Sleep Questionnaires in Screening for Obstructive Sleep Apnoea

Joachim A. Behar[1], Niclas Palmius[2], Jonathan Daly[2], Qiao Li[4], Fabíola G Rizzatti[3], Lia Bittencourt[3], Gari D. Clifford[4]

[1] Faculty of Biomedical Engineering, Technion, Israel Institute of Technology, Haifa, Israel
[2] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK
[3] Department of Psychobiology, Federal University of São Paulo UNIFESP, São Paulo, Brazil
[4] Departments of Biomedical Informatics & Biomedical Engineering, Emory University & Georgia Institute of Technology, Atlanta, GA, USA

Abstract

Introduction: Awareness of the high prevalence of obstructive sleep apnoea (OSA) coupled with the dramatic proportion of undiagnosed individuals has been motivating research in the past two decades for elaborating sleep questionnaires that could help in early detection of OSA. This work aims to assess the predictive value of a subset of features that are used in the STOP-BANG questionnaire (BMI, age, gender, neck size) through a rigorous statistical analysis.

Methods: A clinical database of 856 individuals referred to the sleep clinic for polysomnography was used to estimate the predictive value of the individual and combined demographic features in identifying OSA individuals. Four of the eight STOP-BANG questionnaire features were available in this database. These features were combined using a logistic regression model. The data were divided into train (80%) and test set (20%).

Results: Results on the test set were 83.3% sensitivity, 45.8% specificity, 62.2% positive predictive value, 71.7% negative predictive value. The area under the curve (AUC) had a mean of 0.717. The results showed that combining the available subset of STOP-BANG questionnaire features gave similar performance to the full STOP-BANG questionnaire as reported in a number of studies. This highlights that combining features using a machine learning framework may improve the STOP-BANG questionnaire prediction. Alternatively, it may indicate that there are redundant or uninformative questions in the STOP-BANG questionnaire.

1. Introduction

Sleep disorders are common and have been correlated with a number of cardiovascular diseases, stroke and mental health disturbances. The societal and financial costs of such disorders are consequent. In particular, one of the most common and under-diagnosed sleep disorder is obstructive sleep apnoea (OSA). The prevalence of OSA was estimated to represent 2% to 7% of the adult population globally (1; 2; 3; 4; 5; 6; 7) and a recent study even suggested that the prevalence could be over 30% of the general adult population (8). It is estimated that 90% of the individuals with the condition stay undiagnosed and untreated (9). OSA is characterised by breathing cessation (called apnoea) and periods of overly shallow breathing (called hypopnea). These are due to the partial or complete obstruction of the airway. In a recent review (10) the different signals used in the diagnosis of OSA during polysonmography (PSG) were thoroughly reviewed. The present paper focuses on the sleep questionnaires used as a tool for OSA screening.

Sleep apnoea is associated with hypertension, cerebrovascular disease, myocardial infarction, diabetes and long-term cognitive impairment (11; 12; 13; 14). In addition, a recent study showed that identification and treatment of sleep disorder breathing (SDB) in admitted patients with chronic heart failure with SDB was associated with reduced readmissions over 6 months after discharge (15).

Sleep questionnaires are often used to screen for OSA (16). The common questionnaires are the Epworth Sleepiness Scale (ESS) (17), the Berlin Questionnaire (BQ) (18) or the STOP BANG questionnaire (19). The ESS questionnaire aims to rate the likelihood of falling asleep i.e. daytime sleepiness: eight questions are scored between 0-3 making a total score out of 24. The higher the score the more daytime sleepiness. Typically, ESS < 11, ESS $\in$ [11; 14], ESS $\in$ [15; 18] and ESS > 18 are classified as normal, mild subjective daytime sleepiness, moderate subjective daytime sleepiness and severe subjective daytime sleepiness respectively (20). The correlation between ESS and OSA severity has demonstrated to be relatively weak (21). The BQ consists of a series of ten questions related to: snoring, breathing cessation, tiredness and blood pres-

