Discovering Physical Activity Profiles in COPD Patients Using Topic Models

Gabriele Spina, Piero Casale, Jörg D. Leuppi, Paul S. Albert, Emiel F. M. Wouters, Nidia A. Hernandes, Sally J. Singh, Frank W. J. M. Smeenk, Arnoldus J. R. van Gestel, Judith Garcia-Aymerich, Richard W. Costello, Ruth Tal-Singer, Jennifer Alison, Martijn A. Spruit, and Albertus C. den Brinker

***Abstract*—With the growing amount of physical activity (PA) measures available, the need for methods and algorithms to automatically analyze and interpret unannotated data increases. In this paper PA is seen as a combination of multimodal constructs that can co-occur in different way and proportion during the day. The design of a methodology able to integrate and analyze them is discussed and its operation is illustrated by applying it to a data set comprising data from 997 COPD patients and 66 healthy subjects. The method encompasses different stages. The first stage is a completely automated method of labeling low-level multimodal PA measures. The information contained in the PA labels are further structured using topic modelling techniques, a machine learning method from the text processing community. The topic modelling discovers the main themes that pervade a large set of data, in this case it discovers PA routines that are active in the assessed days of the subjects under study. Applying the designed algorithm to our data provides new learnings and insights. As expected, the algorithm discovers that the PA routines for COPD patients and healthy subjects are substantially different regarding their composition and moments in time in which transitions occur. Furthermore, it shows certain consistent trends relating to disease severity as measured by standard clinical practice.**

***Index Terms*—Activity routine, COPD, topic models.**

G. Spina is with the Department of Electrical Engineering, Eindhoven University of Technology, and with Smart Sensing and Analysis Group, Philips Research, Eindhoven, The Netherlands, e-mail: g.spina@tue.nl.

P. Casale is with the Department of Electrical Engineering, Eindhoven University of Technology, The Netherlands.

J. D. Leuppi, is with Medical University Clinic, Cantonal Hospital Baselland, Liestal and Medical Faculty, University of Basel, Basel, Switzerland.

P. S. Albert is with School of Ageing and Chronic Disease, University Hospital Aintree, Liverpool, United Kingdom.

E. F. M. Wouters is with the Department of Respiratory Medicine, Maastricht University Medical Center+ (MUMC+), Maastricht, The Netherlands.

N. A. Hernandes is with the Laboratory of Research in Respiratory Physiotherapy, Department of Physiotherapy, State University of Londrina (UEL), Londrina, Brazil.

S. J. Singh is with the NIHR EM CLAHRC - Centre for Exercise and Rehabilitation Science, University Hospitals, Leicester, United Kingdom.

F. W. J. M. Smeenk is with the Department of Respiratory Medicine, Catharina Hospital, Eindhoven, The Netherlands.

A. J. R. van Gestel is with the Pulmonary Division, University Hospital of Zurich, Zurich, Switzerland.

J. Garcia-Aymerich is with the Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain.

R. W. Costello is with the Department of Respiratory Medicine, Beaumont Hospital, Dublin, Ireland.

R. Tal-Singer is with GlaxoSmithKline R&D, King of Prussia, PA, United States of America.

J. Alison is with Clinical and Rehabilitation Sciences, The University of Sydney, Sydney, NSW, Australia.

M.A. Spruit is with Department of Research & Education, Center of expertise for chronic organ failure (CIRO+), Horn, The Netherlands.

A.B. den Brinker is with Smart Sensing and Analysis Group, Philips Research, Eindhoven, The Netherlands.

I. Introduction

THE prevalence of chronic diseases in general is rising due to an aging population as well as to environmental and lifestyle changes. This is particularly true for respiratory diseases such as Chronic Obstructive Pulmonary Disease (COPD), which is a progressive and irreversible disease that results in airflow limitation and significant extra pulmonary effects which limit physical activities [11, 20]. Physical activity (PA), defined as any bodily movement produced by skeletal muscles that requires energy expenditure [5], is known to be a relevant indicator of COPD patients health state. Research on physical activity levels of COPD patients has consistently shown that COPD patients have lower physical activity levels than their healthy peers [23]. Moreover, reduced levels of PA have been found to be related to an increased risk of hospital admission and mortality among COPD patients [10, 18, 24]. Outcome variables related to this type of analysis mainly focused on amount and volume parameters, such as number of steps, volume of physical activity as expressed by total number of counts, and total energy expenditure. Although these are important health markers of patients suffering from COPD, interventions thus far have failed to demonstrate important increases in physical activities in patients with COPD. A better insight into daily PAs of patients with COPD needs to be achieved in order to assist in targeted therapeutic strategies and personalized coaching programs. In order to quantify PA it is not only necessary to measure a set of multidimensional parameters as the one mentioned above, but also their co-occurrence and temporal pattern. Moreover, when analyzing PA, physiological responses such as heart rate, temperature or galvanic skin response should be considered complementary constructs that needs to be harmoniously integrated with PA measures into meaningful descriptors. With recent improvements in wearable sensor technologies it becomes easier in daily life to acquire massive amounts of different sensors data. At the same time, however, it is difficult to combine and extrapolate meaningful information in absence of any supervision or annotation. A PA descriptor could be seen as a composite of multiple low-level PA measures, including their physiological responses, that can co-occur in a different way and proportion i.e., PA routines. Routines could also occur at different time and in different proportion across the day for different patients or subgroups of patients characterizing their activity behaviour. The reader might think about PA measures and physiological responses as the letters composing

the words that describe PAs. The co-occurrency of these words creates groups of PA constructs describing the main topics that pervade the day of a patient. This work aims at studying low-level multimodal sensor data in order to find unknown and characteristic PA structures, able to quantify difference among COPD and healthy subjects (matching for gender age and BMI) and within COPD severity classes. This is particularly difficult in this patient population since it is known that COPD patients maintain a constant inactive behaviour during the day that makes standard activity recognition tools not suitable for the purpose.

