6.4.3 Routine discovery

For topic discovery, we used the LDA implementation of [138], 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 three closest levels found. The distances for each bout between the feature point f_j and all the levels L_p are

$$d_p(f_j) = \frac{\left|f_i - L_p^{f_j}\right|}{\sigma_p} \; \forall p = \{1, \ldots, K^{f_j}\}.$$

Once that a term of the vocabulary was assigned to each IB, documents were created by constructing for each day a histogram of terms. We chose the number of routines (R) equal to 15, and set the hyperparameter α equal to 0.01 as in [135]. Empirically we found that a number of routines greater than 15 led to duplicated routines. Hyperparameters are optimized with a variational expectation maximization algorithm initialized by randomly choosing a small number of "seed" documents [141].We selected 18 seeds (nine from healthy subjects and nine from COPD patients). Routines did not change in their overall composition with different seed sets.

6.4.4 Routine inference

Once the routines were calculated, first, we inferred day segments (i.e., in this case, a day segment is considered as the equivalent of a text document) in order to know which routines are active during different parts of the day. For this, we used sliding windows of T = 30 min of duration as suggested in [136]. From the observations (the bouts) in a sliding window, a histogram of terms has to be created as input for the topic inference. This means that the bouts in the window have to be mapped to terms from the dictionary. We did this by soft assignment as follows. For each bout described by feature vector V, the distances d_p (f_j) of each particular feature f_j to the cluster levels $L_p^{f_j}$ were determined. These distances were converted to feature weights according to

$$w(f_j,\ L_p) = \frac{e^{-d_p}}{\sum_{p=1}^{K^{f_j}} e^{-d_p}}.$$

Thus, smaller distances imply higher feature weights, and the sum of the feature weights over the different clusters equals 1. We then create the term weights by summing all the feature weights. The final normalized term weight W_t is the term weight divided by the sum of all term weights. The normalized term weights are, thus, values between 0 and 1. Normalized term weights of terms associated with other intensity categories were set to 0. Finally, we use the normalized term weights to create the histogram. For each term in the dictionary, we sum all the weights stemming from the all the bouts in the window. Second, we applied routine inference on the first 6 h of the assessed days in order to estimate the minutes spent in each routine during the most active part of the day. The same mapping procedure described for sliding windows of 30 min was applied in the case of a unique fixed window of 6 h.