correlated with the disease. The removed term is [L, $Steps_{Yes}$]. The IDF score of the words, and, subsequently, the set of removed words, are related to the wildcarding procedure previously described. If more letters are used, the words created will be more specific with the consequence of a higher IDF score average for the words in the vocabulary. On the other hand, a higher threshold on the IDF score (i.e., IDF equal to the one of the words present in 70% of documents) could cause the removal of all the terms useful to identify specific patients' subtypes. We set a threshold on the IDF equal to the one of the words present in 90% of documents as [153]. The variables for each *IC* and the final associated symbols are shown in Table XV.

Table XV Variables selected for each intensity and associated symbols.

Sleeping			Very light		
ST	GSR	SC	ST	GSR	SC
$Pf_1\ Pf_2\ Pf_3\ Pf_4$	$Pf_1\ Pf_2\ Pf_3\ Pf_4$	$*$	$Pf_1\ Pf_2\ Pf_3\ Pf_4$	$Pf_1\ Pf_2\ Pf_3\ Pf_4$	$No\ Yes$
Light			Moderate to vigorous		
ST	GSR	SC	ST	GSR	SC
$*$	$*$	$No\ Yes$	$*$	$*$	$No\ Yes$

For topic discovery, we used the LDA implementation of [138], and we considered each day of assessment as a separate document. Pre-processed data were received by a symbolization unit which maps the raw, multivariate, continuous-time data stream into a signal which can be handled by LDA. In particular it maps the data contained into each 4-elements vector $V = [IC, ST, GSR, SC]$ representing one assessed minute into a set of discrete symbols which could be interpreted as the letters of text words. Each assessed minute was mapped with an instance of the vocabulary by associating the selected values in V with their partitions and then concatenating them. Once that a term of the vocabulary was assigned to each minute, documents were created by constructing for each day a histogram of terms. We computed the results from a number of sleep modalities that varies from 3 to 20, and set the hyperparameter α equal to 0.01 as in [135]. 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. Once the routines were calculated and each assessed minute of each night mapped to a symbol of the vocabulary we inferred each night in order to estimate the minutes spent in each routine.

7.5 Classification

The evaluation of the results for classifying subjects based on their pathological condition (66 healthy vs. 66 COPD), for distinguishing among different stages of the diseases (healthy, GOLD 1,