In particular, this paper provides the following contributions: (1) we propose a methodology to create a vocabulary of meaningful words from a set of mutimodal PA measures without the need of any supervision or parameter tuning, (2) we discover PA routines that pervade daily life of COPD patients. Finally, (3) we infer the underlying PA routine structure of numerous patients data quantifying differences between COPD patientsa and healthy subjects and among COPD patients. In particular, for each assessed day we infer which is the distribution over the routines in day segments of 30 minutes, describing in such a way the temporal regularities of the multidimodal PA measures.

II. Related works

The automatic monitoring and analysis of chronic diseases has always been central in research on wearable sensors. In particular, the continuous monitoring of COPD patients has gained considerable interests in the recent years. Liao et al. [14] provided a review focused on describing current wearable technologies for measuring the physical activity level of COPD patients. Dimensions such as reliability, validity, advantages and limitations are discussed. Of particular interest is the work of Patel et al. [16] where a comparative study of machine learning techniques is presented in order to track changes in physiological responses of COPD patients with respect to their physical activity level. They used motion data to monitor activities in conjunction with heart-rate and respiration rate to capture the physiological responses of the patients while performing a set of activities. Beattie et al. [1] considered how the early detection of disease exacerbation can lead to earlier provision of intervention advice. These authors focus on important parameters for patients self-management such as autonomy, methods of data transmission and levels of intrusiveness and propose guidelines for the development of a context-aware system aimed at overcoming current limitations in the perspective of a user-friendly system for the patients. More advanced self-management platforms have been recently proposed. For example, Bellos et al. [2] propose an integrated platform aiming at the effective management and real-time assessment of the health status of COPD patients. A combination of machine-learning techniques was able to provide real-time categorizations of COPD episodes and estimate the severity of pathological situations in different levels, triggering an alerting mechanism for the patient and the clinical supervisor.

Topic models represent a class of algorithms able to discover hidden thematic structure in collections of documents. Due to their pattern discovery nature, they have been widely explored in the wearable sensor and activity recognition community. Huyn at al. [13] showed that the activity patterns discovered using topic modelling approaches correspond to high-level users' behavior. Authors used activity patterns based on a learned vocabulary of meaningful events such as walking, using the phone, discussing at whiteboard, etc. These authors also addressed the point of avoiding supervised learning approaches using unsupervised methods for building the vocabulary using a clustering approach. Qualitative results show that high-level structure of the data as well as activity transitions, novelties and anomalies can be discovered using their approach. Seiter et al. [19] investigated unsupervised activity discovery approaches using three topic model approaches. Authors analyzed three public datasets with different properties affecting the discovery such as primitive rate, activity composite specificity, primitive sequence similarity, and composite-instance ratio. They compared the activity composite discovery performance against the performance of a k-means clustering algorithms providing guidelines for optimal parameter selection. Results indicated that LDA shows higher robustness against noise compared to k-means and other topic modelling approaches.

III. Unsupervised Routine discovery

Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Data are treated as observations arising from a generative probabilistic process in which hidden variables reflect the thematic structure of a collection of documents. The intuition behind the use of the Latent Dirichelet Allocation (LDA) to discover PA routines is that each day is a mixture of thematically coherent PA measures as a text document is a mixture of thematically coherent terms. The graphical model for LDA is provided in Fig. 1. All the assessed

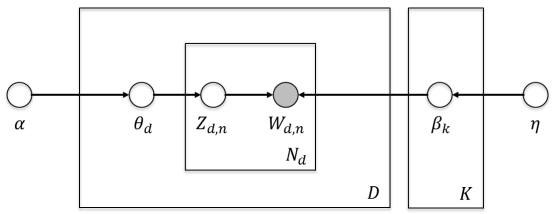


Fig. 1. Graphical model for LDA. Each node is a random variables, edges denote possible dependences. The only observed variables are the words (shaded).

days (also called documents $d_{1:D}$) share the same set of daily routines (also called topics $\beta_{1:K}$) that are defined as Dirichelet distributions over the observed set of PA descriptors (also called words W of a fixed vocabulary). Observing activities in patients is a difficult task since it is time-intensive and intrusive. At the same time patients are not able to accurately self-report their physical activities [17] and the training of a classifier requires annotations that in a daily life scenario are difficult to obtain. In order to make the methodology fully unsupervised we assume that the observed words (input of the model) are composed by multimodal PA measures coming

from a body area sensor network. Each assessed day exhibits PA daily routines in different proportion indicated as $\theta_{1:D}$, i.e. each day has a different distribution over the routines that also follow a Dirichelet distribution. The distribution of the words in a routine and the distribution of the routine in a document depend only on the topic hyper-parameters η and α that control the mean shape and sparsity of the distributions. In such a model the N words ($W_{d,n_{1:N}}$) that compose the D documents are the only random variables observed and depend on the per word topic assignment ($Z_{d,n}$) and all the β_k. The daily routines then are composed indirectly by low-level PA measures that belong, with a certain probability distribution, to different thematic areas. Different routines will have different PA measures with different probabilities. The generative process of the model defines a joint probability distribution over both the observed and hidden random variables, according to:

$$p(\beta_{1:K}, \theta_{1:D}, Z_{1:D}, W_{1:D}) = \prod_{k=1}^{K} p(\beta_k|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N} p(Z_{d,n}|\theta_d) p(W_{d,n}|Z_{d,n}, \beta_{1:K}).$$

