Dublin data description

Goal of study

Examining the breathing response of COPD and asthma patients to second-hand smoking in three cities in Europe: Dublin, Madrid and Liberec (Czech Republic). Subjects wore Respeck and Airspeck Personal (previous version of Airspeck Mini, in the form of a jogger's belt) for around 24h.

There were 60 subjects in total, 30 COPD and 30 asthmatics, with 10 asthmatics in Dublin, 10 asthma and 10 COPD subjects in Spain, and 10 asthma and 20 COPD subjects in Liberec.

Deployment protocol

Subjects wore the sensors for around 24h, but only as long as the sensors would last without being charged. This means that the Airspeck recording stops at around midnight.

Subjects were instructed to spend around one hour minimum in an environment where they are exposed to second-hand smoking. The start and end time of this period was recorded, as well as the number of smokers and whether it was inside or outside of a building.

STRUCTURE AND EXPLANATION OF ANALYSIS:

ANALYSIS

Correlations

- Correlation between the breathing rate and pm values (1, 2.5, 10) is calculated using Pearson's correlation coefficient. This is visualized by line plots for each patient and using heatmaps for all patients together. Different smoothing windows are applied and the size of smoothing *x* is indicated in the filename by *rollx.* This is done for all Daphne's data, then only 8 subjects of Dublin's data who were selected by plotting the signals and observing that the pm values were much higher during the indicated second-hand smoking (SHS) period. Also, for all subjects from Dublin's dataset who were Czech since I was told that the Czech data is verified and ready to use. For this last dataset correlations were found also only for sitting and standing activity types (indicated by a prefix *sitstand_*).

Daphne

- disregard 2%, take x% and 100-x%: Daphne's subjects who showed at least a very slight relationship between the breathing rate and pm1 values. The data used here was only below x percentile and above 100-x percentile of pm1 values corresponding to good (low air pollution) and bad (high air pollution) breathing, respectively.
- Scatter, daphne, breathing rate
    - *20_80_percentiles_best_2_features_meanpm_window10min.png* – the best features found using tsfresh package and having a sliding window of size 10 (10minutes) over the breathing rate. Again, the classes were indicated by taking lower than 20 percentile averages of pm2_5 values in a window to denote good breathing and over 80 – bad one. Insight: these features do not separate the classes as we want.
    - *20_80_percentiles_best_6_features_meanpm_window5min_fdr=1.png* – fdr_level is increased to 1 which gives us more features but unsensible fdr measurement since in the documentation the

fdr_level is denoted to be the ratio of false rejections and all rejections. Similarly as before, there is a window of 5min and classes are taken using percentiles and means of pm2_5 values in a window.
    - *20_80_percentiles_best_155_features_meanpm_window5min, fdr=5.png* – exactly as above but fdr_level=5. There are many more features but none can show what we expect which is red dots (bad breathing) being on the right upper corner and green ones (good breathing) on the lower left corner.
    - *best_20_features_meanpm_window10min.png* – no percentiles as above, the good and bad breathing classes are indicated by checking whether the means of pm2_5 are lower or higher than 12 and 55, respectively (intuition: https://blissair.com/what-is-pm-2-5.htm). Insight: features do not show a separation we want between classes.
    - *best_100 (or 94)_features_meanpm_window10min_200.png* – all features extracted are plotted just to check whether the built-in algorithm in tsfresh selected the features correctly. Window of 10min. None of the features show a separation here.
    - Hex plots with distributions on the size are plotted showing the relationship between the breathing rate and pm1 values.
    - *goodbad_br_rate.png* - the good and bad breathing classes are indicated by checking whether the means of pm2_5 are lower or higher than 12 and 55, respectively (intuition: https://blissair.com/what-is-pm-2-5.htm). This is done for each patient.
- Seasons
    - Plots are self-explanatory. Taking over 70% and below 30% corresponds to taking 70th and 30th percentiles as explained before.
- Violin plots, each patient, breathing rate, percentiles of 20, 80, mean pm
    - Plots of features described and plotted in *20_80_percentiles_best_6_features_meanpm_window5min_fdr=1.png* for each patient.

Dublin

- disregard 2%, take x% and 100-x%: Dublin's data subjects who showed at least a very slight relationship between the breathing rate and pm1 values. The data used here was only below x percentile and above 100-x percentile of pm1 values corresponding to good (low air pollution) and bad (high air pollution) breathing, respectively.
- Insights, comparison with cannula
    - *1-16.png* – showing how the breathing signal misbehaves and does not show any differences in the flow when there are some clear truncations in the cannula signal. The middle plot is accelerometer z axis with a Butterworth filter applied to it with low_cut=1.2 and high_cut=1.3.
    - *DBCA04.png* – accelerometer y axis with Butterworth filter applied (low_cut=1.2 and high_cut=1.3) and the breathing signal during shs and non-shs period.
    - *DBCA06.png* – the top signal is derived by taking the accelerometer y axis with Butterworth filter applied (low_cut=1.2 and high_cut=1.3), finding peaks and fitting a curve on them.
- Patients, questions
    - *axes.png* – analysis on the accelerometer signal, proves that accel z axis is the one we need (up and down movements on the abdomen).
    - *correct.png* – the 8 patients mentioned before were selected by analyzing these plot and checking whether the shs period indicated by blue and orange lines actually corresponds to high increase in pm values (red curve).
    - *during sleep.png* – shows the normal breathing signal with no movements when the person was sleeping.
    - *large window spikes.png,* spike followed by low variance.png, switches.png – the spikes were caused by switching direction/activity.
    - *missing data(…).png* – indicates too much movement which is discarded by the algorithm used to find the breathing signal.
    - *shs x.png, nonshs x.png* – shows the breathing of patient x found manually which is not influenced by movements during shs and nonshs.
    - *not steady.png* – shows the breathing signal probably influenced by other movements than only breathing.