 ISSN: 2325-887X DOI:10.22489/CinC.2017.233-188

sure. Ahmadi *et al.* (22) assessed the BQ on 130 patients from a sleep clinic and reported 62% sensitivity (Se) and 43% specificity (Sp) at the respiratory disturbance index (RDI 1) > 10. Their conclusion was that the BQ was not an appropriate instrument for identifying patients with sleep apnoea in a sleep clinic population. The STOP-BANG questionnaire consists of eight questions related to: snoring, tiredness, breathing cessation, blood pressure, body mass index, age, neck size and gender. Thus a series of eight Yes/No questions which total provides a score. A score greater than or equal to three means that there is a high risk of OSA and a score less than three will mean a low risk of OSA. Chung *et al.* (19) evaluated the STOP BANG questionnaire; the STOP BANG was completed by 2974 patients in the preoperative clinics of Toronto Western Hospital and Mount Sinai Hospital, Toronto, Ontario, Canada. Out of the patients that were invited, 211 patients agreed to undergo polysomnography, 34 for the pilot study and 177 for validation (69% OSA). Respective Se of 83.6%, 92.9% and 100%, Sp of 56.4%, 43% and 37%, PPV of 81.0%, 51.6%, 31.0% and NPV of 60.8%, 90.2%, 100% were found for AHI[1] greater than 5, 15, and 30 for the validation set. A recent study (23) assessed the three questionnaires on 234 patients referred to a sleep clinic. Although their dataset was unbalanced (87.1% of the population had OSA) the results confirmed the rather low specificity of these questionnaires: The STOP-BANG gave Se = 97.55%, Sp = 26.32%, PPV = 93.43%, NPV = 50%; The BQ gave: Se = 95.07% , Sp = 25%, PPV = 92.79%, NPV = 33.33%; The ESS gave: Se = 72.55%, Sp = 75%, PPV = 96.73%, NPV = 21.13%. The threshold for diagnosing OSA was set at an AHI ≥ 5. These statistics highlight the tendency of these questionnaire to 'overdiagnose' OSA. We noted that the classes (no OSA - OSA) are often unbalanced in these studies.

The sleep questionnaires used in the context of OSA screening have been assessed in a number of studies which highlighted their merit and limitations (19; 22; 23; 24). This paper aims to: 1) assess the predictive value of a number of individual features that are used in the STOP-BANG questionnaire; 2) evaluate the performance of the predictive value of the combined features. In particular, we are interested in understanding the statistical improvement that is reached by combining multiple features over a single one by using a more elaborated machine learning approach than simple thresholding over a number of yes/no answer. We use a database of 856 patients with the classes (no OSA - OSA) being balanced (48.1% - 51.9%, see Table 1) which is not the case in most previous studies. The results of our analysis are compared to the ones of other studies (19; 23; 25). The results showed a relatively high Se but low Sp, that neck size was the most predictive individual feature of OSA and that the results improved by fusing all the features available.

[1] AHI: apnea-hypopnea index. This figure corresponds to the average number of apnoeas and hypopnoeas events per hour.

Diagnosis	Number	Percentage (%)
normal	155	18.1
snorer	257	30.0
mild	106	12.4
moderate	123	14.4
severe	215	25.1

Table 1: Study database, n = 856 patients.

2. Methods

2.1. Database

For this study we used a database of 856 patients from the Respiratory Medicine Unit (Churchill Hospital, Oxford, UK). Data were recorded using the Grey Flash home polygraphy recording device (Stowood Scientific Instruments Ltd, Oxford, UK). A clinician analysed each of the recording to produce a diagnostic based on the oxygen desaturation index (ODI), sound recording, movement, heart rate, pulse transit time and respiratory effort. An extended description of the diagnosis process can be found in (26). A number of demographics were recorded from the patients including: gender, age, neck size, height and weight. More details on the data can be found in (27) and, (28). Table 1 summarises the statistics on the diagnosis of the patients included in this database. Because the end goal of the screening test is to identify patients with OSA versus non-OSA then only the following two class classification problem was considered: OSA (mild, moderate, severe) versus non-OSA (snorer, normal). In practice, a mild OSA patient may not be systematically treated. However, the role of a screening test is to identify patients at risk and although mild OSA individuals may not be treated in the first place their identification is important given that OSA can further develop and so that these individuals may make some lifestyle changes such as sleeping on the side or a diet.

2.2. Features

The demographic features that were available and used in this study are: body mass index (BMI), weight, height, age, neck size and gender. BMI, age, neck size and gender are four of the eight features used in the STOP-BANG questionnaire. Figure 1 shows the distribution obtained for each feature for the 856 patients of the database which gives some insights into what features are likely to be the most useful in classifying the population of patients. In order to train the classifier, the missing data were replaced by the average value of the corresponding missing feature

across all the training set individuals when considering multiple features. When building the model with only a single feature, the patients for which this feature was missing were excluded.

