m:58s)	18m (11m)
26-30	1h:56m (5m:58s)	20m (12m)
31-35	1h:38m (3m:48s)	17m (14m)
36-40	1h:32m (5m:41s)	24m (16m)
41-50	1h:35m (4m:28s)	21m (13m)
51-60	1h:36m (5m:12s)	22m (12m)
60+	1h:22m (4m:4s)	29m (12m)
Overall	1h:37m (4m:55s)	21m (14m)

TABLE II
DURATION BASED MEASURES BY GENDER AND AGE

Table II shows the overall mean and standard deviation of the 2 duration based measures for different age groups and genders. It can be seen that, on average, a participants' phone moved for a total of 1 hour and 37 minutes per day. Additionally, it can be seen that, on average, a participants' phone stayed stationary for an average period of 21 minutes. For the remainder of this work, we denote TMD and ASP as Υ and Λ respectively.

A qualitative analysis of movement data was performed in order to investigate potential links between movement and health status. Figure 2 shows movement duration data (TMD), and individual SF-36 self-ratings, for two female participants (both aged 40-50). A potential link between health status and movement can be seen. Participant A has low SF-36 self-ratings and relatively little movement, while Participant B has high SF-36 self-ratings and significant and regular movement between 10am and 11pm.

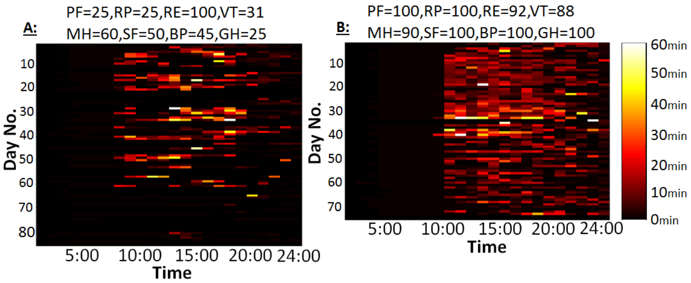


Fig. 2. Sample Movement Durations, Υ, for each day and hour, for 2 participants. Blank (black) areas of the graph denote no motion recorded for that hour. This can be due to the sensor being turned off, or because the phone remained stationary for the entire hour

Further to the qualitative analysis above, we performed a quantitative evaluation to further investigate potential links between SF-36 self-ratings and movement durations. Correlation between each of the SF-36 self-ratings and the two duration measure Υ and Λ were calculated. Results showed an average correlation of 0.16 and 0.03 between the 8 SF-36 scales and Υ and Λ respectively. The largest correlation between SF-36 and Υ was for the PCS component, with $r = 0.201$. The largest correlation between SF-36 and Λ was for the BP component, with $r = 0.042$. The Λ measures showed no statistically significant correlation with the health status measures. However, results did show that while correlation

		PCS	**MCS**	**PF**	**RP**	**BP**	**GH**	**VT**	**SF**	**RE**	**MH**
	N	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Gender											
Female	446	48.9 (9.0)	48.5 (9.4)	72.5 (26.0)	76.2 (29.9)	65.9 (28.1)	52.5 (22.7)	45.1 (20.9)	62.6 (28.7)	61.0 (32.9)	54.6 (23.2)
Male	312	51.4 (7.8)	52.0 (7.9)	76.5 (27.1)	79.6 (29.6)	73.3 (23.2)	58.8 (21.1)	53.0 (18.6)	68.5 (27.8)	68.6 (30.7)	62.5 (20.9)
Age											
18-21	132	50.4 (8.0)	49.6 (8.8)	76.4 (25.7)	76.5 (30.3)	72.9 (25.3)	56.0 (20.3)	48.4 (19.2)	64.0 (28.6)	62.4 (33.3)	56.3 (22.7)
22-25	108	51.0 (7.7)	49.1 (8.9)	79.7 (22.3)	83.4 (23.6)	70.8 (23.8)	54.4 (22.5)	46.5 (19.6)	65.1 (27.7)	61.3 (31.4)	54.5 (22.6)
26-30	97	50.2 (8.0)	49.5 (9.8)	77.0 (24.2)	78.3 (30.2)	72.5 (24.6)	53.9 (21.4)	46.6 (21.4)	64.4 (27.8)	62.2 (31.5)	57.9 (24.5)
31-35	89	49.1 (8.1)	48.9 (8.5)	71.5 (26.4)	74.9 (27.8)	70.0 (24.4)	52.6 (22.3)	48.1 (21.2)	59.6 (28.1)	61.3 (33.1)	54.8 (20.9)
36-40	79	48.6 (8.7)	49.2 (8.6)	68.9 (27.4)	73.9 (31.0)	67.9 (28.7)	53.0 (22.7)	48.1 (19.1)	61.4 (29.3)	57.5 (32.1)	57.8 (22.3)
41-50	124	48.6 (10.0)	49.6 (8.7)	71.1 (31.8)	75.3 (35.8)	63.8 (29.0)	51.9 (24.0)	46.1 (20.0)	63.3 (29.5)	67.8 (30.6)	58.6 (21.7)
51-60	73	50.7 (9.1)	52.8 (9.8)	71.0 (26.2)	78.3 (29.2)	65.5 (29.2)	61.4 (23.2)	52.9 (21.8)	74.3 (28.4)	73.0 (32.5)	64.2 (22.7)
60+	52	51.3 (8.3)	53.0 (7.5)	74.6 (23.6)	81.3 (24.3)	64.8 (24.8)	61.6 (19.1)	55.2 (19.4)	72.9 (24.1)	72.1 (29.1)	64.3 (21.2)
Country											
UK	227	50.1 (9.2)	49.5 (9.4)	76.6 (28.1)	81.1 (30.0)	68.2 (28.2)	52.9 (21.9)	46.6 (20.6)	64.5 (29.9)	64.8 (33.7)	57.1 (23.6)
USA	139	49.8 (8.8)	50.9 (9.6)	72.4 (27.2)	76.2 (29.9)	66.0 (24.9)	58.0 (23.5)	48.6 (22.6)	68.6 (29.5)	68.7 (32.5)	60.1 (24.2)
Ireland	88	51.1 (8.7)	50.4 (9.1)	77.4 (24.7)	80.0 (28.7)	73.1 (28.4)	57.3 (22.9)	49.2 (20.8)	65.8 (29.1)	67.2 (31.8)	57.7 (23.9)
Canada	82	50.1 (7.2)	49.3 (8.1)	77.8 (22.1)	80.7 (28.2)	70.4 (22.7)	51.6 (20.0)	45.1 (19.6)	63.4 (27.5)	63.8 (30.9)	57.8 (20.2)
Spain	30	52.8 (7.0)	53.3 (8.0)	78.6 (20.6)	81.1 (25.5)	78.2 (27.3)	62.7 (16.3)	55.7 (18.6)	70.8 (28.2)	71.1 (31.6)	64.7 (20.0)
Australia	55	48.5 (8.5)	48.7 (7.8)	69.0 (26.8)	73.5 (29.4)	65.1 (24.5)	52.5 (22.7)	48.4 (18.0)	64.1 (22.7)	55.0 (29.6)	55.3 (19.5)
New Zealand	39	50.3 (6.8)	49.1 (7.8)	76.9 (23.1)	80.5 (27.3)	69.0 (24.7)	56.0 (19.9)	46.3 (16.5)	61.6 (30.1)	63.3 (28.3)	55.9 (19.7)
Other	100	48.4 (8.4)	49.8 (8.4)	65.8 (27.0)	67.3 (31.1)	69.7 (25.6)	55.1 (24.1)	52.3 (18.5)	61.5 (25.4)	57.6 (30.5)	56.8 (21.4)
Overall	760	49.9 (8.6)	49.9 (9.0)	74.2 (26.5)	77.6 (29.8)	69.0 (26.4)	55.0 (22.4)	48.3 (20.3)	65.0 (28.5)	64.2 (32.2)	57.9 (22.6)
Burholt et al. [20]	*13917*	*N/A*	*N/A*	*77.8 (30.0)*	*78.3 (32.3)*	*70.1 (28.9)*	*66.2 (24.0)*	*57.3 (22.3)*	*80.2 (28.1)*	*87.0 (26.0)*	*74.0 (19.9)*

