Inferring Physical Function from Wearable Activity Monitors: Analysis of Activity Data from Patients with Knee Osteoarthritis

Authors: Vibhu Agarwal, MS\textsuperscript{1}, Matthew Smuck, MD\textsuperscript{2}, Nigam H. Shah, MBBS, PhD\textsuperscript{3}

\textsuperscript{1}Biomedical Informatics Training Program, Stanford University, Stanford, USA  
\textsuperscript{2}Stanford University Hospital and Clinics, Stanford, CA, USA;  
\textsuperscript{3}Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, USA

Corresponding author  
Email: vibhua@stanford.edu

Abstract

\textbf{Background:} Clinical assessments for physical function do not objectively quantify routine daily activities. Wearable activity monitors enable objective measurement of routine daily activities, but do not map to clinically measured physical performance measures.

\textbf{Objective:} The objective of our study was to derive a representation of physical function from daily measurements of routine activity obtained via a wearable activity monitor. We also evaluated our derived measure against objectively measured function using an ordinal classification setup.

\textbf{Methods:} We defined function profiles — representing average time spent in a set of pattern classes over consecutive days. We constructed a function profile using minute-level activity data from a wearable activity monitor available from the Osteoarthritis Initiative (OAI). Using the function profile as input, we trained statistical models that classified subjects into quartiles of objective measurements of physical function as measured via the 400m walk test, the 20m walk test and 5 times sit-stand test. We evaluated model performance on held-out data from the same calendar year as that used to train the models, as well as using activity data two years into the future.

\textbf{Results:} The function profile derived from minute-level activity data can accurately predict physical performance as measured via clinical assessments. Using held-out data, the area under the ROC curve (AUC) obtained in classifying performance values in the 1st quartile was 0.79, 0.78 and 0.72, for the 400m walk, the 20m walk and 5 times sit-stand tests. For classifying performance values in the 4\textsuperscript{th} quartile, the AUC obtained was 0.77, 0.66 and 0.73 respectively. Evaluated on data from two years into the future, for the 20m pace test and the 5 times sit-stand tests, the highest AUC
obtained was 0.77 and 0.68 for the 1st quartile and 0.75 and 0.70 for the 4th quartile respectively.

**Conclusions:** Function profiles accurately represent physical function, as demonstrated by the relationship between the profiles and clinically measured physical performance. Daily estimation of physical performance via function profiles derived from routine activity data may enable remote functional monitoring of patients.

**Keywords:** Physical function profile; wearable activity monitoring; Generalized additive models; passive monitoring

**Introduction**

Physical function is an important indicator of physiological well-being. Recently, physical status has become an outcome of interest within most medical specialties, [1–9] and is increasingly regarded as the “sixth vital sign”[10]. Attempts at arresting and managing functional decline must start with an evaluation of baseline functional status. For example, maximizing improvement in advanced osteoarthritis requires knowing a patient’s baseline function so as to detect any improvement. Therefore, valid metrics to monitor physical function are necessary [11–13].

Physical function is characterized by the ensemble of daily activities an individual is capable of doing. These activities include bathing, dressing, toileting, transfer, continence, and feeding. Collectively, they are referred to as Activities of Daily Living (ADL) [14,15] and are typically measured via surveys. Disability indexes based on ADLs can differentiate healthy aging patients, patients with mild cognitive impairment, and patients with dementia [16]. However, ADL scores may have a response bias from self-reporting, and low sensitivity to changes in high functioning older adults[17]. By contrast, physical performance measures (such as walking and sit-to-stand speeds and grip strength observed under supervision) capture variation across a wider range of physical function, including initial changes in the early stages of decline[18–20]. The chief drawback of such performance measures is that they require substantial time and effort and access to specialized facilities. The relationship between physical activity and performance measures is a topic of active research[21–24].

Therefore, although the need to measure physical function is widely appreciated, self-reported assessments of physical function are inadequate due to biases and difficulties in recalling historical activity, and physical performance measures require adherence to specific test protocols and are usually limited to research settings.

