Prediction of patient-reported physical activity scores from wearable accelerometer data: a feasibility study

Ines Bahej, Ieuan Clay, Martin Jaggi and Valeria De Luca

***Abstract*— Many diseases are characterized by limitations in mobility, including a wide range of musculoskeletal and neurological conditions. Reduced mobility impacts a patients ability to perform activities of daily living, which in turn reduces health-related quality of life. Mobility can be assessed by collecting patient-reported outcome scores from standardized questionnaires and by directly measuring physical activity parameters from wearable accelerometer data. In this work, we explored the relationship between subjectively and objectively measured mobility by training machine learning models to predict patient responses based on features derived from real-world acceleration data. Our method achieved up to 82% accuracy using a random forest classifier and set the basis to develop novel data-driven digital biomarkers for objective, quantitative and more frequent evaluation of patients' mobility.**

I. Introduction

Improving the quality of life of patients is a core objective of all clinical programs. Patient mobility, physical activity (PA) and the ability to perform activities of daily living (ADLs), such as walking, carrying groceries or bathing, are important indicators of quality of life (QoL) [1]. This is particularly relevant for patients affected by musculoskeletal diseases [2]. Current clinical assessment of QoL factors relies on patient-reported outcome (PRO) scores, standardized questionnaires and diaries, which are often condition-specific. A common questionnaire is the 36-Item Short-Form Health Survey (SF-36). PROs are an important basis for assessing the efficacy of clinical interventions [3], yet these scores are subjective, biased and infrequently collected. Recently, wearable accelerometers have been increasingly used to monitor and quantify PA [4] and gait parameters [5], and for activity recognition [6]. While not yet fully translated into clinical applications, this technology allows for passive monitoring over a long period of time, outside the clinic, and hence it is highly valuable in evaluating the patient's everyday mobility [7]. Yet the relationship between the aforementioned objective measures and the patient perceived mobility and related impact on QoL has not been thoroughly investigated [8]. In this work, we explore the feasibility of predicting PROs on PA (from a subset of answers to SF-36) from passively acquired, real-world mobility data from a wearable accelerometer, using machine learning models. We aim to provide novel objective scores of mobility to evaluate the patient status and quantify treatment efficacy.

We thank Arne Müller and Danny Tuckwell for preparing clinical trial data, Ronenn Roubenoff, Jason Laramie and Scott Kennedy for support.

Ines Bahej, Ieuan Clay and Valeria De Luca are with the Novartis Institutes for Biomedical Research, Basel, Switzerland (corresponding author: valeria.de_luca@novartis.com). Martin Jaggi is with the Department of Computer Science, EPFL, Switzerland.

II. Material and Methods

Data: We considered a total of 165 patients (51% female and 49% male, 70-95 years of age, median 79) affected by a chronic and degenerating musculoskeletal condition. These patients are enrolled in an ongoing clinical trial and hence further information cannot be disclosed. All Ethics Committees involved in the study granted approval and written consent was given by every participant. Patients were asked on a monthly basis, for up to 6 months, to fill in the SF-36 Health Survey and to wear the *actibelt* sensor (Trium Analysis Online GmbH, Germany) around their waist for at least one continuous week prior to each survey answer. The *actibelt* system continuously collects 3D accelerations via tri-axial accelerometer of range ± 6 g from the center of gravity of the subject at 100 Hz. Gait speed, activity counts, walked distance, number of steps and wear-time are then derived at 1 Hz. To validate the proposed prediction of general mobility scores, we selected the following subset of SF-36 questions that are directly related to PA: (walking) "Does your health now limit you in walking *more than a mile* [Q1] / *one hundred yards* [Q2] / *several hundred yards* [Q3]? If so how much?"; (ADLs) "During the past week, how much of the time have you *cut down on the amount of time you spent on* [Q4] / *had difficulty performing* [Q5] work or other activities as a result of your physical health?". At each assessment time t, multiple-choice answers were converted to an integer score $q(t)$: (Q1-3) Yes, limited a lot ($q = 0$) / yes, limited a little (1) / no, not limited at all (2); (Q4-5) All of the time ($q = 0$) / most of the time (1) / some of the time (2) / a little of the time (3) / none of the time (4).

