and 1930 (33%) weekend days. In median, 982 min were analysed 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.

Table XI Subject group characteristics. Data are summarized as absolute frequency (n), relative frequency (%), or median and quartiles [Q1–Q3].

	All COPD n = 977	Matching healthy n = 66	Matching COPD n = 66
Male/Female (n)	636/341	30/36	30/36
Age (years)	66 [61–72]	65 [61–70]	65 [61–70]
BMI (Kg/m2)	25.8 [22.5–29.4]	25.2 [23–27.3]	25 [22.5–27.8]
FEV1 (% predicted)	49 [34–64]	107 [97–117]	42 [29–63]
GOLD 1–2–3–4 (n)	89–385–330–173	-	8–16–23–19
Assessed days (n)	5846	411	395
Weekend days (%)	33	32.6	33.6
Weekdays (%)	67	67.4	66.4
Days per subject	6 [6–6]	6 [6–6]	6 [6–6]
Minutes per subject	982 [913–1060]	965 [918–1014]	972 [919–1035]

6.4.2 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 limit the heuristics, we used a data-driven methodology to automatically create the vocabulary without specifying its size beforehand. Sensor data from the 66 healthy subjects and the matched COPD patients subsample were used to automatically create the vocabulary of words without specifying its size beforehand. Each 1-min data point consists of a 7-D measurement vector comprising: MET, ST, GSR, Longitudinal Acceleration (Acc_L), Transversal Acceleration (Acc_T), 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 (IC) using the thresholds proposed by the American College of Sports Medicine [81]: 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}), $\geq$ 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 refer to the IC only with subscripts S, VL, L, MV. Figure 38 shows an example of a one day METs data stream with the respective IC superimposed. Consecutive minutes exhibiting the same IC_s are then grouped together in 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 IBs, where each bout is fully characterized by a six-elements feature vector $\tilde{V}$

$$\tilde{V} = [IC, d, \mu ST, \mu GSR, \mu VM, \mu SC].$$

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