2.3. Statistics

The statistics computed in this study are sensitivity (Se), specificity (Sp), positive predictive value (PPV) and negative predictive value (NPV). In the context of the study these are defined as: Se, the percentage of individuals with OSA that have been correctly identified as OSA out of the whole OSA population; Sp, the percentage of individuals without OSA that have been identified as such out of the whole non-OSA population; PPV, the percentage of people identified as OSA and that actually are OSA; NPV, percentage of people flagged as non-OSA that are actually non-OSA.

2.4. Machine learning

The purpose of the sleep questionnaires is to screen for individuals with OSA. It must identify the highest number of individuals with OSA even to the detriment of having a higher proportion of false positive (in other words we seek a high Se). However, too many false positive (i.e. characterised by a low Sp) will overload sleep clinics with non-OSA individuals which is time consuming and costly. Typically OSA sleep questionnaires are evaluated using heuristics or simple thresholding over a number of scored answers. In this work, we use logistic regression to build our machine learning model. The model was built for each individual feature and for the combination of a subset or all the features. Data were divided into 80% for the training set and 20% for the test set. 3-fold cross validation (200 runs) with stratification was performed for each model on the training set. The receiver operating characteristic curve (ROC) was produced by averaging the statistics obtained on the validation sets of the cross validation while varying the decision threshold. The area under the ROC curve (AUC) was also computed for each individual features and the combined features. A threshold was chosen so that a Se of about 85% was obtained on the ROC curve. Given the chosen threshold, the performance of the classifier was evaluated on the test set.

3. Results

When combining the neck size, BMI, age and gender features (denoted 'All' in Table 2) an AUC of 0.717 was obtained (Table 2). Adding height and weight to the model did not increase the classifier performance. The best single-feature-AUC was obtained for neck-size (0.690). When choosing the threshold on the validation set so that a Se of about 85% was obtained, the results on the test set for all features were: Se=83.3%, Sp=45.8%, PPV=62.5% and NPV=71.7%. The results on the test set are summarised in the Table 3. The ROC curves obtained on the validation sets are plotted in Figure 2.

Feature	AUC $\pm$ STD
gender	0.590 ± 0.025
age	0.608 ± 0.026
neck size	0.690 ± 0.028
height	0.537 ± 0.028
weight	0.654 ± 0.033
BMI	0.644 ± 0.027
neck size+BMI	0.678 ± 0.026
neck size+BMI+age	$\underline{0.710 \pm 0.028}$
neck size+BMI+age+gender (All)	$\underline{0.717 \pm 0.026}$

Table 2: Area under curve (AUC) for individual and combined features with the standard deviation (STD) of the AUC obtained over the 200 runs. Since height and weight are already 'contained' in the BMI feature then these were excluded from the model denoted 'All'.

4. Discussion and conclusion

The results showed a rather high Se but a low Sp (Table 3) for the threshold selected on the training set. These results are in agreement with the statistics reported by previous studies assessing the performance of sleep questionnaires (19; 23; 25). A low specificity means that a number of individuals without OSA are predicted as having OSA which can result in a high number of unnecessary referral to the sleep clinic for PSG testing or a an increased number of perioperative PSG test for surgical patients (24). The larger variation in the statistics across studies (Table 3) can be attributed to the imbalance of the classes (see right hand side of Table 3) that was present in the other studies or to the additional features available in the STOP-BANG than in this study. For example, the higher proportion of OSA individuals in (19; 23) combined with the tendency of the STOP-BANG to over predict OSA, likely leaded to a higher PPV in their studies (81.0% and 93.4% respectively) than in the present study (PPV=62.5%) and in the study from Cruces *et al.* (25) (PPV=66.5%). These variations in the statistics and their association with the imbalance of the classes highlights the challenges in interpreting and concluding on statistical analysis conducted on unbalanced classes. In addition, it was found that the prevalence of OSA in men was higher than in women and that the prevalence was increasing with age (see Figure 1) similar as in Chung *et al.* (19).