We have created a novel method for inferring physical function based on objective measurements of daily physical activity obtained from a wearable device. Our work enables quantitative monitoring of physical function — the first step toward improved precision in clinical research and practice.

Wearable activity monitors (WAM), typically equipped with one or more accelerometers, provide a convenient way to measure physical activity objectively [25–31]. However, attempts to use WAMs to link measured physical activity and physical function have been limited [32,33]. van Lummel et al. [34] suggest that physical activity and physical performance are associated but independent domains of physical function, and a change in performance (for example, on the 6 minute walk test) need not imply a corresponding
Because WAMs produce an aggregate value proportional to velocity change over a period of time—which is inadequate for distinguishing pattern classes—we hypothesized that patterns recorded by a WAM will correlate with physical function, and may be used to estimate objective measures of physical function. We define instances of pattern classes from daily activity data using an unsupervised approach and used this information to create a function profile, which represents mean allocation of time in different pattern
classes. Using machine learning techniques, we classified activity profiles into discrete quartiles of commonly used measures of physical function, such as the 400 m walk test. To the best of our knowledge, we are the first group to use WAM data to characterize daily physical activity into pattern classes, and infer physical function based on the pattern classes. Our study demonstrates the feasibility of distinguishing physical function categories with high sensitivity and specificity, and there are potential uses in medical research and treatment. Figure 1 illustrates our overall workflow.

Methods

Data
We used publicly available data from the Osteoarthritis Initiative, (OAI, http://www.oai.ucsf.edu), which follows a cohort of subjects who either had clinical diagnosis of osteoarthritis or were at risk at baseline. The OAI has daily accelerometer measurements for subjects who participated in a physical activity study. Participants use an ActiGraph GT1M uniaxial accelerometer (ActiGraph; Pensacola, Florida) continuously for up to 7 consecutive days, except during sleep and water activities. The Actigraph GT1M is a compact, hip-worn device that measures dynamic acceleration in the range of 0.05 g – 2.0 g. Its validity and reliability is established [35–40].

Participants maintained a daily log of water and cycling activities, as the accelerometer may not have been able to capture these accurately. Post facto analysis revealed that participants spent little time in these activities (median 0 minutes/day, interquartile range = 0.0 to 3.4 minutes/day), indicating that little activity was missed. Table 1 summarizes key attributes of the physical activity study data.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Key attributes of the physical activity data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of subjects</td>
<td></td>
</tr>
<tr>
<td>Incidence</td>
<td>1490</td>
</tr>
<tr>
<td>Progression</td>
<td>505</td>
</tr>
<tr>
<td>Control</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>2001</td>
</tr>
<tr>
<td>Gender (% male)</td>
<td>44.53</td>
</tr>
<tr>
<td>BMI (mean, sd)</td>
<td>(28.52, 4.87)</td>
</tr>
<tr>
<td>Mean Comorbidity Index</td>
<td>0.52</td>
</tr>
<tr>
<td>Median days of activity</td>
<td>7</td>
</tr>
</tbody>
</table>

The accelerometer data in OAI consists of activity counts per minute. An activity count is a weighted sum of discretely sampled (30 Hz) values of one-dimensional acceleration. Since zero or low values of activity counts could also arise from non-wear time, we excluded non-wear periods. Continuous runs of zero counts more than 90 minutes long (allowing for interruptions of up to 2 consecutive minutes with fewer than 100 counts) were discarded as non-wear periods [41]. A day with a wear
Objective Measures of Function
The Osteoarthritis Research Society International (OARSI) [42] recommends testing of activities that are typically affected by OA. We selected three OAI performance measures that had equivalents in OARSI’s recommended tests: 1) The 400m walk test (400MWT), which predicts mobility impairment. Longer completion times are associated with a higher risk of mobility limitation and disability (adjusted hazard ratios 4.43, p < 0.001), as well as a higher risk of death (adjusted hazard ratio 3.23, p < 0.001) for subjects in the highest quartile [43]; 2) the 20m walk pace (20MPACE) is the closest available short walk-length evaluation in the OAI dataset. Gait speed, an important functional outcome in knee OA research, is generally evaluated using shorter walk-lengths; 3) the 30 second sit-to-stand test measures sit-to-stand function in subjects with knee or hip OA [42]. The OAI data provided the number of sit to stands per second measured over 5 repetitions (5CSPACE) as a measure of sit-to-stand function, which has good test-retest reliability [44].