Reversing the generative process, it is possible to calculate the hidden structure that likely generated the observed collection of document. More formally the joint probability distribution is used to compute the conditional distribution of the hidden variables given the observed variables by:

$$p(\beta_{1:K}, \theta_{1:D}, Z_{1:D}|W_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, Z_{1:D}, W_{1:D})}{p(W_{1:D})}. \quad (1)$$

This conditional distribution is also called the posterior distribution and it is intractable to compute. Topic modelling algorithms compute an approximation of (1) by finding a distribution over the latent topic structure to be close to the true posterior. In particular, for this analysis, we use variational inference that posits a parameterized family of distributions over the hidden structure and then finds the member of that family that is closest to the posterior according to *Kullback-Leibler* divergence.

IV. Implementation and evaluation

A. Dataset

Physical activity of a subject was assessed during daily life by mean of the SenseWear Armband and SenseWear Mini Armband activity monitors. These devices combine an accelerometer with different physiological sensors (a heat flux sensor, a galvanic skin response sensor, a skin temperature sensor, and a near-body ambient temperature sensor). Together with demographic characteristics, such as gender, age, height and weight, energy expenditure (EE) and metabolic equivalent of task (MET) were estimated using proprietary algorithms developed by the manufacturer. Moreover, information about the sleeping status of a subject (0=awake, 1=sleeping) is also provided. The SenseWear Armband has been shown to be valid both in field [7, 15] and in laboratory studies [6, 8, 12].

Data from 1001 COPD patients (65% men; age, 67 years; FEV_1, 49% predicted; BMI, 25.8) were recorded across 10 different countries. Subjects wore the sensor both during day time and night time so that continuous, non-scripted activities were recorded in a natural environment with 1 minute resolution. A minimum of 4 days (2 weekdays + Saturday + Sunday) was considered acceptable to include a patient in the analysis [25], with the device being used for at least 22 hours per day. In order to minimize the intrinsic variability of data recorded days were synchronized according to the morning time instants after the longest period of sleep (awakening points). Data prior to the awakening point was discarded from the analysis.

The median number of days analyzed per patient was 6 (4 weekdays - 2 weekend days), resulting in a total of 5846 valid PA days assessed, of which 3916 (67%) were weekdays and 1930 (33%) weekend days. A median of 982 minutes were analyzed per patient each day (992 weekdays - 961 weekend days). Ethics Board approval was obtained from the local ethics committees, and written informed consent was provided by participants.

B. Vocabulary

One of the most important choices one has to make when applying LDA to activity data is the nature and number of terms forming the vocabulary. In order to relax this assumption we used a data-driven methodology to automatically create the vocabulary without specifying its size beforehand. Sensor data from a sub-sample of 66 COPD patients and 66 healthy subjects pairwise matched for gender, age and body mass index (BMI) were used for this purpose with a median of 972 minutes analyzed per subject each day (983.5 weekdays - 946 weekend days). Each one-minute data point consists of a 7-dimension measurement vector comprising: *MET*, *Skin Temperature (ST)*, *Galvanic Skin Response (GSR)*, *Longitudinal Acceleration (Acc$_L$)*, *Transversal Acceleration (Acc$_T$)*, *Step Counts (SC)* and *Sleeping Status (SL)*. Acc_L and Acc_T were combined to compute Vector Magnitude (VM). METs data were first divided into activity intensity categories (ICs) using the following thresholds proposed by the American College of Sports Medicine [9]: very light intensity (IC_{VL}), < 2.0 METs; light intensity (IC_L), 2.0 to 2.9 METs; and moderate-to-vigorous intensity (IC_{MV}), ≥ 3.0 METs. Minutes marked by the sensor as sleeping and with METs < 2.0 formed a separated category named sleeping (IC_S). For simplicity, we will be referring to the ICs only with subscripts S, VL, L, MV. Figure 2 shows an example of 1 day METs data stream with the respective ICs superimposed.

Consecutive minutes exhibiting the same ICs are then grouped together in intensity bouts (IB) of variable duration (d). In each IB we calculated the mean μ of ST, GSR, VM and SC. The original sensor data stream is then represented by a series of intensity bouts where each bout is fully characterized by a 6-elements feature vector $\tilde{V}$

$$\tilde{V} = [IC, d, \mu_{ST}, \mu_{GSR}, \mu_{VM}, \mu_{SC}]. \quad (2)$$

Subsequently, for each intensity category (S, VL, L and VL) the most relevant subset of features was selected such that the multi-cluster structure of the data can be best preserved. Features were selected using the Multi-Cluster Feature Selection (MCFS) that deploys spectral regression with l_1-norm