Features: A total of 162 clinically relevant acceleration features were considered as predictors, namely: mean, median, standard deviation, 5th and 95th percentiles of speed, activity count and walked distance; and total number of steps, activity counts and walked distance. We computed daily summary features from the day when the patient filled in the SF-36 survey d_0 to one week prior to it $\{d_{-7}, \ldots, d_{-1}\}$, and weekly ones over $[d_{-7}, \ldots, d_0]$. For this population strong day-to-day variation is not expected, therefore data for missing days were replaced by averages over the previous 7 days, allowing for gaps of at most 2 consecutive days.

Prediction: For each patient k, assessment t and question i, we converted answers $q_{k,i}(t)$ to binary classes to define targets $y_{k,i}(t)$, see Tab. I. This conversion was based on the distribution and clinical relevance of answers. We also fused the targeted questions into walking (Q1-3) and other ADLs (Q4-5) questions, calculated new scores as

TABLE I

SUMMARY OF THE RF CLASSIFICATION RESULTS.

PRO question	Target	Class distribution	F1 scores	Accuracy
Q1	$y=0$ if $q=0$ $y=1$ if $q\in\{1,2\}$	0.28 0.72	0.40 0.84	0.75 ± 0.08
Q2	$y=0$ if $q\in\{0,1\}$ $y=1$ if $q=2$	0.27 0.73	0.29 0.85	0.76 ± 0.09
Q3	$y=0$ if $q\in\{0,1\}$ $y=1$ if $q=2$	0.38 0.62	0.43 0.77	0.69 ± 0.08
Q1/Q2/Q3	$y=0$ if $q\in\{0,1,2\}$ $y=1$ if $q\in\{3,\ldots,6\}$	0.21 0.79	0.32 0.89	$\mathbf{0.82 \pm 0.07}$
Q1/Q2/Q3	$y=0$ if $q\in\{0,1,2\}$ $y=1$ if $q\in\{3,4\}$ $y=2$ if $q\in\{5,6\}$	0.21 0.22 0.57	0.42 0.07 0.72	0.60 ± 0.10
Q4	$y=0$ if $q\in\{0,1,2\}$ $y=1$ if $q\in\{3,4\}$	0.24 0.76	0.10 0.84	0.74 ± 0.11
Q5	$y=0$ if $q\in\{0,1,2\}$ $y=1$ if $q\in\{3,4\}$	0.32 0.68	0.35 0.79	0.70 ± 0.09
Q4/Q5	$y=0$ if $q\in\{0,\ldots,4\}$ $y=1$ if $q\in\{5,\ldots,8\}$	0.22 0.78	0.09 0.85	$\mathbf{0.75 \pm 0.11}$
Q4/Q5	$y=0$ if $q\in\{0,1,2\}$ $y=1$ if $q\in\{3,4,5\}$ $y=2$ if $q\in\{6,7,8\}$	0.05 0.26 0.69	0.00 0.06 0.79	0.66 ± 0.12

the sum of the individual ones, and defined new binary and 3-class targets. Based on the aforementioned features $\mathbf{X} \in \mathbb{R}^{162\times516}$, we predicted PRO targets $y(q)$ by training binary and 3-class random forest (RF) classifiers. Individual models were trained and validated using patient-stratified 10-fold cross-validation (CV), to predict each $y_{k,i}(t)$, with $i \in \{1;\ldots;5;walk;ADL\}$. Hyperparameters were selected via nested 10-fold CV randomized search.

III. RESULTS

Table I summarizes the mean accuracy and classes F1 score for all considered targets on the left-out patient groups. For individual questions on walking (other ADLs) accuracy ranged from 69% to 76% (70% to 74%). When fusing answers to similar questions, accuracy increases to 82% and 75% for walking and other ALDs outcome scores, respectively. For all prediction tasks, F1 scores of the majority class are higher than the minority ones. Binary RF classifiers were compared to support vector machine (SVM) models with 3rd-degree radial basis function kernel, trained using the same approach. SVM achieved slightly lower accuracies of 0.76 ± 0.07 and 0.72 ± 0.07 for walking and other ADLs, respectively, but higher F1 scores for $y = 0$, with F1s of [0.40, 0.85] and [0.32, 0.81].