TABLE I
SF-36 SELF-RATINGS FOR PARTICIPANT DEMOGRAPHICS (GENDER, AGE, COUNTRY).

between Υ and different SF-36 components were not strong, the correlations were statistically significant.

Due to only small correlations between movement duration and health status, we conclude that movement duration alone cannot be utilized to consistently infer health status. We postulate that this is due to the real world and inherent uncontrolled nature of this study, where participants use the sensing modality without researcher supervision. It is possible that periods of inactivity relate to periods where the phone was simply not being worn/carried by the participant. During these periods, the sensor would infer that the person was being sedentary when it is possible that the person was in fact being active. We refer to these as "periods of unknown".

The preliminary investigates therefore informed our overall approach for extracting meaningful health related information from motion data. In particular, "periods of unknown" must be accounted for in our approach. An "unknown" occurs when no movement is recorded from the sensor. During this period, it cannot be determined whether (a) the participant is wearing the phone and being sedentary or (b) not wearing the phone (and being active or sedentary). In this work, we address this problem by dealing only with periods of time when we can be almost certain that the participant is wearing the phone. These periods relate to periods when movement is occurring and the screen is off.

Our approach must consider the following 2 points: 1) Utilize data during periods of activity, while removing unknowns by discarding data during stationary periods. 2) Include features which are not based on quantifying the duration of activity, as it is possible that a participant is active during discarded periods. Based on these 2 points, features will therefore describe the type of movement a person performs and will discard movement information during periods of inactivity.

2) Stage 1 (Smart-phone processing): As previously mentioned, a two-stage procedure was implemented in order to process motion data. This section describes the first stage. It was not feasible to upload all raw motion data due to participant network constraints and server storage constraints. Stage 1 of data processing was therefore performed on the Smartphone in order to reduce the quantity of data uploaded to the server. At the end of each day, raw motion data was automatically converted into hourly summary measures and uploaded to a central server. Stage 2 of data processing was conducted on the server and computations were performed to generate an overall behaviour descriptor for each participant using hourly summary measures from Stage 1.

a) Signal Processing: In order to measure a participants' motion, 3-axis Accelerometer data A_x, A_y, A_z and 3-axis Gyroscope data G_x, G_y, G_z, were recorded during periods when the was App enabled. The Madgwick Attitude and Heading Reference System (AHRS), where beta $= 0.2$, was used to combine accelerometer and gyroscope data to calculate the orientation quaternion $Q_{\theta\phi}$ representing the pitch (θ) and roll (ϕ) of the phone [21]. Yaw was not found to be of relevance due to the unconstrained sensor placement. Overall magnitude of the acceleration is defined as $A^m = \sqrt{A_x^2 + A_y^2 + A_z^2}$ and the overall magnitude of the angular velocity is defined as $G^m = \sqrt{G_x^2 + G_y^2 + G_z^2}$.

Data collection was performed with real world uncontrolled conditions. It was therefore highly probable that participants used the sensing device in many different ways, including placing the phone in many different positions and orientations on their body. Due to the unconstrained sensor orientation, useful information such as movement in a particular direction could be lost. In order to overcome this problem, while still retaining information relating to directional movement, a set of orientation independent features were used. We utilized the technique described by Kelly et al. [22] to compute orientation independent features by using a global reference frame to measure acceleration and rotation with respect to gravity. A rotation matrix $R_{\theta\phi}$ was computed from the orientation quaternion $Q_{\theta\phi}$ and the global acceleration frame, calculated using a matrix transformation, defined as $\bar{A} = A \times R_{\theta\phi}$, where $A = \{A_x, A_y, A_z\}$. The acceleration vector $\bar{A}$ represents acceleration relative to gravity. Due to the unconstrained orientation of the phone, there is no way of determining the mediolateral and dorsoventral axis. We therefore combine horizontal axis x and z into a single horizontal measure. Acceleration on the horizontal plane, A^h, is therefore defined as $A^h = \sqrt{\bar{A}_x^2 + \bar{A}_z^2}$ while vertical acceleration, A^v, is defined as $A^v = \bar{A}_y$. Similarly, the global gyroscope frame is defined as $\bar{G} = G \times R_{\theta\phi}$, where $G = \{G_x, G_y, G_z\}$. Rotation around the vertical plane, G^v, and horizontal rotation, G^h, are defined as $G^v = \bar{G}_y$ and $G^h = \sqrt{\bar{G}_x^2 + \bar{G}_z^2}$ respectively.