Relationship with Daily Activity
Physical function is defined as the repertoire and relative proportion of activities that a subject can perform in a given environment. We recovered segments representing homogenous activities from the daily sequences of counts-per-minute obtained from a WAM and defined \textit{pattern classes} on the basis of similar segments. A subject’s function profile was computed as average minutes allocated to each pattern class. Finally, we inferred mappings from the function profile to the objective measurements of function described in the earlier section using supervised learning.

Pattern Classes and Function Profile
We used the change-point analysis algorithm by Matteson and James [45] to segment counts-per-minute sequences. This algorithm searches for segment boundaries such that each segment represents a change in the distribution of the time-ordered counts with respect to preceding and subsequent segments. Figure 2A illustrates a counts-per-minute sequence for a typical subject on a given day and the segments that are recovered via change-point analysis. Each segment was indexed using mean and standard deviation (SD) of the counts-per-minute values. This representation improves discrimination between classes of activity patterns — henceforth referred to as \textit{pattern classes} [46].

A \textbf{pattern class} is a bounded region in the segment feature space. Our feature space $F$ consists of all $(m, s)$ vectors: $m \in [0, M], s \in [0, S]$, where $M$ and $S$ are the maximum mean and SD over all segments found via the segmentation. A pattern class is defined
Figure 2. A) Segmentation of counts per minute sequences. Horizontal red lines mark segments recovered via change point analysis. Each segment is an instance of a pattern class whose mean and SD is estimated by the sample mean and SD of the segment. B) The mean and SD space containing all segments is partitioned into bounding regions, each defined by a mean and an SD interval. The shaded region represents pattern class \((m_1, m_2, [s_1, s_2])\). C) Scatter plot of the mean, sd and duration of the segments by a pair of intervals such as \(([m_1, m_2], [s_1, s_2])\). A segment with mean \(x\) and SD \(y\) \((m_1 \leq x < m_2, s_1 \leq y < s_2)\) is an instance of the pattern class so defined. Figure 2B illustrates such a segment represented in the mean-sd space spanned by all segments and its assignment to a pattern class, indexed by the mean-sd interval pair for the pattern class. Based on the pattern classes obtained from partitioning \(F\), we define a function profile for each subject as the average time allocated to each pattern class per day. The function profile for a subject \(i\) is given by

\[A_i = (a_{i1}, a_{i2}, \ldots a_{ij})\]

where
As seen in figure 2C, the number of instances of a pattern class decrease as the mean and SD increase resulting in a sparse daily activity profile.

**Supervised Learning**

We defined a composite descriptor $D_i = \{\text{BMI}_i, \text{Age}_i, \text{Sex}_i, \text{Height}_i, \text{OA}_i, A_i\}$ for each subject $i$ in our data, where $\text{OA}_i$ refers to a subject’s baseline status (healthy, at-risk or progressive disease) and $A_i$ is the function profile. A regression function $f(D_i)$ that maps $D_i$ to an objective measure of physical performance can be obtained by minimizing the expected squared error loss.

Medical studies commonly group continuous variables into quantiles for ease of interpretation and analysis\[43,47,48\]. We therefore defined our response variable by grouping the objective measure of physical performance into ordered categories $1 < 2 < 3$. As shown in Figure 1, categories 1 and 3 represented values in the lowest and highest quartiles respectively and 2 represented values spanning the inter-quartile range for a specific response.

Generalized Additive Models (GAM) can identify and characterize non-linear regression effects through an additive specification of non-parametric functions of the predictors. We use GAMs because fits from quantitative regressions suggest that at higher values, linearity in the predictors may not be a justifiable modeling assumption (supporting information S1). A GAM may be specified as follows

$$g(\mu(X)) = \alpha + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p)$$

where

- $\mu(X)$ denotes the conditional mean of the response i.e $E[Y|X]$
- $g(\cdot)$ is the link function
- $f_1, \ldots, f_p$ are the unspecified smooth functions for each of the $p$ predictors.