- *pat1.png* – the timestamp does not match.
- same position, pat4, lying on the back.png, same.png – shows that even if the same activity is being performed, the breathing signal patterns are different and the accurate breathing flow cannot be measured.
- *sub11 other date.png* – recording from a different date, can be discarded; however, if matching of timestamps with the airspeck data is performed, these will be disregarded automatically.

- PCA pairplots
  - PCA was performed on manually defined features (not tsfresh) and 10principal components were extracted and plotted against each other. Prefix All_ denotes that the shs windows were taken from participant details csv file and the rest was taken to be normal breathing, prefix Intervals means taking shs and non-shs intervals from Intervals.csv which was created manually by selecting periods of time when most of the uninterrupted breathing was evident. The last number denotes the sliding window size (60-around one breath 4.8s).
- Violin plots for each feature all
  - For each patient the distributions of two classes (0:normal, 1:shs) were plotted w.r.t. each feature. Shs windows were taken from participant details csv file and the rest was taken to be normal breathing.
- Violin plots for each feature intervals
  - For each patient the distributions of two classes (0:normal, 1:shs) were plotted w.r.t. each feature. Taking shs and non-shs intervals from Intervals.csv which was created manually by selecting periods of time when most of the uninterrupted breathing was evident.
- *Each patient, just standing or sitting.docx* – some initial analysis for each patient

Medians, differences

- For each patient x and 100-x percentiles were taken and the medians were found of those which are below the xth percentile and above 100-xth percentile to denote good and bad breathing medians, respectively. Then these two were subtracted in an order given by the folder's name: good-bad means subtracting the "bad" median from the "good" one and vice versa. high(100-x)_low(x) in the filenames show numbers x and 100-x which are the lower and higher percentiles. The number following the dataset name is the number of subjects involved (Daphne: all, Dublin 8: 8 selected subjects as described before, Dublin 29: only Czech subjects). Note that in the first subplot of each figure the labels <x percentile and >100-x percentile are switched in the order, so for percentiles 30 and 70, the order of labels would be >70 percentile and <30 percentile. The second subplot shows lines for each patients connecting the medians during high air pollution and the low one, respectively for 0.0 and 1.0 in x axis. The third subplot is showing the difference of these two medians and the boxplot for all patients is plotted. On the x-axis the mean of all difference of "good" and "bad" medians is indicated.
- *Dublin 15_high70_low30_pm2_5_participantdetailsincluded.png* – shows the correct labeling and also Dublin's data subjects which showed fev1, fvc and pefr values decreasing after shs (these metrics are given in the participant details).

BREATH DETECTION

- In order to remove all the unwanted data, we want to find the periods which resemble the breathing the most. This was done by looking at the standard deviation of peaks, standard deviation of the actual values and the number of peaks separated by a particular distance. All parameters were selected manually by visualizing the plots and checking how well the breath detection is done. Note that some values from the beginning and ending of a window are disregarded since they might contain large spikes (noticed in the plots). The details can be found in *Dublin, Daphne analysis by visualizing the data, breath detection.pynb.* Equally balanced in the filenames mean that the equal number of samples in shs and non-shs periods was taken. The features are extracted and filtered using tsfresh package.

DUBLIN RAW DATA

- Self-explanatory. Includes the breathing signal and the accelerometer axes.

FEATURES WRITTEN MANUALLY

- Prefix All_ indicates that shs periods were between the timestamps indicated in the participant details file and the rest being nonshs periods while prefix Intervals means taking shs and non-shs intervals from Intervals.csv which was created manually by selecting periods of time when most of the uninterrupted breathing was evident. These features were written manually and are defined in *Feature analysis, p-values, PCA, boxplots, violin plots.ipynb*. The last number in the filenames indicates the size of a sliding window. Note that no overlap of windows was taken here. All the features are plotted in the boxplots with corresponding filenames. The deletion of outliers is based by taking small quantiles and disregarding anything that is too small or too high, the details are in the same notebook.

SPECKLED STUDENTS

- Datasets downloaded from datasync.

TSFRESH FEATURES

accel

- Features extracted from the accelerometer axes using tsfresh package and their built-in selection technique (https://tsfresh.readthedocs.io/en/latest/api/tsfresh.feature_selection.html). Butterworth filter is applied to each of them. The data is taken from the intervals defined in Intervals.csv except the ones starting with "morning". The latter means that the shs period was taken from Intervals.csv but non-shs breathing was selected to contain all the morning's data. The number in filenames indicated the size of non-overlapping sliding window. If there is a word "data" in the filename, it means that the raw windows are written to the csv, if there is "filtered_ft" these are the selected features using tsfresh and if there is "p_values" the p values are reported of each feature in increasing order.

tsfresh all data

- All data taken: shs windows from participant details file and the rest would be denoted as nonshs. The non-overlapping window size is 60 and the features extracted using tsfresh and the breathing signal.

tsfresh intervals

- Shs and non-shs periods were taken from the *Intervals.csv* file. The non-overlapping window size is 60 and the features extracted using tsfresh and the breathing signal.

The rest of the files are self-explanatory, everything is in the names and explained before.

Intervals.csv – manually defined shs and non-shs breathing periods were the breathing was most evident.
Respeck activity_types mapping.txt – shows how each activityType number maps to the actual activity.