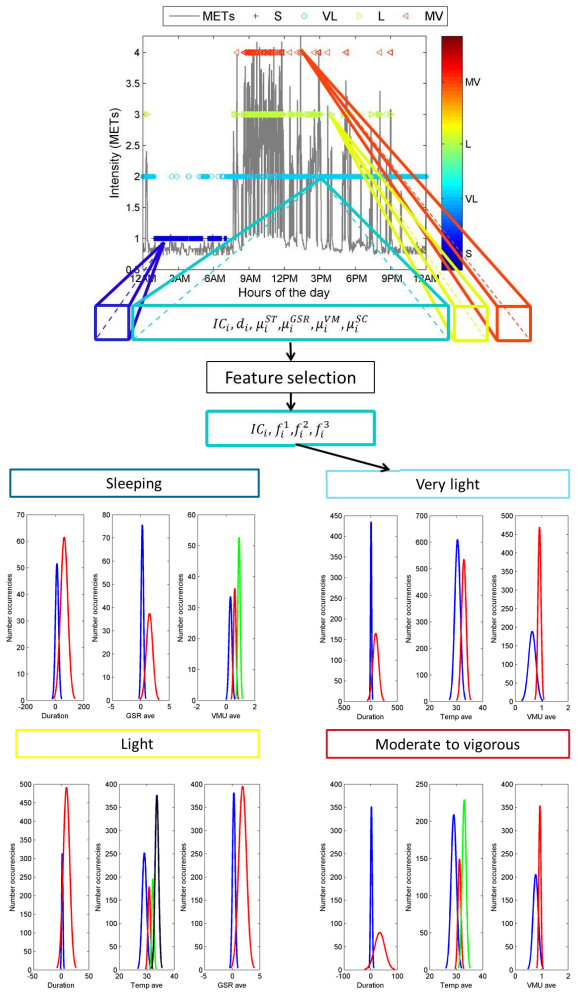


Fig. 2. A continuous data stream representing METs value (grey line) is converted in bouts of PA intensity (blue=sleeping, ciano=verylight, green=light, and red=moderate to vigorous intensities). Features are computed within the bouts and the most relevant will be combine to create PA descriptors.

regularization in order to select features jointly instead of evaluating each feature independently [4]. The feature vector $\tilde{V}$ can then be simplified according to

$$V = [IC, f_1, f_2, f_3]. \tag{3}$$

The reader might think about this step as a stemming procedure that removes redundant letters that might confuse the model. The selected features $f_{j\in\{1,2,3\}}$ obviously might be different for each intensity and are shown in the bottom part of Fig. 2.

To generate the vocabulary of words each of the selected features was first standardized and then mapped into a set of discrete levels using a K-means clustering algorithm. The algorithm automatically selects for each feature the number of levels K^{f_j} (which could be interpreted as the letters of our words) in a way that the corresponding clustering results $L_{p\in\{1:K\}}$ are the most stable under small perturbations of the input dataset as described in [22].
Levels are sorted according to their mean value $\bar{L_p}$ in ascending order such that: first level (L_1) represents clusters with the smallest feature values and the last level (L_K) represents clusters with the highest feature values. Mean value and variance of the levels were stored and used to create the documents as described later in section IV-D. The vocabulary of terms was built by allowing all the possible combinations between levels sharing the same IC. For the sleeping category, for example, the feature $f_1 = d$, $f_2 = \mu_{GSR}$ and $f_3 = \mu_{VM}$ were selected and divided in $K^{f_1} = K^{f_2} = 2$ and $K^{f_3} = 3$ levels respectively. The N^S ($N^S = K^{f_1} \cdot K^{f_2} \cdot K^{f_3}$) terms of the vocabulary describing the sleeping intensity category ($t^S_{i\in\{1:N^S\}}$) are:

$$\begin{aligned} t^S_1 &= \{L^{f_1}_1\_L^{f_2}_1\_L^{f_3}_1\} \\ t^S_2 &= \{L^{f_1}_1\_L^{f_2}_1\_L^{f_3}_2\} \\ &\dots \\ t^S_{12} &= \{L^{f_1}_2\_L^{f_2}_2\_L^{f_3}_3\} \end{aligned} \tag{4}$$

The sum of terms across different activity levels is the total number of words. In particular a total of 48 terms were created (12 for S, 8 for VL, 16 for L and 12 for MV intensity). Note that we did not need to specify the number of unique artificial words (vocabulary size) beforehand. As a last step frequent words (occurring at least in the 94% of the documents) were removed because they will be placed by the model with high probability in all the topics. In other words they occur so frequently that they are more likely to obscure than facilitate a meaningful decomposition of the collection of documents.

C. Topic discovery

For topic discovery we used the LDA implementation of [3] and we considered each day of assessment as a separate document. Each IB was mapped with an instance of the vocabulary by associating the selected features in V with their closest levels and then concatenating the 3 closest levels found. The distance for each bout between the feature point f_j and all the levels L_p are

$$d_p(f_j, L_p) = \frac{\left| f_j - \bar{L}^{f_j}_p \right|}{\sigma_p}, \forall p = \{1, ..., K^{f^j}\}. \tag{5}$$

Once that a term of the vocabulary was assigned to each IB, documents were created constructing for each day a histogram of terms occurrences. We chose the number of topics (T) equal to 18 and set the hyperparameter α equal to 0.01 as in [13]. Hyperparameters are optimized iteratively within a variational expectation maximization (EM) algorithm based on observed words from 18 randomly selected documents (6 healthy subjects, 6 COPD patients).