Unspecified functions of the predictors are smoothers (typically kernels or cubic splines) that are estimated simultaneously using a backfitting algorithm\[49\]. The estimated $f_i$ reveal the nature of the predictor-response relationship. The function profile depends on the pattern classes, which are defined as intervals in the mean-SD space covering all segments. The size of the 2D interval in feature space that defines our pattern classes is a tuning parameter. Small intervals allow instances from adjacent classes to be in close proximity, increasing correlation between activity profile elements. We ascertained the optimal size of the 2D interval through repeated 5-fold cross validation on our training data as shown in Figure 3. We evaluated cross validation performance using the mean Goodman Kruskal Gamma\[50\], which measures the rank correlation between the true and the predicted categories (supporting information 1). For optimal bin sizes, GAMs were
refit using the full training data and features based on the optimal bin size, and ordered categorical for the response family using the mgcv package[51]. We evaluated the area under the receiver operating curve (AUC) for discriminating category 1 versus others, and category 3 versus others, using the held-out data.

**Results**

As described in Materials and Methods, we found homogeneous segments from daily activity sequences of counts-per-minute and defined pattern classes on the basis of similar segments. A subject’s function profile was average daily minutes allocated to each pattern class. Finally, we inferred mappings from the function profile to the objective measurements of function. The function profile obtained via a wearable activity monitor can predict clinically measured physical performance.

**Predicting Ordered Categories**

Table 2 summarizes classifier performance for the GAMs on the held out data using the function profiles based on the optimal interval sizes. AUC 1 and 3 refer to the performance in the 1 versus rest and 3 versus rest classification tasks, respectively.

<table>
<thead>
<tr>
<th>Interval size</th>
<th>500,500</th>
<th>700,700</th>
<th>2100,2100</th>
<th>Without Activity Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>400MWT</td>
<td>0.62</td>
<td>0.62</td>
<td>0.60</td>
<td>0.52</td>
</tr>
<tr>
<td>20MPACE</td>
<td>0.49</td>
<td>0.52</td>
<td>0.56</td>
<td>0.46</td>
</tr>
<tr>
<td>5CSPACE</td>
<td>0.43</td>
<td>0.55</td>
<td>0.50</td>
<td>0.47</td>
</tr>
<tr>
<td>AUC 1</td>
<td>0.73</td>
<td>0.75</td>
<td>0.78</td>
<td>0.74</td>
</tr>
<tr>
<td>AUC 3</td>
<td>0.76</td>
<td>0.77</td>
<td>0.77</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Table 2 Gamma and AUC for GAMs evaluated on held-out data from Y4. Shaded columns represent models with the optimal interval size obtained in cross validation.

Including the activity profile improved the held out AUC by 1% – 4% and Gamma by 3%-10%, compared to classifiers in which activity profile was excluded from the predictors, with higher improvement in classification of walking performance.
(400MWT and 20MPACE). The lowest quartile classification in 20MPACE and 5CSPACE improved the least with the inclusion of the function profile. At the optimal cutoff point, the AUCs for the 400MWT function profile based on an interval size of (700,700) had sensitivity = 0.72 and specificity = 0.71 for the lowest quartile. In the highest quartile, sensitivity was 0.67 and specificity was 0.74. The optimal cut-off points for classifying gait speed in the lowest quartile for 20MPACE with the same interval size were 0.76 for sensitivity and 0.68 for specificity.

**Predictor of Physical Performance**

GAMs fit smooth functions for each predictor in the model that additively contribute to the value of a latent variable. The model fitting algorithm[49] also estimates thresholds, whose values in relation to the latent variable computed from the smooth functions determines the ordered categorical response. Relationships between response and predictors in a GAM may be studied via plots of smoothers fitted by the GAMs. We studied predictors that were significant at p = 0.05 in the GAMs. These are presented in figure 4. We refer to a specific pattern class using the mean and SD interval pairs, as defined in Materials and Methods. The smooth function plots for pattern classes ([0,700), (701,1400)) & ([701,1400), [701,1400)) (numbered 2 and 3 respectively on the mean-sd grid in figure 4) show that up to 25 daily average minutes in these pattern classes were monotonically associated with improved response in the 400MWT and the 20MPACE.