In order to describe behaviour at a given time t, a number of different features were calculated from 2 seconds windows of the accelerometer signals A^m, A^v, A^h, the gyroscope signals G^m, G^v, G^h and the orientation angles θ and ϕ. A sliding window system was used to calculate features for each 2 second window. A set of features were calculated utilizing the aforementioned signals above and the following feature processing techniques. Feature processing techniques were based on methods described by Kelly et al. [22] for a smartphone activity recognition system.

- $Min(x)$: Min value of signal x.
- $Max(x)$: Max value of signal x.
- $\mu(x)$: Mean of signal x.
- $\sigma(x)$: Variance of signal x - variance is higher for more dynamic activities.
- $s(x)$: Skewness of signal x.
- $k(x)$: Kurtosis of signal x.
- $IQR(x)$: Refers to the Interquartile range of signal x. IQR of Gyro can be important for identifying sit to stand activity.
- $ROC(x)$: Refers to the rate of change of a signal x.
- $FFT(x)$: Frequency domain analysis of signal x, calculating frequency with the greatest amplitude.
- $MSV(x)$: Refers to the most significant velocity of the signal. This is computed by identifying zero crossings from the mean subtracted signal. The velocity is then defined as the maximum rate of change at zero crossing points.

Using the 8 motion signals and the above 10 measurements, a set of 80 features were computed. Additionally, two correlation based features were utilized in order measure interactions between vertical and horizontal motion signals:

- $Corr(A^v, A^h)$: Refers to the correlation of vertical acceleration and horizontal acceleration.
- $Corr(G^v, G^h)$: Refers to the correlation of vertical rotation velocity and horizontal rotation velocity.

In total, a set of $(8 \times 10) + (2) = 82$ features were computed to create an overall feature vector $f(t)$.

b) Feature Summary: In order to generate a feature summary, only features, $f(t)$, which had a corresponding accelerometer variance, $\sigma(A_t^m)$, greater than a pre-set threshold, $T = 0.35$, were used in the generation of a behaviour profile vector. Additionally, features which were recorded during periods when the participant was interacting with the phone were discarded. For a specific day, d, feature vectors $f(t)$ were calculated using a 2 second sliding window. Therefore, for each hour, a maximum of 1800 feature vectors would be calculated if the participant was active for the entire hour and did not interact with the phone. For each hour, h, all feature vectors, which have an acceleration variance greater than the threshold, were averaged to compute a single feature vector F_{dh} which described the overall behaviour profile of a participant for hour h on day d. If there was no motion for an entire hour, then no behaviour profile was generated for that hour. Duration measures, Υ and Λ, were then normalized in order to represent duration measures as a fraction of an hour, where $\tilde{\Upsilon} = \frac{\Upsilon}{1Hour}$ and $\tilde{\Lambda} = \frac{\Lambda}{1Hour}$ such that $0 \leq \tilde{\Upsilon} \leq 1$ and $0 \leq \tilde{\Lambda} \leq 1$. The hourly behaviour profile, F_{dh}, was then augmented with the normalized duration measures such that $\overline{F_{dh}} = \{F_{dh}, \tilde{\Upsilon}_{dh}, \tilde{\Lambda}_{dh}\}$. Where $\tilde{\Upsilon}_{dh}$ and $\tilde{\Lambda}_{dh}$ represent normalized TMD and ASP measures respectively, for day d and hour h. The overall hourly behaviour profile, $\overline{F_{dh}}$ was therefore comprised of $(82 + 2) = 84$ features. The normalized duration measure $\tilde{\Upsilon}_{dh}$ was also used to determine the weighting of each hour block in generating an overall behaviour profile. For example, if there were two hourly behaviour profiles, one which comprised 3 minutes of movement and the other which comprised 53 minutes of movement, the behaviour profile which comprised 53 minutes should be given more influence in the generation of an overall behaviour profile. For each day d, the entire set of behaviour profiles, $F = \{\overline{F_{d0}}, \overline{F_{d1}}, ..., \overline{F_{d23}}\}$, was uploaded to the server along with the set of duration weights $W = \{\tilde{\Upsilon}_{d0}, \tilde{\Upsilon}_{d1}, ..., \tilde{\Upsilon}_{d23}\}$.

3) Stage 2 (Server processing): The second data processing stage was performed on the central server. A database on the server stores hourly behaviour profiles F, and movement duration weights W, as uploaded by each participant's smartphone App (described in Section II-B2). The aim of stage 2 processing was to generate a single overall behaviour profile which described the average behaviour for each participant.

The first server processing step grouped hourly behaviour profiles, for each participant p, into hourly bins, F_h^p. Where $F_h^p = \{F_{0h}^p, F_{1h}^p, ..., F_{Dh}^p\}$, and F_h^p represents all behaviour profiles for participant p for a specific hour h for all days from day 0 to day D and D was the total number of days recorded. A time specific average behaviour profile, which represents the average behaviour of a participant during a specific hour h for all days, was computed using a weighted average as defined in Equation 1.

$$\overline{F}_h^p = \sum_{i=0}^{D} \left(F_{ih}^p \times \frac{\tilde{\Upsilon}_{ih}^p}{W_h^p} \right) \tag{1}$$

$$W_h^p = \sum_{i=0}^{D} \tilde{\Upsilon}_{ih}^p \tag{2}$$

$$\Psi_\alpha^p = \sum_{i=0}^{23} (F_i^p \times \omega_i^p) \tag{3}$$

$$\Psi_\sigma^p = \sum_{i=0}^{23} \left((\Psi_\alpha^p - F_i^p)^2 \times \omega_i^p \right) \tag{4}$$

$$\Psi_\delta^p = \sum_{i=1}^{23} \left((F_i^p - F_{i-1}^p) \times \omega_i^p \right) \tag{5}$$

$$\omega_i^p = \frac{\sum_{j=0}^{D} \tilde{\Upsilon}_{ji}^p}{\sum_{k=0}^{23} \sum_{j=0}^{D} \tilde{\Upsilon}_{jk}^p} \tag{6}$$

The average hourly behaviour profile, $\overline{F}^p_h$, represents the average behaviour of a participant at a certain time of the day. For example, $\overline{F}^p_{10}$ represents the average behaviour of a participant between 10am and 11am for all the days a participant had the App enabled. The sequence of average hourly behaviour profiles, $\overline{F}^p = \{\overline{F}^p_0, ..., \overline{F}^p_{23}\}$, represents all hourly behaviour profiles over the course of an average day for participant p. The hourly sequence of behaviour profiles, $\overline{F}^p$ was used to calculate the overall behaviour profile, Ψ^p for participant p. Equations 3 - 6 detail the different components used to create the overall behaviour profile, which we define as $\Psi^p = \{\Psi^p_\alpha, \Psi^p_\sigma, \Psi^p_\delta\}$.