D. Topic inference

Once the topics are calculated, to know which one was active during the different parts of the day, we inferred documents composed by day segments. Differently from the documents used to discover the topics, each document represents a mixture of terms over a window of time D. We used sliding windows of 30 minutes as suggested in [19]. For each

window we constructed a histogram of terms occurrences by mapping the bouts in D to the words in the vocabulary giving soft assignments. For each feature point in V we calculated the distances $d_{1...K^{f_j}}(f_j, L_p)$ from the mean values of associated levels as in (5). The distances were converted in weights according to

$$w_p(f_j, L_p) = \frac{e^{-d_p}}{\sum_{p=1}^{K^{f_j}} e^{-d_p}} \tag{6}$$

Thus smaller distances imply higher weights and the weights for different levels of one selected feature sum up to one. We then concatenate the weights as we did in (4) creating combination of weights, each of one assigned to the related term of the vocabulary. Summing up the weights of a specific term across the feature selected and dividing by the sum of the weights of all the terms we get values comprised between 0 and 1 that indicate the probability that the term appear in the document segment D. Weights of terms referring to other intensities will be set to 0. Recalling the example with a sleeping bout we have 12 weights G_i^S

$$\begin{aligned} G_1^S &= \frac{w_1^{f_1} + w_1^{f_2} + w_1^{f_3}}{\sum_{t=1}^{N^S} G_t} \\ &\dots \\ G_2^S &= \frac{w_1^{f_1} + w_1^{f_2} + w_2^{f_3}}{\sum_{t=1}^{N^S} G_t} \\ &\dots \\ G_{12}^S &= \frac{w_2^{f_1} + w_2^{f_2} + w_3^{f_3}}{\sum_{t=1}^{N^S} G_t} \end{aligned} \tag{7}$$

We next use the weights associated to each term to construct documents of size D. More specifically, for each term we sum up the weights W_t over a feature window of length D, and then generate m_{W_t} instances of the term by multiplying the sum of this weights by the document length D and rounding to the next integer.

V. Qualitative analysis

Since PA measures during the weekdays and the weekend days are known to be different [25], results were computed separately and only the ones relative to weekdays will be presented. Figure 3 illustrates the distribution of the dicovered routines $\beta_{1,..,18}$ over the terms of the vocabulary. Four routines ($R2$, $R7$, $R13$, $R18$) related to low intensity levels, seven routines ($R4$, $R6$, $R9$, $R11$, $R14$, $R16$, $R17$) related high intensity levels and six routines ($R1$, $R3$, $R5$, $R8$, $R10$, $R12$) composed by a combination of VL, L and MV descriptors were discovered in data from 66 COPD patients and 66 healthy subjects. A separate routine ($R15$) characterizing the sleeping behavior was also found. Each of these routines is characterized by a combination of words that in turn varies in relation to the feature levels. A detailed description of the routines with the 3 most important descriptors associated can be found in table I. The words composing a particular routine are listed together with their occurrence probability (e.g. the first word of $R1$: *[2 1 1 1]* refers to the descriptor $VL\_1^d\_1^{ST}\_1^{VM}$ and has an associated probability equal to 25%). Higher the probability, higher the importance of the descriptor for the routine.

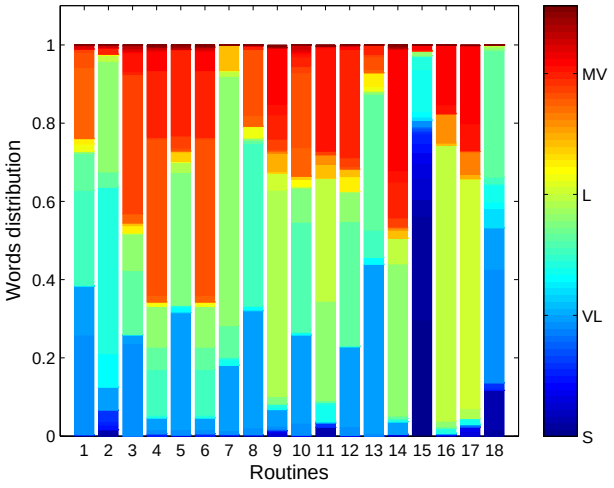


Fig. 3. Distribution over the words

R1	%	R2	%	R3	%	R4	%	R5	%	R6	%
2 1 1 1	0.25	2 2 2 2	0.43	4 1 2 1	0.36	4 1 1 2	0.40	3 1 2 2	0.34	4 1 1 2	0.40
3 1 1 1	0.24	3 1 3 1	0.28	2 1 1 1	0.24	4 1 2 2	0.17	2 1 1 2	0.30	4 1 2 2	0.17
4 1 1 1	0.18	2 2 1 2	0.09	3 1 2 1	0.16	3 1 1 1	0.12	4 1 2 2	0.22	3 1 1 1	0.12
R7	%	R8	%	R9	%	R10	%	R11	%	R12	%
3 1 3 1	0.64	3 1 1 1	0.42	3 1 3 2	0.53	3 1 1 2	0.28	3 1 4 1	0.31	3 1 2 1	0.32
2 1 1 2	0.17	2 1 1 2	0.30	4 1 3 2	0.15	2 1 1 2	0.23	4 1 3 1	0.26	4 1 2 2	0.28
3 1 2 1	0.08	4 1 1 2	0.17	4 1 2 2	0.06	4 1 1 2	0.19	3 1 3 1	0.25	2 1 1 2	0.20
R13	%	R14	%	R15	%	R16	%	R17	%	R18	%
2 1 1 2	0.43	3 1 3 1	0.39	1 1 1 1	0.30	3 1 4 1	0.70	3 1 4 2	0.59	3 1 2 1	0.32
3 1 2 1	0.35	4 1 3 2	0.31	1 1 1 2	0.27	4 1 3 2	0.15	4 1 3 2	0.20	2 1 1 1	0.29
3 1 1 1	0.07	4 1 2 2	0.09	2 2 2 1	0.11	3 2 4 1	0.07	4 1 3 1	0.07	2 1 1 2	0.11