The first main limitation is that this study utilises a database of patients from a hospital based respiratory unit, that is from a preselected population whereas a good screening questionnaire should be intended to be used as a scoring tool for the general population. Thus assessment on a database of an unselected population is necessary for

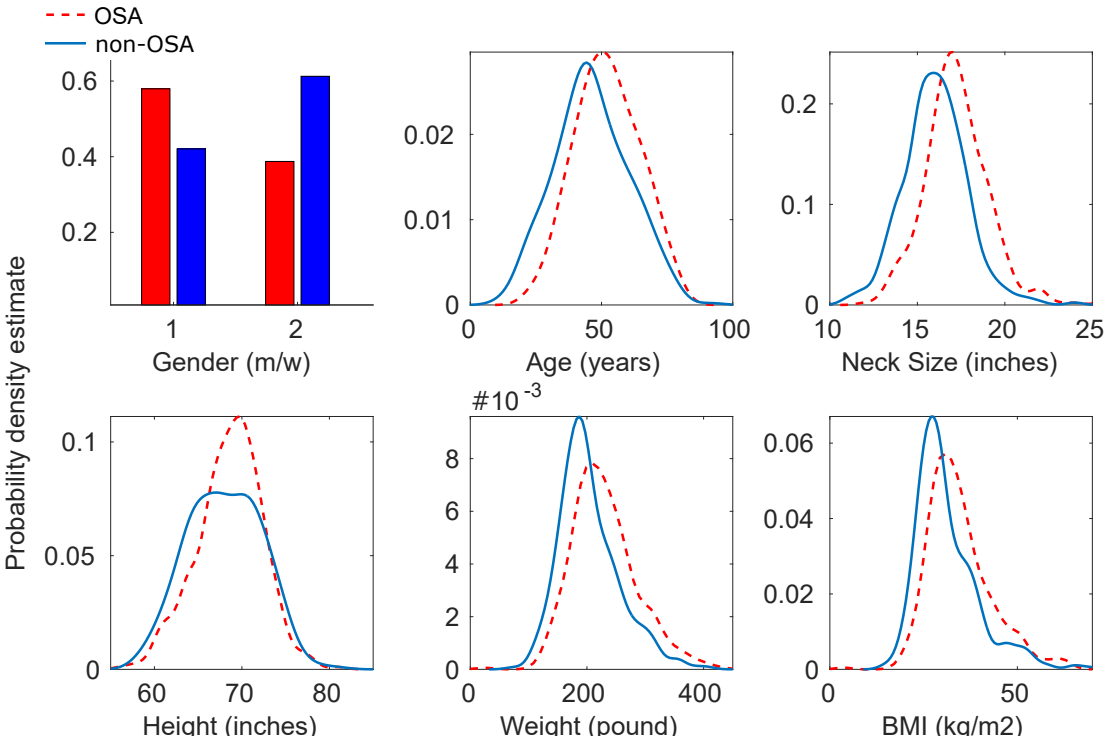

Figure 1: Distribution for the six features in the database: gender, age, neck size, height, weight and body mass index (BMI). The distributions for the two groups (OSA/non-OSA) are plotted on top of each other. For visualisation purposes, probability density estimate was used to produce these plots with exception of the gender feature. Number of patients in the two classes OSA/non-OSA for the six features: 444/412 (gender), 444/412 (age), 430/355 (neck size), 400/375 (height), 433/387 (weight) and 405/377 (BMI).

	Se (%)	Sp (%)	PPV (%)	NPV (%)	non-OSA (%)	OSA (%)
Chung et al. (19) (177)[a]	83.6	56.4	81.0	60.8	31.0	69.0
El-Sayed et al. (23) (234)[a]	97.6	26.3	93.4	50.0	12.9	87.1
Cruces et al. (25) (178)[a]	96.4	23.9	66.5	73.4	39.8	60.2
This study, all features, (171)	83.3	45.8	62.5	71.7	48.1	51.9

Table 3: Summary of the results on the test set for a threshold giving a sensitivity of about 85% on the validation set (i.e. chosen on the ROC curve of Figure 2). Results are compared with already published studies (19; 23; 25). Statistics are reported for the test set. Number of study subjects indicated in parenthesis are the one used in the test set.[a] The threshold chosen for diagnosing of OSA was set at an AHI $\geq$ 5 that is OSA versus non-OSA (two classes). The statistics on the right-hand side ('non-OSA' and 'OSA') indicates the repartition of the individuals in the two classes. All features refers to the model including: neck-size, BMI, age and gender.

a straightforward comparison.

The second main limitation of this study was the unavailability of the answers of some of the STOP-BANG questions (Snoring/Tired/Observed/Pressure). Ideally, the same rigorous statistical analysis should be conducted with the missing STOP-BANG features taken into account to assess whether the additional features add to the predictive value obtained in this study.

