combinations between levels sharing the same IC. For the sleeping category, for example, the feature f_1 = d, f_2 = μ_{GSR}, and f_3 = μ_{VM} were selected and divided in $K^{f1} = K^{f2} = 2$ and $K^{f3} = 3$ levels, respectively. The N^S ($N^S = K^{f1} \cdot K^{f2} \cdot K^{f3}$) terms of the vocabulary describing the sleeping intensity category ($t^S_{1\epsilon\{1:N^S\}}$) can be represented by a simplified feature vector according to

$$\begin{aligned} t_1^S &: [S, \quad L_1^d, \quad L_1^{GSR}, L_1^{VM}] \\ t_1^S &: [S, \quad L_1^d, \quad L_1^{GSR}, L_2^{VM}] \\ &\vdots \\ t_{12}^S &: [S, \quad L_2^d, \quad L_2^{GSR}, L_3^{VM}] \end{aligned}$$

Table XII Features selected for each intensity and associated levels.

Sleeping			Very light		
d	μ_{GRS}	μ_{VM}	d	μ_{ST}	μ_{VM}
$L_1\ L_2$	$L_1\ L_2$	$L_1\ L_2\ L_3$	$L_1\ L_2$	$L_1\ L_2$	$L_1\ L_2\ L_3$
Light			Moderate to vigorous		
d	μ_{ST}	μ_{GRS}	d	μ_{ST}	μ_{VM}
$L_1\ L_2$	$L_1\ L_2\ L_3\ L_4$	$L_1\ L_2$	$L_1\ L_2$	$L_1\ L_2\ L_3$	$L_1\ L_2$

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, we weighted the informativeness of the words in the vocabulary based on their inverse document frequency (IDF) score. Those terms that have a high IDF are considered more informative, because they rarely occur in the collection. In particular, we removed the words occurring in at least 90% of the documents since, occurring so frequently, they are more likely to obscure than facilitate a meaningful decomposition of the collection of documents. The term frequency (TF), usually used in combination with IDF to form the TF-IDF score [141], was not considered because it penalizes words that appear rarely within a document such as words of moderate to vigorous intensity, these words are important in the identification of routines correlated with the disease. The removed terms are [VL, L_1^d, L_1^{ST}, L_2^{MV}], [VL, L_1^d, L_2^{ST}, L_1^{MV}], and [VL, L_1^d, L_2^{ST}, L_2^{MV}]. 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. Although a rigorous methodology would be necessary to select the optimal combination of feature selected and words removed in order to generate sparse topics, we empirically found out that setting a strict threshold on the IDF (90% of documents) and performing feature selection at the level of the letters works well for our purpose.