such that any bout has a probability to belong to each of the terms created. We extract the vocabulary in a subset of COPD and healthy patients and show that the vocabulary chosen can fit the model on a larger cohort of patients. The dataset and the methodology developed are described in detail in the following sections.

6.4.1 Dataset

Data from 1001 patients suffering from mild to very severe COPD were collected across ten countries (United Kingdom, Ireland, The Netherlands, Germany, Switzerland, Italy, Spain, The United States of America, Brazil, and Australia) as part of previous studies (references can be found in the appendix) without overlaps with the current post hoc analysis. Subjects were included if they met the following inclusion criteria: COPD with a postbronchodilator forced expiratory volume in the first 1 s (FEV_1) / forced vital capacity (FVC) ratio < 0.70 and stable condition (i.e., no symptoms of increased shortness of breath and sputum production compared to usual). The dataset comprises only baseline data, which means that the COPD patients were not undergoing any specific intervention by the time of the assessment. Centres from The Netherlands and UK also provided data on 66 healthy control subjects that were matched for age, gender, and BMI with a subgroup of 66 COPD patients. On the basis of a 1:1 multivariate matching, the closest possible *case:control* matches were determined. Subjects matched exactly for age and gender, the median error between BMI values of matching subjects was 0.58 [0.29–1.2] Kg/m^2. Subject group characteristics are presented in Table XI. PAs of COPD and healthy subjects were assessed during daily life by mean of the SenseWear Armband and SenseWear Mini Armband activity monitors [140]. These devices combine an accelerometer with different physiological sensors: a heat flux sensor, a galvanic skin response (GSR) sensor, a skin temperature (ST) sensor, and a near-body ambient temperature sensor. Data are sampled in 1-min intervals and together with demographic characteristics (such as gender, age, height, and weight) were used to estimate EE and metabolic equivalent of task (MET) using proprietary algorithms developed by the manufacturer. The use of multisensory data in combination with pattern recognition algorithms ensures that the MET estimation is insensitive to noise and random motion artefacts [104]. For each minute, the associated steps count (SC) and information about the sleeping status of a subject (0 = awake, 1 = sleeping) are also provided by the sensor. The SenseWear Armband has been shown to be valid both in field [77] and in laboratory studies [79]. COPD patients and healthy subjects wore the sensor both during daytime and night-time so that continuous nonscripted activities were recorded in a natural environment. A minimum of four days (two weekdays + Saturday + Sunday) was considered acceptable to include a subject in the analysis [44], with the device being used for at least 22 h/day. From the minutes coded by the activity monitor as “sleeping,” the longest period of night sleep was extracted, and the awakening point defined as the time instant after such period. Brief awake periods (<10 min) of very light intensity (MET < 2.0) within time intervals coded as “sleeping” longer than 2 h were considered part of the sleeping time. Recorded days were synchronized according to this point in order to minimize the intrinsic variability of the data. Data prior to the awakening point were discarded from the analysis. Subjects with at least 12 h of data after the awakening point were included for a total of 977 COPD patients and 66 healthy controls. The median number of days analysed per patient was 6 (four weekdays, two weekend days), resulting in a total of 5846 valid PA days assessed, of which 3916 (67%) were weekdays