The first behaviour profile component, Ψ^p_α, defines the overall weighted average behaviour profile over the course of the average day. The weighting factor ω^p_i is defined as the movement duration ratio of movement occurring during hour i versus the total movement for all hours and days for participant p. The overall average behaviour profile, Ψ^p_α, therefore combines all hourly behaviour profiles into a single behaviour profile, where hours containing longer movement durations have more influence on the final average behaviour profile when compared to hours which contained short movement durations. The second behaviour profile component, Ψ^p_σ, defines the overall weighted variance of the hourly behaviour profiles over the course of the average day. This measure takes into account the overall variability of individual hourly behaviour profiles over the course of an average day. Finally, the third behaviour profile component, Ψ^p_δ, defines the weighted rate of change of the hourly behaviour profiles over the course of the average day. This measure takes temporal patterns into account and evaluates how the behaviour of a participant changes from morning to night. As discussed in Section II-B2b, an hourly behaviour profile is comprised of a total of 84 different features. The overall behaviour profile, generated from the three feature vector components, therefore contains $(84 \times 3) = 252$ features. While it is possible that different smart-phone models will have sensors with varying levels of accuracy, potential issues relating to accuracy are filtered out during the feature summary process.

C. Feature Selection

Feature selection is performed in order to reduce the number of features used, enhance the generalization of models and reduce the chances of over fitting during training [23]. A Steepest Ascent Hill-Climbing based feature selection method was implemented in this work. In order to denote chosen features, and discarded features, a binary feature selection mask $B = \{b_0, ..., b_{216}\}$ is defined. Where $B[n] = 1$ denotes that a feature is selected and $B[n] = 0$ denotes that a feature is discarded. An initial solution is chosen such that $B[i] = 0$ for all i. A fitness function $f(x, y, B)$ is then used to evaluate the fitness of B, where x and y are the training and test data sets respectively.

At each iteration, a candidate mask B^i is generated for all i, where $0 \leq i \leq 252$. Each candidate mask B^i is a copy of mask B, with the exception that bit i is inverted. In other words, each candidate mask represents a different individual feature, i, being added to the overall set of selected features. For each candidate mask, B^i, the fitness $f(x, y, B^i)$ is computed. At the end of the current iteration, if $f(x, y, \overline{B}_{max}) > f(x, y, B)$ then the feature selection mask is updated such that $B = \overline{B}_{max}$. Where $\overline{B}_{max}$ represents the candidate mask with the maximum fitness as defined in Equation 7. If $f(x, y, \overline{B}_{max}) < f(x, y, B)$ then the termination criteria has been reached and the selection process will stop. At this point, the bits which are enabled in B represent the selected features.

$$\overline{B}_{max} = \underset{B^i}{\arg\max} f(x, y, B^i) \tag{7}$$

During preliminary experiments, a number of different regression modelling techniques were evaluated in order to determine the best technique to carry out detailed experiments on. Results from preliminary results showed that Support Vector Machine (SVM) regression [24], using a Radial Basis Function (RBF) kernel, performed best. We therefore utilize SVMs, using RBF, for the core experiments of this work.

SVMs will be used to build the regression models such that health status predictions can be made from behaviour profiles Ψ. In order for feature selection to find features that best suit the learning algorithm, a wrapper subset feature selection is performed [25]. The fitness function $f(x, y, B)$ is therefore based on an SVM regression model, trained on training set x and tested on test set y using features defined in B. The fitness function measures the Pearson Correlation between predicted health status $\delta(M^B_x, y^\Psi_i)$ and the actual health status $y^{'}_i$. Where δ is the prediction function, M^B_x is the regression model trained using feature vectors x and feature mask B, and y^Ψ_i is the i^{th} behaviour profile in the test set.

D. Repeated Double Cross Validation

In order to avoid feature selection bias in the regression performance evaluation, we implemented a repeated (M = 10) double k-fold (k = 5) cross-validation structure modelled after that of Filzmoser et al. [26], and also utilized in a recent study by Reynolds et al. [27]. Three nested loops where implemented: 1) a repetition loop, 2) an outer cross-validation loop, and 3) an inner cross-validation loop which is contained within a Hill-Climbing Wrapper Subset feature selection procedure. Since each participant has only a single feature describing their behaviour, the cross validation for this study is inherently subject-independent. Figure 3 gives a visual overview of the cross validation method used. The outermost loop (the repetition loop) is used to repeat the double k-fold cross validation 10 times to assess the variability associated with the particular data segmentation. Within each iteration of the repetition loop, the data are randomly split into five random segments. For each iteration of the outer cross validation loop, one segment is set aside as the test group. The other four segments are considered the calibration set. The calibration set is sent to the inner cross validation loop and repartitioned into 5 segments. Each iteration uses 4 of the calibration segments to perform feature selection and uses the remaining test segment to compute the fitness. A set of 5 individual feature masks, B_k, are generated for each iteration k. The individual masks are then combined to create a single feature mask, $\overline{B}$, by selecting the most common features from all individual masks. A feature is deemed common, if it is enabled in at least 60% of the individual feature masks.

E. Evaluation Metrics

A number of experiments were conducted to evaluate our overall hypothesis and test if, and to what extent, motion data can be used to predict health status. Specifically, experiments were conducted to test the ability of our proposed behaviour profiles to predict the 10 different SF-36 self-ratings using behaviour profiles Ψ. Experiments used the repeated double cross validation protocol, described in Section II-D, to perform feature selection, training, testing and ultimately calculate overall evaluation metrics. Three different evaluation metrics are utilized. Firstly, Pearson correlation (ρ) is used to measure the linear correlation between predicted health status and ground truth health status. Secondly, Mean Absolute Error (MAE) is used to calculate the average absolute difference between predicted health status and ground truth health status. Finally, Relative Absolute Error (RAE) computes the MAE as a percentage of the standard deviation of the health status measure.