![Figure 4](image_url)

**Figure 4** The grid boxes represent pattern classes labelled with the mean interval (X axis) and the SD interval on the Y axis. Predictor response relationships are shown by the smooth function plots arranged around the grid and linked to the corresponding predictor.
Most pattern classes with a higher mean and SD have low duration, resulting in fewer degrees of freedom for estimating the smooth functions at high values. This explains the wider confidence bands for the function estimates at higher values of daily average minutes. Inspection of a sample of instances from the class ([701 – 1400), [701 – 1400)) revealed that long duration instances were typically spells of rest punctuated by frequent interruptions. A plausible explanation may be that such interruptions involve sit-stand transitions and therefore, these instances are associated with improved 5CSPACE response. Pattern classes with the mean interval (0 – 700) are not associated with 5CSPACE performance.

Pattern classes ([2801,3500), [0,700)) and ([2801,3500), [1401,21000)), numbered 4 and 5 respectively, are associated with 400MWT performance. The smooth function plots for these classes suggest that higher daily average minutes in both classes were associated with improved long walk performance. The association with increased completion times with > 20 daily average minutes in the class ([2801,3500),[0,700)) was due to instances comprising of mostly sedentary activity. Infrequent occurrences of such instances also resulted in wide estimate intervals for the smooth function.

Approximations of the distribution of activity counts in any given pattern class can be obtained via tail probability bounds. For example, use Chebyshev’s inequality,

\[ P(|X - \mu| > k\sigma) < \frac{1}{k^2}, \]

where \( \mu \) and \( \sigma \) are the mid-interval values of the mean and SD intervals respectively for a given pattern class, for the class ([2801,3500), [0,700)) we obtained:

\[ P(|X - 3150| < k \cdot 350) = 0.69 \text{ for } k = 1.8, \]

implying that at least 70% of activity counts per minute were between 2520 and 3780. Thus, most of the activity in the class ([2801,3500), [0,700)) was likely to be in the lower moderate intensity range. Similarly, for the class ([2801,3500], [1401,2100)), we noted that at least 70% of activity counts per minute were below 6300, which indicates a mix of activity moderate and vigorous activity.

**Moderate to Vigorous Activity with Malalignment**

In the pattern class ([3501,42000), [701,14000)), numbered 6 in the mean-sd grid of Figure 4, an increase in daily average minutes was monotonically associated with improved responses in all three performance measures up to 20 minutes per day. However, an increase of more than 20 minutes was associated with a decline. Unlike the classes with low mean and SD, instances longer than 20 minutes did not represent sedentary activity. A drop in physical function with increased time in moderate to vigorous activity is counter-intuitive. To understand this finding, we reviewed patient-reported outcomes on the Physical Activity Scale for the Elderly (PASE). The PASE measures engagement in different kinds of daily activities related to leisure, household, and occupational work in the elderly [52]. We also reviewed joint exam results reporting varus (bow-legged) and valgus (knock-kneed) alignments for the same subjects. This information is summarized in table 3. It
suggests that subjects with more than 20 daily average minutes in the pattern class ([3501,4200), [701,1400]) had a higher prevalence and severity of knee deformity, higher time in the pattern class (minutes as well as frequency), and fewer sitting hours along with more walking hours per week.

Table 3 Knee deformity and PASE results of subjects with at least one instance of the pattern class (3501-4200, 701-1400).