TABLE I
ROUTINES MATRIX

Figure 4 illustrates the activation probability of the extracted routines when day segments of 30 minutes are sequentially inferred. The top plot and the bottom plot show respectively the routine patterns for a COPD patient and a healthy subject. It can be seen that 3 PA routines ($R2$, $R3$ and $R15$) pervade the day of a COPD patient (top plot of Fig. 4). In particular this patient spent most of his time performing activities that involve a VL intensity descriptor of long duration, with high μ_{ST} and high μ_{VM} and a L intensity descriptor of short duration, with high μ_{ST} and low μ_{GSR}. After 5 hours from the patient awakening we can see that $R13$ becomes dominant for 1 hour. This routine includes a VL descriptor of short duration, small μ_{ST} and high μ_{VM}, and a L intensity descriptor of short duration, medium-low μ_{ST} and low μ_{GSR}. It can also be seen that around 12 hours after the awakening of the subject $R12$ starts decreasing and $R15$, characterizing the sleeping behavior, starts to become the most active routine. On the other hand the day of a healthy subject shows a more variety of active routine patters. It can be seen from the bottom plot of Fig. 4 that in the first phase of the day an alternating sequence of the routines $R13$ and $R12$ is present. $R12$ in particular is composed by descriptors related to L and MV intensities of short duration, medium-low μ_{ST} and low μ_{GSR}. Around 5 hours after his awakening and in the evening before sleeping this subject assumes a similar behavior if compared to his

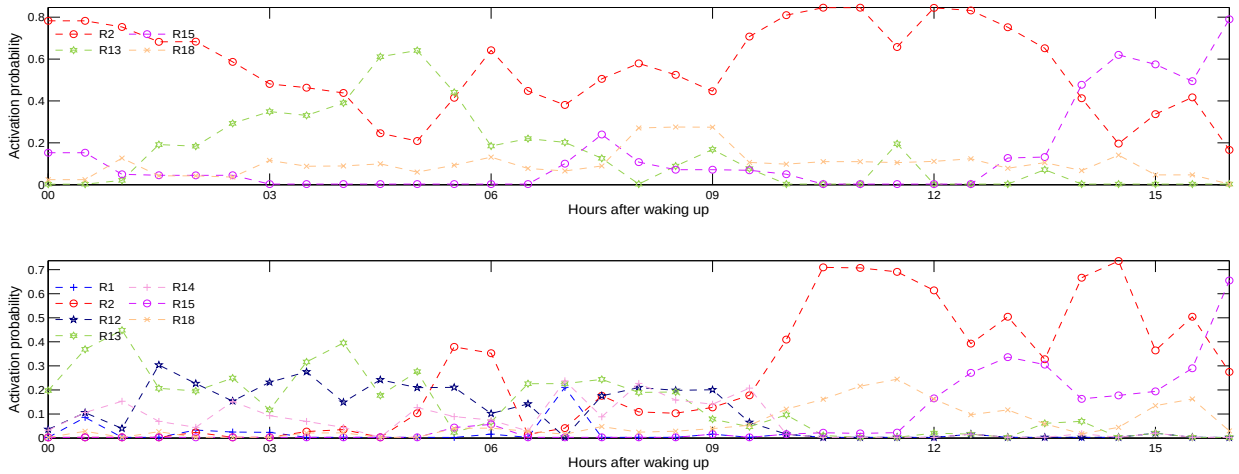
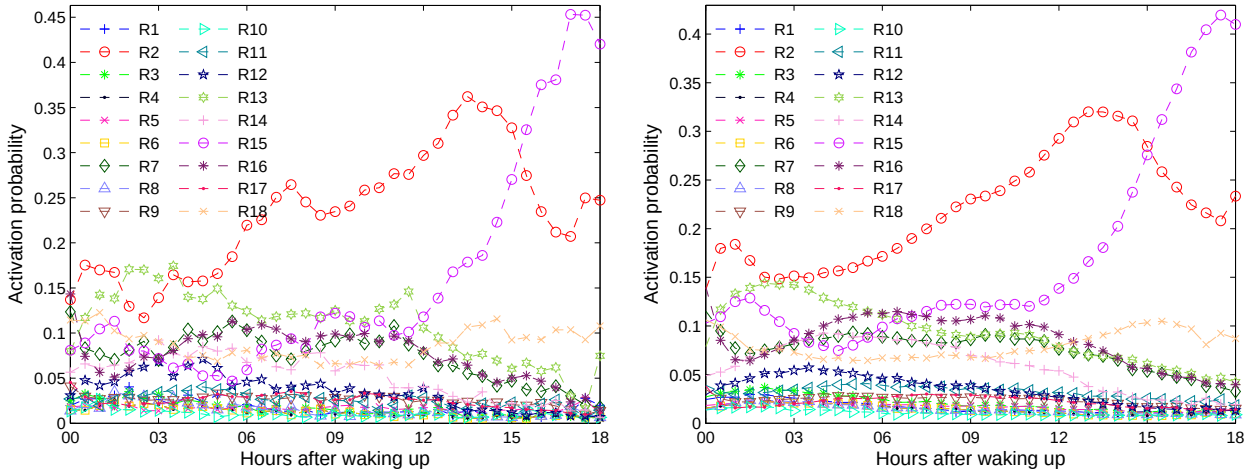


Fig. 4. Routines found inferring a COPD patient day (top) and a healthy subject (bottom). Only topics reaching an activation > 0.2 are shown.



(a) 66 COPD patients.

(b) 977 COPD patients.