III. Results

In this Section, we discuss the results of experiments. As discussed in Sections II-A and II-B, an Android App was utilized to collect health status and behaviour profile data from a set of participants from the general population. In total, the App was downloaded 1751 times and an average of 114 hours of data was uploaded by each participant. Of the 760 participants completing the SF-36 questionnaire, a total of 371, 249 and 196 of these uploaded at least 1, 24 and 48 hours of motion data respectively.

The proposed behaviour measures rely upon constructing an average behaviour profile over a number of days. In order for the average behaviour profile to accurately reflect the actual average behaviour of a participant, the measure must be generated from data recorded over large enough period of time such that generalizations can be made about behaviour. We first discuss an experiment aimed at discovering the minimum number of hours required to generate an average behaviour profile. An hour threshold is utilized in order to specify the minimum number of hours of data a participant must provide in order to be included in the experiment. For this experiment, we evaluate 4 possible hour thresholds (24, 48, 72 and 96). For each hour threshold, the repeated double cross validation protocol was used to compute performance metrics for each of the 8 SF-36 concepts and the 2 component scores. Different hour thresholds produce different size training sets. For example, an hour threshold of 24 hours will result in more participants being utilized when compared to an hour threshold of 96. In order to remove any bias in results caused by training set sizes, each of the training sets, for each of the 4 hour thresholds, are sub-sampled such that all training sets have the same number of participants. The 96 hour threshold produces the smallest training set (N=148). Therefore, participants were removed from other training sets such that each training set comprised 148 participants. Random participants are removed at each iteration of the repetition loop during the execution of the double cross validation protocol. Priority was given for participants to remain in a sub-sampled training set if the number of hours for that participant is close to the hour threshold currently being tested.

	N=148		
Threshold	ρ	MAE	RAE
> 24 Hours	0.568	13.73	61.1%
> 48 Hours	0.639	13.49	60.3%
> 72 Hours	0.696	11.62	51.0%
> 96 Hours	0.692	11.81	51.7%

TABLE III
Hour Thresholds - Prediction Metrics

	72 Hour Minimum *N=171*		
Measure	ρ *(SD)*	**MAE** *(SD)*	**RAE** *(SD)*
PCS	0.707 *(0.042)*	4.99 *(0.33)*	49.9% *(3.3%)*
MCS	0.690 *(0.059)*	5.89 *(0.74)*	58.9% *(7.4%)*
PF	0.632 *(0.013)*	13.8 *(0.62)*	52.0% *(2.3%)*
RP	0.658 *(0.053)*	14.6 *(1.19)*	48.9% *(3.9%)*
RE	0.670 *(0.018)*	16.6 *(0.81)*	51.5% *(2.5%)*
VT	0.645 *(0.048)*	12.1 *(1.31)*	59.6% *(6.4%)*
MH	0.705 *(0.022)*	11.2 *(0.51)*	49.5% *(2.2%)*
SF	0.705 *(0.045)*	13.6 *(1.12)*	47.7% *(3.9%)*
BP	0.697 *(0.009)*	13.8 *(0.35)*	52.2% *(1.3%)*
GH	0.752 *(0.021)*	10.6 *(0.64)*	47.3% *(2.8%)*
Average	0.686 *(0.033)*	11.7 *(0.76)*	51.9% *(3.6%)*

TABLE IV
Regression Prediction Evaluation Metrics

Table III details the average valuation metrics, for the 8 individual SF-36 self-ratings, achieved by training the system on the 4 different hour thresholds. Results show that 72 and 96 hour thresholds perform best with correlations of 0.696 and 0.692 respectively. While both these thresholds appear to perform well, we utilize a 72 hour threshold for the remaining experiments in this work. This is due to the fact that the 72 hour threshold will produce a bigger dataset, compared to 96 hour threshold, to carry out additional evaluations.

We now describe an in depth experiment, evaluating the ability of the proposed behaviour profiles to predict individual SF-36 self-ratings. Experiments utilize an hour threshold of 72 hours, producing a dataset comprising 171 participants. Behaviour profiles were computed for each participant and the repeated double cross validation protocol was used to compute performance metrics for each of the 8 SF-36 concepts and the 2 component scores. Table IV details the individual performance metrics for each of the SF-36 self-ratings. Individual performance metrics shown in Table IV represent the average, and standard deviation, evaluation metrics obtained from the 10 repetition loops during the repeated double cross validation protocol.

Results show that, on average, the proposed behaviour profile can be utilized to predict health status with an average MAE of 11.7. Component scores, PCS and MSC, were predicted with a correlation of 0.707 and 0.69 respectively. Figure 4 plots the PCS predictions for 1 of the repetition loops in the repeated double cross validation protocol. The most accurate predictions were made based on the GH scale, with a correlation and MAE of 0.752 and 10.6 respectively. Figure