However, the results obtained using our machine learning framework already obtained comparable results with other studies (Table 3). This highlights that combining features using a machine learning framework may improve the STOP-BANG questionnaire prediction. Alternatively, it may indicate that there are redundant or uninformative questions in the STOP-BANG questionnaire.

Acknowledgements

JB acknowledges the support of an Aly-Kaufman Postdoctoral Fellowship. GC and QL are partially funded by the US National Institutes of Health grants R01HL136205 'Sleep Disturbance as a Mechanism for Ischemic Heart Disease in PTSD' and P50 HL117929 'Morehouse Cardiovascular Research Center of Excellence', from the National Heart, Lung and Blood Institute (NHLBI). NP and JD acknowledge the support of the RCUK Digital Economy Programme grant number EP/G036861/1 (Oxford Centre for Doctoral Training in Healthcare Innovation).

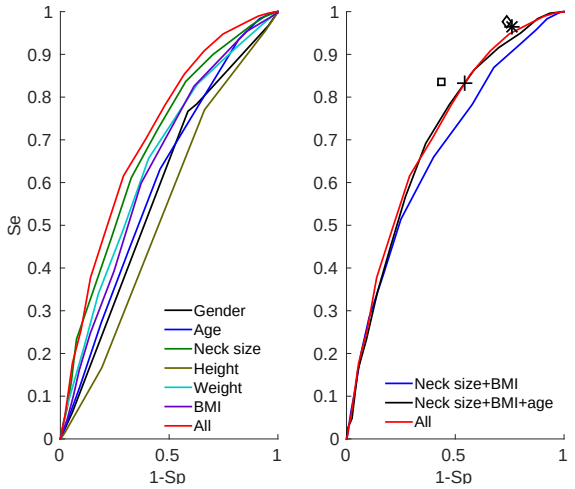

Figure 2: The receiver operating characteristic curve for all individual features, the combination of a subset of the features and for all features. 'All': neck size+BMI+age+gender. Symbols on the right figure, Cross: the statistics reported for the test set of this study; Square shape: results from (19); Asterisk: results from (25); Diamond: results from (23).

References

[1] Bearpark H, Elliott L, Grunstein R, Cullen S, Schneider H, Althaus W, Sullivan C. Snoring and sleep apnea. a population study in australian men. American journal of respiratory and critical care medicine 1995;151(5):1459–1465.

[2] Bixler EO, Vgontzas AN, Lin HM, TEN HAVE T, Rein J, Vela-Bueno A, Kales A. Prevalence of sleep-disordered breathing in women: effects of gender. American journal of respiratory and critical care medicine 2001;163(3):608–613.

[3] Ip MS, Lam B, Lauder IJ, Tsang KW, Chung Kf, Mok Yw, Lam Wk. A community study of sleep-disordered breathing in middle-aged chinese men in hong kong. CHEST Journal 2001;119(1):62–69.

[4] Lam B, Lam D, Ip M. Obstructive sleep apnoea in asia. The International Journal of Tuberculosis and Lung Disease 2007;11(1):2–11.

[5] Udwadia ZF, Doshi AV, Lonkar SG, Singh CI. Prevalence of sleep-disordered breathing and sleep apnea in middle-aged urban indian men. American journal of respiratory and critical care medicine 2004;169(2):168–173.

[6] Young T, Palta M, Dempsey J, Skatrud J, Weber S, Badr S. The occurrence of sleep-disordered breathing among middle-aged adults. New England Journal of Medicine 1993;328(17):1230–1235.

[7] Sharma SK, Kumpawat S, Banga A, Goel A. Prevalence and risk factors of obstructive sleep apnea syndrome in a population of delhi, india. CHEST Journal 2006;130(1):149–156.

[8] Tufik S, Santos-Silva R, Taddei JA, Bittencourt LRA. Obstructive sleep apnea syndrome in the sao paulo epidemiologic sleep study. Sleep medicine 2010;11(5):441–446.

[9] Young T, Evans L, Finn L, Palta M. Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women. Sleep 1997;20(9):705–706.

[10] Roebuck A, Monasterio V, Gederi E, Osipov M, Behar J, Malhotra A, Penzel T, Clifford G. A review of signals used in sleep analysis. Physiological measurement 2014;35(1):R1.

[11] Lavie P, Herer P, Hoffstein V. Obstructive sleep apnoea syndrome as a risk factor for hypertension: population study. Bmj 2000; 320(7233):479–482.