<table>
<thead>
<tr>
<th>Daily average minutes</th>
<th>&gt;= 20</th>
<th>&lt; 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of subjects with:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Varus or valgus deformity in both knees</td>
<td>12</td>
<td>255</td>
</tr>
<tr>
<td>Varus or valgus deformity in either knee</td>
<td>10</td>
<td>179</td>
</tr>
<tr>
<td>Joint laxity (mild – severe) in either knee</td>
<td>8</td>
<td>109</td>
</tr>
<tr>
<td>Average number of days per week in activity</td>
<td>4.3</td>
<td>1.7</td>
</tr>
<tr>
<td>Average number of minutes per week in activity</td>
<td>161</td>
<td>43.4</td>
</tr>
<tr>
<td>% with sitting hours &lt;2, 2 – 4, &gt;4 per day in last 7 days (PASE)</td>
<td>25, 50, 25</td>
<td>19, 55, 24</td>
</tr>
<tr>
<td>% walking &lt; 1, 1 – 2, 2 – 4 hours a day in last 7 days (PASE)</td>
<td>25, 33, 33</td>
<td>40, 40, 11</td>
</tr>
</tbody>
</table>

Studies have suggested that in subjects with knee malalignment or laxity, altered tibiofemoral loading could be responsible for biomechanical damage and OA progression[53–55]. A much debated view on the role of the quadriceps in OA is that greater muscle strength in malaligned or lax knees increases the risk of OA progression[56,57]. If the relationship between lower extremity strength and the risk of OA progression is confounded by knee alignment status, a plausible explanation for the decreasing trend discussed above may be that regular investment in the pattern class ([3501,4200), [701,1400]) promotes muscle strength but advances OA in subjects with malaligned knees. Though the current guidelines for knee OA management recommend muscle strengthening, our analysis highlights the need for a mechanistic investigation of greater power, given that muscle strength is a modifiable risk factor in OA.

Discussion
To infer physical function from a daily activity trace, it is necessary to derive a representation that conveys information about the daily activity mix. We defined distinct segments from daily activity traces as instances of a set of pattern classes. Doing so transforms a sequence of activity counts into a sequence of pattern classes. Pattern classes provide an informative view of daily physical activity from the perspective of functional ability. Our approach of unsupervised segmentation and the subsequent definition of a set of pattern classes allows a function based
comparison among subjects without the overhead of obtaining annotated activity traces from subjects. This comparison is based on objective measurements, and is perhaps the first effort to interpret functional capability from wearable device data, within a clinical research use case.

There are two main limitations of our methods. First, the mean and SD are likely to be inadequate representations of the activity generating processes, as they ignore temporal relationships between activity counts. Modeling class instances as subsequences generated by a random process has been proposed, [58] and may improve detection of pattern classes. Second, our approach ignores time ordering between pattern class instances in the daily activity profile. One way to address these limitations may be to learn within- and inter-class relationships for a set of daily activity sequences, as a single Bayesian network.

Classifying physical function may be useful in several areas. For example, evaluating alternative rehabilitation programs for OA patients recovering from knee surgery requires a comparative assessment of outcomes that include functional recovery. Alternatives to outpatient physical therapy[59] are a topic of active research, as care providers focus on improving utilization while assuming financial accountability for the full care episode. Remote monitoring of physical function in daily living may allow rehabilitation programs to be evaluated in populations larger than those in clinical trials.

Conclusions

An assessment of physical function based on the ability to perform routine tasks in daily living is desirable. Widely available wearable motion sensors can record daily activity objectively and unobtrusively. We have created an approach for deriving a function profile that represents time spent in different tasks encountered in daily living. Classifiers trained on the function profile were able to predict highest and lowest quartile results of clinically used physical performance measures. We recovered associations between pattern classes and physical performance measures, some of which have support in prior OA research. Capturing physical function as a function profile can enable remote monitoring of patient physical function.

Acknowledgements

This work was support by gifts and unrestricted funds. Additionally the authors thank Jason Fries for his suggestions while editing the manuscript and Professor Trevor Hastie for his advice pertaining to data analysis.

Conflicts of Interest

None declared.

References


29. Cadmus-Bertram L, Marcus BH, Patterson RE, Parker BA, Morey BL. Use of the Fitbit to Measure Adherence to a Physical Activity Intervention Among Overweight or Obese, Postmenopausal Women: Self-Monitoring Trajectory During 16 Weeks. JMIR mHealth uHealth. JMIR mHealth and uHealth;


50. Agresti A. An Introduction to Categorical Data Analysis [Internet]. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2007. PMID: 14562513