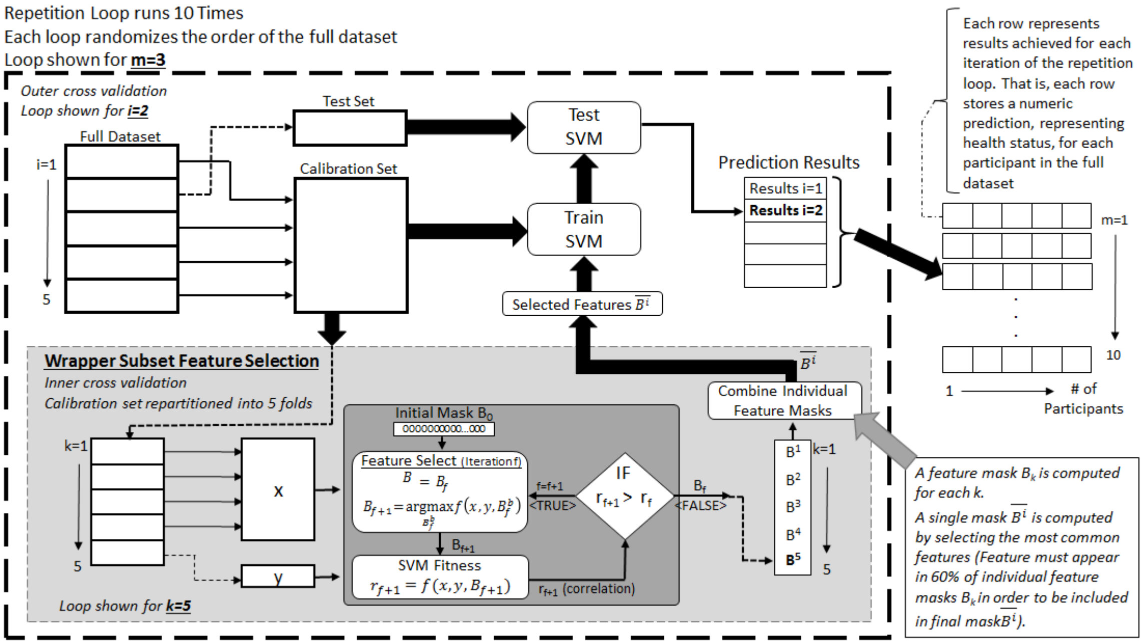

Repetition Loop runs 10 Times
Each loop randomizes the order of the full dataset
Loop shown for m=3
Outer cross validation
Loop shown for i=2
Full Dataset
i=1
5
Test Set
Calibration Set
Test SVM
Train SVM
Prediction Results
Results i=1
Results i=2
Selected Features $\overline{B^i}$
$\overline{B^i}$
Wrapper Subset Feature Selection
Inner cross validation
Calibration set repartitioned into 5 folds
k=1
5
x
y
Loop shown for k=5
Initial Mask B_0
0000000000...000
Feature Select (Iteration f)
$B = B_f$
$B_{f+1} = \underset{B_f^b}{\text{argmax}} f(x, y, B_f^b)$
B_{f+1}
SVM Fitness
$r_{f+1} = f(x, y, B_{f+1})$
r_{f+1} (correlation)
f=f+1
<TRUE>
IF $r_{f+1} > r_f$
B_f
<FALSE>
Combine Individual Feature Masks
B^1
B^2
B^3
B^4
B^5
k=1
5
Each row represents results achieved for each iteration of the repetition loop. That is, each row stores a numeric prediction, representing health status, for each participant in the full dataset
m=1
10
1
# of Participants
A feature mask B_k is computed for each k.
A single mask $\overline{B^i}$ is computed by selecting the most common features (Feature must appear in 60% of individual feature masks B_k in order to be included in final mask $\overline{B^i}$).

Fig. 3. Feature Selection, Training and Testing Protocol Overview

5 plots the GH predictions for 1 of the repetition loops in the repeated double cross validation protocol. Predictions for the PF and VT scores produced the lowest evaluation metrics, with correlations of 0.632 and 0.645 respectively. Figure 6 plots the VT predictions for 1 of the repetition loops in the repeated double cross validation protocol.

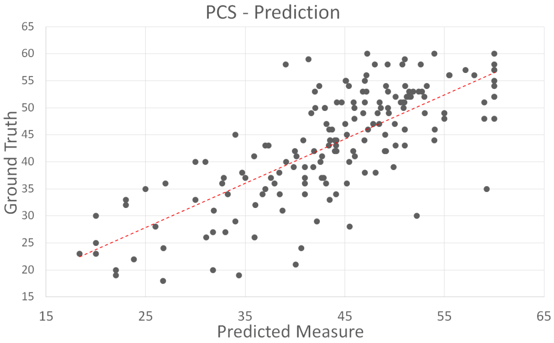


Fig. 4. Sample PCS Predictions: Correlation (ρ) = 0.726, RAE = 47.8%.

Results indicate that the proposed behaviour profiles can be used to predict SF-36 self-ratings with an average correlation and MAE and 0.686 and 11.7 respectively.

A. Feature Selection

As described in Section II-C, a feature selection process is performed in order to improve prediction performance. In this section we give details relating to features which were selected, for the different SF-36 self-ratings, by the hill-climbing wrapper subset feature selection process during the repeated double cross validation protocol. Table V details the most common features which were selected for different SF-36 self-ratings. It can be seen that correlation between horizontal and vertical acceleration is an important feature in

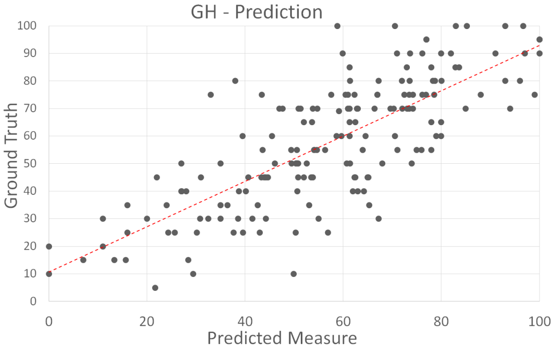


Fig. 5. Sample GH Predictions: Correlation (ρ) = 0.747, RAE = 46.9%.

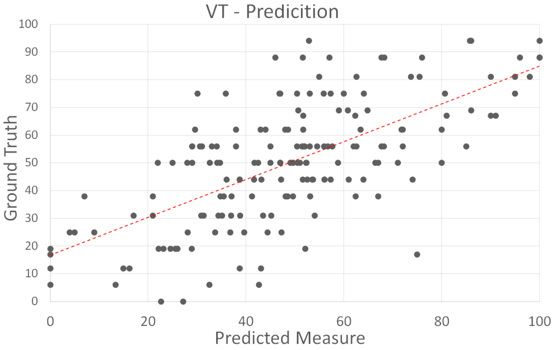


Fig. 6. Sample VT Predictions: Correlation (ρ) = 0.647, RAE = 59.5%.

SF-36	**x**	**F**	Ψ	**SF-36**	**x**	**F**	Ψ
PCS, MCS, BP, GH, PF, RE, RP, SF	A^v, A^h	COR	δ	PCS, MCS, BP, VT, GH, MH	A^m	Min	σ
MCS, MH, PF, RE, RP	A^m	K	α	PCS, MCS, BP, VT, PF	A^v	MSV	α
PCS, MH, PF, RP, SF	A^v	MSV	δ	MCS, BP, RE, SF	A^v	μ	α
PCS, BP, PF, RE	A^m	μ	σ	PCS, VT, GH, PF	A^v, A^h	COR	σ
BP, PF, RP, SF	A^v	μ	δ	GH, RE, SF	A^v, A^h	COR	α
PCS, MCS, SF	G^v	IQR	α	MCS, BP, MH	A^m	Min	σ
PCS, BP, VT	A^v	MSV	σ	PCS, BP, GH	θ	ROC	α
PCS, MCS, GH	Λ	-	σ	PF, RP, SF	A^m	Min	δ
PCS, BP, PF	A^m	FFT	σ	PCS, MCS, RE	Λ	-	δ
MCS, BP, SF	Υ	-	σ	MCS, MH, RE	A^m	FFT	δ

x = Motion Signal
F = Signal Processing Method
Ψ = Component Processing Method, as detailed in Equations 3 - 6

TABLE V
FEATURES SELECTED FOR DIFFERENT SF-36 SELF-RATINGS

predicting health status, and was selected to predict 8 of the 10 SF-36 self-ratings (PCS, MCS, BP, GH, PF, RE, RP, SF). Additionally, the vertical acceleration signal, A^v, was the basis for a number of features.