Fig. 5. Routines found inferring days from 66 COPD patients (left) and a 977 COPD patients (right). The figures show the average activation probability across the patients samples.

COPD match. In particular, first routine $R12$ becomes the dominant routine for about 1 hour, and then it is active permanently until the sleeping time. The validity and robustness of the routines estimated are qualitatively highlighted in Fig. 5. Indeed, the routines discovered on a sub-sample of 66 patients were also inferred in the same way on data from 977 patients showing similar patterns. We can see that, on average, for COPD patients the PA descriptor $R2$ is almost always active across the days. This is consistent with other studies showing physical inactivity in patients with COPD [21]. Fig. 6 shows the average values of the routines' activation probabilities when the inference is performed on the first 6 hours after the subjects awakenings. Subjects were stratified according to their disease severity where GOLD1 and GOLD4 indicates respectively the least and most severe stage of COPD. Each point identifies the mean across all the subjects belonging to one of the 4 categories. Healthy subjects formed a separated category. We can observe that routines are organized according to COPD severity. In particular we can observe 4 main trends. $R2$ and $R15$ are increasing with the increase of COPD severity. In particular $R2$ represents a medium-inactive PA routine composed for the 43% by $VL\_2^d\_2^{ST}\_2^{VM}$ and for the 28% by $L\_1^d\_3^{ST}\_1^{GSR}$. The first PA descriptor represents very light intensity movements that cause a moderate increase of the temperature and that last for a long duration. The second descriptor represent light intensity movements characterized by short duration and a high physiological response (high body temperature). The positive trend is interrupted in the most severe group of patients that compensate a smaller value for $R2$ with a higher value of $R16$. This PA routine is characterized from PA descriptors including L and MV intensities and characterized by higher physiological responses if compared to $R2$. This might indicate a bigger effort in performing activities. Another positive trend is shown by $R15$ representing the time spent while performing a very inactive behavior (mainly sleeping). On the other hand we can note

that $R13$, $R12$ and $R1$ decrease with an increase in COPD severity. These 3 routines indicate movements performed with medium ($R13$) and high ($R1$ and $R12$) activity intensities characterized by small physiological responses. Of particular

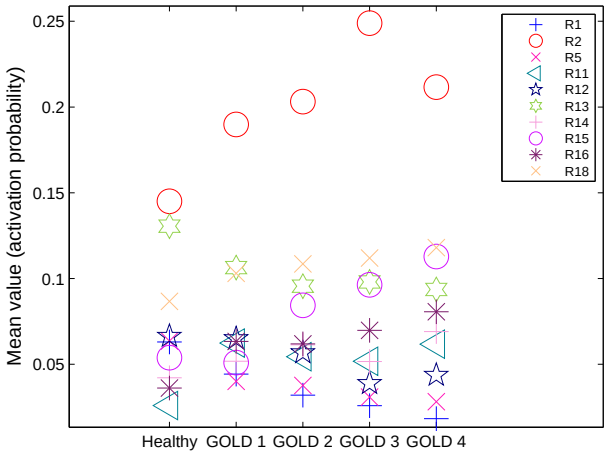


Fig. 6. Activation topic averages after stratification for COPD severity across 6h weekday. Inference was performed in the first 6 hours after the awakenings from the night sleep.

interest are $R1$ and $R13$ since they are weakly but significantly correlated with $FEV_1, \%predicted$ ($\rho = 0.2$, $p = 4e^{-10}$ and $\rho = 1.8$, $p = 1.84e^{-8}$ respectively). No correlation was found with BMI and age for $R1$. $R13$ is weakly correlated with age ($\rho = -0.18$, $p = 5.3e^{-9}$). Differently from other results in literature where PA measures were highly correlated with BMI, the routines discovered seems to be decoupled from it and, although a more detailed analysis is necessary, they seem reflecting the stage of the disease as measured by common clinical practice. In more, performing ANOVA test statistical differences have been found in the percentage of activation of $R1$ between GOLD1, GOLD3 and GOLD4 and between GOLD2 and GOLD4. Regarding $R13$ statistical differences have been found between GOLD1, GOLD3 and GOLD4, and between GOLD2, GOLD3 and GOLD4.

VI. Conclusion

Unsupervised discovery of latent structures in data from activity sensors is becoming of increased relevance due to the increasing amount of available activity data. The paper contributes to this field in different ways. First of all, real-life data is used concerning a relatively large population involving healthy and COPD patients. Secondly, the design and usage of tools differs in a number of ways from that reported so far. Using relatively simple assumptions and settings, it is shown that interpretable and consistent results can be obtained using the large set of real-life data. As such it is a encouraging step into the direction of practical applications of these techniques in daily life.

References

[1] Mark Beattie, Huiru Zheng, Chris Nugent, and Paul McCullagh. Self-management of copd: A technology driven paradigm. In *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication*, ICUIMC '14, pages 53:1–53:8, 2014.

[2] C.C. Bellos, A. Papadopoulos, R. Rosso, and D.I. Fotiadis. Identification of copd patients health status using an intelligent system in the chronious wearable platform. *Biomedical and Health Informatics, IEEE Journal of*, 18(3):731–738, May 2014.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[4] Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *16th International conference on Knowledge discovery and data mining*, 2010.

[5] C.J. Caspersen, K.E. Powell, and G.M. Christenson. Physical activity, exercise, and physical fitness: Definitions and distinctions for health-related research. *Public Health Reports*, 100(2):126–131, 1985.