[12] Young T, Finn L, Peppard PE, Szklo-Coxe M, Austin D, Nieto FJ, Stubbs R, Hla KM. Sleep disordered breathing and mortality: eighteen-year follow-up of the wisconsin sleep cohort. Sleep 2008; 31(8):1071–1078.

[13] Punjabi NM, Newman AB, Young TB, Resnick HE, Sanders MH. Sleep-disordered breathing and cardiovascular disease: an outcome-based definition of hypopneas. American journal of respiratory and critical care medicine 2008;177(10):1150–1155.

[14] Gami AS, Olson EJ, Shen WK, Wright RS, Ballman KV, Hodge DO, Herges RM, Howard DE, Somers VK. Obstructive sleep apnea and the risk of sudden cardiac death: a longitudinal study of 10,701 adults. Journal of the American College of Cardiology 2013; 62(7):610–616.

[15] Sharma S, Mather P, Gupta A, Reeves G, Rubin S, Bonita R, Chowdhury A, Malloy R, Willes L, Whellan D. Effect of early intervention with positive airway pressure therapy for sleep disordered breathing on six-month readmission rates in hospitalized patients with heart failure. The American Journal of Cardiology 2015;.

[16] Abrishami A, Khajehdehi A, Chung F. A systematic review of screening questionnaires for obstructive sleep apnea. Canadian Journal of AnesthesiaJournal canadien danesthesie 2010; 57(5):423–438.

[17] Jones S. Stop questionnaire: A tool to screen patients for obstructive sleep apnea chung f, yegneswaran b, liao p (univ of toronto, on; toronto western hosp, univ health network, on) anesthesiology 108: 812-821, 2008. Year Book of Pulmonary Disease 2009;2009:271–272.

[18] Netzer NC, Stoohs RA, Netzer CM, Clark K, Strohl KP. Using the berlin questionnaire to identify patients at risk for the sleep apnea syndrome. Annals of internal medicine 1999;131(7):485–491.

[19] Chung F, Yegneswaran B, Liao P, Chung SA, Vairavanathan S, Islam S, Khajehdehi A, Shapiro CM. Stop questionnaire. Anesthesiology 2008;108(5):812–21.

[20] Parkes J, Chen S, Clift S, Dahlitz M, Dunn G. The clinical diagnosis of the narcoleptic syndrome. Journal of sleep research 1998; 7(1):41–52.

[21] Network Scottish Intercollegiate Guideline. Management of obstructive sleep apnoea/hypopnoea syndrome in adults. A national clinical guideline 2003;.

[22] Ahmadi N, Chung SA, Gibbs A, Shapiro CM. The berlin questionnaire for sleep apnea in a sleep clinic population: relationship to polysomnographic measurement of respiratory disturbance. Sleep and Breathing 2008;12(1):39–45.

[23] El-Sayed IH. Comparison of four sleep questionnaires for screening obstructive sleep apnea. Egyptian Journal of Chest Diseases and Tuberculosis 2012;61(4):433–441.

[24] Chung F, Abdullah HR, Liao P. Stop-bang questionnaire: A practical approach to screen for obstructive sleep apnea. CHEST Journal 2015;.

[25] Cruces-Artero C, Martin-Miguel M, Herves-Beloso C, Lago-Deibe F, Hernaiz-Valero S, Claveria-Fontan A, Guglielmi O. Validation of the stop and stop bang questionnaire in primary health care. Journal of sleep research 2012;21:226–226.

[26] Craig SE, Kohler M, Nicoll D, Bratton DJ, Nunn A, Davies R, Stradling J. Continuous positive airway pressure improves sleepiness but not calculated vascular risk in patients with minimally symptomatic obstructive sleep apnoea: the mosaic randomised controlled trial. Thorax 2012;(67):1090–1096.

[27] Behar J, Roebuck A, Shahid M, Daly J, Hallack A, Palmius N, Stradling J, Clifford GD. Sleepap: An automated obstructive sleep apnoea screening application for smartphones. In Computing in Cardiology Conference (CinC), 2013. IEEE, 2013b; 257–260.

[28] Roebuck A. A comparative analysis of polysomnographic signals for classifying obstructive sleep apnoea. PhD dissertation Univ Oxford Oxford UK 2013;.

Address for correspondence:

Joachim A. Behar, PhD
Technion-IIT, Haifa, Israel
jbehar@technion.ac.il