IV. DISCUSSION

No matter how health status is measured, it is important that the results are interpretable. In the context of a clinical trial, for example, a treatment group might show a 5 point SF-36 difference when compared to a control group. However, it is important to have a benchmark to evaluate whether this difference actually matters. This benchmark is referred to as Clinically Important Difference (CID) or Minimal Clinically Important Difference (MCID). Research has shown, based on a systematic review of 38 studies using different HR-QOL instruments, that the MCID was consistently close to half a standard deviation of the health status measure [28][29]. Half a standard deviation equates to approximately 5 points for the SF-36 component scores (PCS and MCS) and approximately 10-16 points for individual SF-36 concepts. This has been backed up in the literature, where approximately 10, 20 and 30 points have been suggested to represent a small, moderate and large CID respectively for COPD patients for the 8 individual SF-36 self-ratings [30].

Some SF-36 self-ratings performed better than others. Predictions for PCS, RP, MH, SF and GH achieved error rates under 50% standard deviation. Two results of particular interest were GH and PF, with contrasting correlations of 0.753 and 0.632 respectively. Initially, we would have assumed that PF would perform better due to the assumed direct relationship between the PF and measures extracted from motion sensors (i.e. physical movement). Results contradict this however. Investigating this further, we postulate that PF performs poorer due to the distribution of training data. Specifically, it can be seen in Table I that the averages and standard deviations for PF are skewed such that the majority of the distribution is concentrated on the upper end of the scale. The limited number of participants on the lower end of the PF scale could possibly skew the training of the model and result in higher PF results. This could be a potential limitation of this work where there is relatively low number of participants which measure on the lower end of the SF-36 scale. Additional participant numbers would be required in order to evaluate the performance of regression models trained on non-skewed subsets.

While overall RAE results are not consistently below 50%, as advised in the literature, general results certainly indicate that automatically generated behaviour data can be used to make predictions relating to health status. To our knowledge, this is the first work which investigates the use of sensors for the prediction of health status and the results indicate that this work is an important step towards automatic and objective health status measures.

In terms of specific features that indicate health status, experiments revealed that the vertical acceleration signal, A^v, was significant in making SF-36 predictions. We postulate that vertical acceleration, particularly in conjunction with the MSV processing method, could indicate regular performance of activities with significant vertical motion. Significant vertical motion could relate to activities such stair climbing and sit-to-stand movements. Measures of low vertical motion could indicate impaired ability to perform these activities. While measures of high vertical motion could indicate performance of these activities with good strength and balance and could also indicate the occurrence of more intense activities such as running and jumping. While preliminary results indicated that Λ on its own showed no correlation with health status, it is likely that it complimented a number of other features as results show Λ was the basis for two of the selected features.

A potential limitation of this work relates to the method of SF-36 administration, where questionnaires are self-completed by participants. Research has shown that self-completed SF-36 results are likely to be lower than if scores were obtained through interviewer-administration [31]. In Table I it can be seen that the majority of the average scores are lower when compared to the study conducted in 2007 by Burholt et al. [20]. Another limitation relates to the method of dealing with "periods of unknown". Our approach solves the problem of assigning sedentary features to a participant who is potentially active. However, in doing this, we are discarding some data during sedentary periods which could potentially hold relevant health status information. Future work should consider alternative solutions of discovering when a participant is actually wearing the phone and is sedentary vs. not wearing the phone.

V. CONCLUSION

Research has shown that there is a need for new accurate and objective methods for measuring patient health status. Current state of the art methods of making automatic predictions relating to health status rely on costly and time consuming manual observations about a patient's behaviour. This paper investigates the use of modern smartphones as a means of computing automatically generated behaviour observations. Moreover, investigations of how automatically generated behaviour observations can be used to infer health status are carried out. Methods were proposed to compute a behaviour profile for a participant utilizing motion sensor data from a

Smartphone. Evaluations were then performed using a crowdsourced data set comprising motion sensor data and health status information from 171 participants. Results show that, on average, the 10 different SF-36 self-ratings could be predicted within 51.9% of the standard deviation of the SF-36 self-ratings. This is comparable to the suggested CID of 50%. It can be concluded that, while prediction accuracy should be further improved before use in a clinical setting, this work represents an important first step towards making health status predictions without the need for manual observations. The key innovation of this work is that health status can be measured using unobtrusive, inexpensive and already available hardware. The significance of this is that it could have major benefits for clinicians in treating patient with chronic conditions where health status and daily life function is of relevance to treatment. Without additional cost or infrastructure, it could enable clinicians to accurately and objectively assess the daily life benefits of treatments on an individual patient basis. Moreover, due to the automatic nature of the system, health status could be tracked on an ongoing basis. This could allow clinicians to assess health status trajectory over time without the need for a healthcare professional to record information or without the need for patients to actively participate in the recording of information.

References

[1] M. Viswanathan, C. E. Golin, C. D. Jones, M. Ashok, S. J. Blalock, R. C. M. Wines, E. J. L. Coker-Schwimmer, D. L. Rosen, P. Sista, and K. N. Lohr, "Interventions to improve adherence to self-administered medications for chronic diseases in the United States: A systematic review," pp. 785–795, dec 2012.

[2] UK Department of Health, "Long Term Conditions Compendium of Information: Third Edition," 2012.

[3] J. Curtis and D. Patrick, "The assessment of health status among patients with COPD," *European Respiratory Journal*, vol. 21, no. Supplement 41, pp. 36S–45s, 2003.

[4] J. W. H. Kocks, M. G. Tuinenga, S. M. Uil, J. W. K. van den Berg, E. Ståhl, and T. van der Molen, "Health status measurement in COPD: the minimal clinically important difference of the clinical COPD questionnaire." *Respiratory research*, vol. 7, no. i, p. 62, 2006.

[5] P. W. Jones, "Health status measurement in chronic obstructive pulmonary disease," *Thorax*, vol. 56, no. 11, pp. 880–887, 2001.