[6] V. Cavalheri, L. Donaria, T. Ferreira, M. Finatti, CA. Camillo, EM. Cipulo Ramos, and F. Pitta. Energy expenditure during daily activities as measured by two motion sensors in patients with copd. *Respir. Med.*, 105:922–929, 2011.

[7] LH. Colbert, CE. Matthews, TC. Havighurst, K. Kim, and DA. Schoeller. Comparative validity of physical activity measures in older adults. *Med. Sci. Sports Exerc.*, 43:867–876, 2011.

[8] KC. Furlanetto, GW. Bisca, N. Oldemberg, TJ. Sant'anna, FK. Morakami, CA. Camillo, V. Cavalheri, NA. Hernandes, VS. Probst, and EM. Ramos. Step counting and energy expenditure estimation in patients with chronic obstructive pulmonary disease and healthy elderly: Accuracy of 2 motion sensors. *Arch. Phys. Med. Rehabil.*, 91:261–267, 2010.

[9] CE. Garber, B. Blissmer, MR. Deschenes, BA. Franklin, MJ. Lamonte, IM. Lee, DC. Nieman, and DP. Swain. American college of sports medicine position stand. quantity and quality of exercise for developing and maintaining cardiorespiratory, musculoskeletal, and neuromotor fitness in apparently healthy adults: Guidance for prescribing exercise. *Med. Sci. Sports Exerc.*, 43:1334–1359, 2011.

[10] J Garcia-Aymerich, P Lange, M Benet, P Schnohr, and J M Ant. Regular physical activity reduces hospital admission and mortality in chronic obstructive pulmonary disease: a population based cohort study. *Thorax*, 61(9):772–778, 2006.

[11] Lidwien Graat-Verboom, Ben EEM van den Borne, Frank WJM Smeenk, Martijn A Spruit, and Emiel FM Wouters. Osteoporosis in copd outpatients based on bone mineral density and vertebral fractures. *Journal of Bone and Mineral Research*, 26(3):561–568, 2011.

[12] Kylie Hill, Thomas E Dolmage, Lynda Woon, Roger Goldstein, and Dina Brooks. Measurement properties of the sensewear armband in adults with chronic obstructive pulmonary disease. *Thorax*, 65(6):486–491, 2010.

[13] Tâm Huynh, Mario Fritz, and Bernt Schiele. Discovery of activity patterns using topic models. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, UbiComp '08, pages 10–19, 2008.

[14] S. Liao, R. Benzo, A.L. Ries, and X. Soler. Physical activity monitoring in patients with chronic obstructive pulmonary disease. *J COPD F.*, 1(2):155 – 165, 2014.

[15] DC. Mackey, TM. Manini, DA. Schoeller, A. Koster, NW. Glynn, BH. Goodpaster, S. Satterfield, AB. Newman, TB. Harris, and SR. Cummings. Validation of an armband to measure daily energy expenditure in older adults. *J. Gerontol. A Biol. Sci. Med. Sci.*, 66:1108–1113, 2011.

[16] Sanjay A. Patel, Roberto P. Benzo, William A. Slivka, and Frank C. Sciurba. Activity monitoring and energy expenditure in copd patients: A validation study. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 4(2):107–112, 2007.

[17] F. Pitta, T. Troosters, M. Spruit, M. Decramer, and R. Gosselink. Activity monitoring for assessment of physical activities in daily life in patients with chronic obstructive pulmonary disease. *Archives of Physical Medicine and Rehabilitation*, 86:1979–1985, 2007.

[18] Fabio Pitta, Thierry Troosters, Vanessa S. Probst, Martijn A. Spruit, Marc Decramer, and Rik Gosselink. Physical activity and hospitalization for exacerbation of copd*. *Chest*, 129(3):536–544, 2006.

[19] Julia Seiter, Oliver Amft, Mirco Rossi, and Gerhard Troster. Discovery of activity composites using topic models: An analysis of unsupervised methods. *Pervasive and Mobile Computing*, 15(0):215 – 227, 2014.

[20] JM Seymour, MA Spruit, NS Hopkinson, A Sathyapala, WD-C Man, A Jackson, HR Gosker, AMWJ Schols, J Moxham, MI Polkey, and EFM Wouters. The prevalence of quadriceps weakness in copd and the relationship with disease severity. *AJRCCM*, 179, 2009.

[21] Thierry Troosters, Frank Sciurba, Salvatore Battaglia, Daniel Langer, Srinivas Rao Valluri, Lavinia Martino, Roberto Benzo, David Andre, Idelle Weisman, and Marc Decramer. Physical inactivity in patients with copd, a controlled multi-center pilot-study. *Respiratory Medicine*, 104(7):1005 – 1011, 2010.

[22] Ulrike von Luxburg. Clustering stability: An overview. *Foundations and Trends in Machine Learning*, 2(3):235–274, 2010.

[23] SigridNW Vorrink, HeliantheSM Kort, Thierry Troosters, and Jan-WillemJ Lammers. Level of daily physical activity in individuals with copd compared with healthy controls. *Respiratory Research*, 12(1), 2011.

[24] Benjamin Waschki, Anne Kirsten, Olaf Holz, Kai-Christian Mller, Thorsten Meyer, Henrik Watz, and Helgo Magnussen. Physical activity is the strongest predictor of all-cause mortality in patients with copd: A prospective cohort study. *Chest*, 140(2):331–342, 2011.

[25] H. Watz, B. Waschki, and T.and Magnussen H. Meyer. Physical activity in patients with copd. *Eur. Respir. J.*, 33:262–272, 2009.