IV. DISCUSSION

The proposed RF classifiers can accurately and robustly predict PRO scores on ADLs, with an accuracy between 69% and 82%. The most predictive features were speed standard deviation, walked distance and activity counts features, and number of steps. When observing individual classes, results for class $y = 0$ are significantly poorer. This might be due to the much smaller representation of this class in our dataset. Principal component analysis (PCA) allows us to visualize the original high-dimensional $\mathbf{X}$ by projecting it onto two dimensions. Fig.1 shows $\mathbf{X}$ after projection onto the first 2 PCA components for Q1. The subjective nature of SF-36 and the limited amount of observations, especially for the minority class, could contribute to the fact that no clear clusters can be easily identified and reflect the lower F1 scores. Compared to SVM, the success of RFs in our experiments is partially enabled by the robustness of RF to class imbalances. Data augmentation and oversampling of the minority class might improve performance. Grouping similar questions defines more general mobility scores, which are less subject to patient's biases, reflected in the better results versus individual questions.

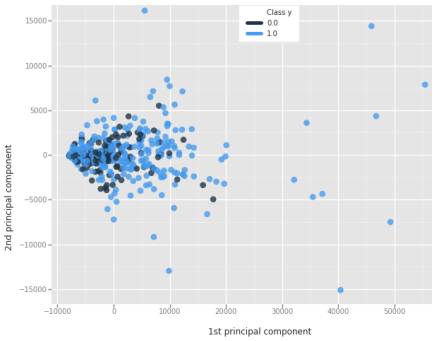

Fig. 1. First and second PCA components of $\mathbf{X}$ and corresponding y_1 (Q1).

V. CONCLUSIONS

We presented a novel approach towards quantifying and objectifying scores to evaluate patient mobility and capability of performing ADLs. Our method is based on predicting PROs from the SF-36 Health Survey, using random forest classifiers trained on features extracted from wearable accelerometer data. An accuracy of up to 82% was achieved. The subjective nature of the survey, the difficulty for patients to objectively answer to quantitative questions, and their impact on our results should be further investigated. Future work will focus on collecting more data, exploring deep learning approaches and temporal prediction, and formulating different definitions of QoL mobility functions.

REFERENCES

[1] B. Ainsworth, L. Cahalin, M. Buman, and R. Ross, "The current state of physical activity assessment tools," *Prog Cardiovasc Dis*, vol. 57, no. 4, p. 387, 2015.

[2] A. Banerjee, S. Jadhav, and J. Bhawalkar, "Limitations of activities in patients with musculoskeletal disorders," *Ann med health sc res*, vol. 2, no. 1, p. 5, 2012.

[3] P. R. Deshpande, S. Rajan, B. L. Sudeepthi, and C. A. Nazir, "Patient-reported outcomes: a new era in clinical research," *Perspect clin res*, vol. 2, no. 4, p. 137, 2011.

[4] J. Staudenmayer, S. He, A. Hickey, J. Sasaki, and P. Freedson, "Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements," *J Appl Physiol*, vol. 119, no. 4, p. 396, 2015.

[5] U. Della Croce, A. Cereatti, and M. Mancini, "Gait parameters estimated using inertial measurement units," *Handb Hum Motion*, p. 245, 2018.

[6] A. Jordao, L. A. B. Torres, and W. R. Schwartz, "Novel approaches to human activity recognition based on accelerometer data," *Signal Image Video P*, p. 1, 2018.

[7] J. Goldhahn, "Need for digital biomarkers in musculoskeletal trials," *Digital Biomarkers*, vol. 1, no. 1, p. 82, 2017.

[8] S. Kumar, T. Quisel, L. Foschini, and J. Juusola, "Digital trackers show that high intensity excercising and consistent sleep patterns are associated with positive self-reported health status," in *Value Health*, vol. 20, no. 5. Elsevier, 2017, p. A54.