[6] B. Gandek, J. E. Ware, N. K. Aaronson, J. Alonso, G. Apolone, J. Bjorner, J. Brazier, M. Bullinger, S. Fukuhara, S. Kaasa, A. Leplège, and M. Sullivan, "Tests of data quality, scaling assumptions, and reliability of the SF- 36 in eleven countries: Results from the IQOLA Project," *Journal of Clinical Epidemiology*, vol. 51, no. 11, pp. 1149–1158, 1998.

[7] S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers, "A review of wearable sensors and systems with application in rehabilitation." *Journal of neuroengineering and rehabilitation*, vol. 9, p. 21, jan 2012.

[8] K. Hung, Y. T. Zhang, and B. Tai, "Wearable medical devices for tele-home healthcare." *Conference of the IEEE Engineering in Medicine and Biology Society.*, vol. 7, pp. 5384–7, jan 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17271560

[9] S. Del Din, A. Godfrey, and L. Rochester, "Validation of an Accelerometer to Quantify a Comprehensive Battery of Gait Characteristics in Healthy Older Adults and Parkinson's Disease: Toward Clinical and at Home Use," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 3, pp. 838–847, may 2016.

[10] B. R. Greene, A. Odonovan, R. Romero-Ortuno, L. Cogan, C. N. Scanaill, and R. A. Kenny, "Quantitative falls risk assessment using the timed up and go test," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 12, pp. 2918–2926, 2010.

[11] D. J. Cook, M. Schmitter-Edgecombe, and P. Dawadi, "Analyzing Activity Behavior and Movement in a Naturalistic Environment Using Smart Home Techniques," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 6, pp. 1882–1892, nov 2015.

[12] J. Juen, Q. Cheng, and B. Schatz, "A Natural Walking Monitor for Pulmonary Patients Using Mobile Phones," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 4, pp. 1399–1405, 2015.

[13] D. Kelly, S. Donnelly, and B. Caulfield, "Smartphone derived movement profiles to detect changes in health status in COPD patients - A preliminary investigation." *Conference of the IEEE Engineering in Medicine and Biology Society.*, vol. 2015, pp. 462–5, 2015.

[14] Y. Yi Yang, A. Hauptmann, M.-Y. Ming-Yu Chen, Y. Yang Cai, A. Bharucha, and H. Wactlar, "Learning to predict health status of geriatric patients from observational data," in *2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB).* IEEE, may 2012, pp. 127–134.

[15] S. Pakhomov, N. Shah, P. Hanson, S. Balasubramaniam, S. A. Smith, and S. A. Smith, "Automatic quality of life prediction using electronic medical records." *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, vol. 2008, pp. 545–9, 2008.

[16] R. Bize, J. A. Johnson, and R. C. Plotnikoff, "Physical activity level and health-related quality of life in the general adult population: A systematic review," *Preventive Medicine*, vol. 45, no. 6, pp. 401–415, 2007.

[17] F. M. Boueri, B. L. Bucher-Bartelson, K. A. Glenn, and B. J. Make, "Quality of life measured with a generic instrument (Short Form-36) improves following pulmonary rehabilitation in patients with COPD." *Chest*, vol. 119, no. 1, pp. 77–84, jan 2001.

[18] E. Ståhl, A. Lindberg, S.-A. Jansson, E. Rönmark, K. Svensson, F. Andersson, C.-G. Löfdahl, and B. Lundbäck, "Health-related quality of life is related to COPD disease severity." *Health and quality of life outcomes*, vol. 3, p. 56, 2005.

[19] S. S. Farivar, W. E. Cunningham, and R. D. Hays, "Correlated physical and mental health summary scores for the SF-36 and SF-12 Health Survey, V.1," *Health and Quality of Life Outcomes*, vol. 5, no. 1, p. 54, 2007.

[20] V. Burholt and P. Nash, "Short Form 36 (SF-36) Health Survey Questionnaire: normative data for Wales." *Journal of public health (Oxford, England)*, vol. 33, no. 4, pp. 587–603, 2011.

[21] S. Madgwick, "An efficient orientation filter for inertial and inertial/magnetic sensor arrays," *Report x-io and University of Bristol*, p. 32, 2010.

[22] D. Kelly and B. Caulfield, "An investigation into non-invasive physical activity recognition using smartphones." *Conference of the IEEE Engineering in Medicine and Biology Society.*, vol. 2012, pp. 3340–3343, 2012.

[23] M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, P. Navarro, and C. S. Haley, "Application of high-dimensional feature selection: evaluation for genomic prediction in man," *Scientific Reports*, vol. 5, p. 10312, may 2015.

[24] C.-c. Chang and C.-j. Lin, "LIBSVM : A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, pp. 1–39, 2011.

[25] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[26] P. Filzmoser, B. Liebmann, and K. Varmuza, "Repeated double cross validation," in *Journal of Chemometrics*, vol. 23, no. 4, 2009, pp. 160–171.

[27] J. Reynolds, W. Goldsmith, J. Day, A. Abaza, A. Mahmoud, A. Afshari, J. Barkley, E. Petsonk, M. Kashon, and D. Frazer, "Classification of voluntary cough airflow patterns for prediction of abnormal spirometry," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2015.

[28] G. R. Norman, J. a. Sloan, and K. W. Wyrwich, "The truly remarkable universality of half a standard deviation: confirmation through another look." *Expert review of pharmacoeconomics & outcomes research*, vol. 4, no. 5, pp. 581–585, 2004.

[29] S. S. Farivar, H. Liu, and R. D. Hays, "Half standard deviation estimate of the minimally important difference in HRQOL scores?" *Expert review of pharmacoeconomics & outcomes research*, vol. 4, no. 5, pp. 515–523, 2004.

[30] K. W. Wyrwich, W. M. Tierney, A. N. Babu, K. Kroenke, and F. D. Wolinsky, "A comparison of clinically important differences in health-related quality of life for patients with chronic lung disease, asthma, or heart disease," *Health Services Research*, vol. 40, no. 2, pp. 577–591, 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15762908

[31] R. A. Lyons, K. Wareham, M. Lucas, D. Price, J. Williams, and H. A. Hutchings, "SF-36 scores vary by method of administration: Implications for study design," *Journal of Public Health Medicine*, vol. 21, no. 1, pp. 41–45, 1999.