Journal of Biomedical Analytics
Vol. 1, No. 1 (2018), pp. 15–28
doi:10.30577/jba.2018.v1n1.2

Research Article

A Comparison of Various Data Reduction Procedures in a Multiple Sclerosis Sleep Study

Aniqa Tasnim Hossain[1], Daria Trojan[2], John Kimoff[3], and Andrea Benedetti[4,*]

[1]Department of Epidemiology, Biostatistics & Occupational Health, McGill University, Canada
[2]Department of Neurology and Neurosurgery, Montreal Neurological Institute and Hospital, McGill University Health Centre, Canada
[3]Respiratory Division and Sleep Laboratory, McGill University Health Centre, Canada
[4]Department of Epidemiology, Biostatistics & Occupational Health, McGill University; Department of Medicine, McGill University Health Centre; Respiratory Epidemiology and Clinical Research Unit, Montreal Chest Institute, McGill University, Canada
[*]*Corresponding author:* andrea.benedetti@mcgill.ca

Received: May 4, 2017; revised: August 28, 2017; accepted: October 1, 2017.

Abstract: Clinical studies often deal with data sets with numerous variables. As a result of the similarities between the variables, we frequently observe the presence of multicollinearity in the data. This study aimed to apply different data reduction strategies to sleep study variables in multiple sclerosis (MS) patients. The main objective was to use various data reduction strategies to explain a subjective measure of sleep quality (Pittsburgh Sleep Quality Index: PSQI) by the objective measures of sleep quality obtained during complete in-laboratory overnight polysomnography. Overall, we found that few objective measures of sleep quality were important in explaining the subjective PSQI, based on the results of various well accepted statistical methods. Total sleep time was found to be the most important feature of objective sleep quality for explaining subjective sleep quality among all other investigated objective sleep quality variables in most of the approaches investigated in this study. The LASSO method for estimation worked best in terms of interpretability among all the approaches considered.

Keywords: Sleep study, data reduction, LASSO, principal components, model selection

1 Introduction

In health research, we often face data sets with a huge number of variables. Some of these variables may be highly correlated or even collinear. Many of them may share similar information which is redundant in terms of describing the variability in data. In the presence of multicollinearity, variable selection becomes critical. There are many techniques of variable selection, and some are still under development. See Cox and Snell (1974), Hocking (1976, 1983), Hocking and Leslie (1967), Myers (1990) and Miller (1984) for general discussions on variable selection approaches.

A 'sleep study' also known as polysomnography (PSG) is a test used to investigate sleep patterns and to diagnose sleep disorders including obstructive sleep apnea-hypopnea (OSAH). PSG records the changes that take place during sleep and includes monitoring the electrocortical activity, eye movements, muscle activity, heart rhythm, oxygenation, breathing pattern during sleep, and leg movements (Agarwal and Gotman, 2002).

The variables derived from a PSG may be substantially correlated to each other and may cause difficulties in analyses (Silva *et al.*, 2007). Moreover, many of these variables measure related characteristics and as such, substantial correlation may exist between them. In this context, data reduction techniques may be especially attractive. We applied a number of popular variable selection techniques, such as Principal Component Analysis (PCA), Subset Regression methods and Least Absolute Shrinkage and Selection Operator (LASSO) techniques to reduce the set of variables first, and then performed further analyses to address our substantive objectives.

In this work, we explored whether there is any association between objective and subjective measures of sleep quality in subjects with multiple sclerosis (MS). MS is believed to be a chronic autoimmune demyelinating and degenerative disorder of the central nervous system. MS can lead to diverse clinical manifestations including visual loss, weakness, sensory loss, ataxia, bladder and bowel dysfunction, cognitive changes, and fatigue. These difficulties can lead to permanent disability. Four clinical courses of MS have been identified: relapsing remitting, secondary progressive, primary progressive and progressive relapsing (Compston and Coles, 2008). Many MS patients have poor sleep which can have an impact on their quality of life. It would be important to know the objective predictors of poor sleep for the clinical care of MS patients, and the best methods to perform study data analysis in MS patient populations.

We have previously reported that OSAH is common in MS patients and is strongly associated with the important symptom of fatigue (Kaminska *et al.*, 2011). In the current work we investigate: (i) different data reduction strategies in the context of a sleep study for the purpose of explaining subjective sleep quality (as measured via the Pittsburgh Sleep Quality Index, PSQI (Buysse *et al.*, 1989); and (ii) which features of objective sleep quality may account for subjective sleep quality.

2 Methods

This study was a cross-sectional study of 61 patients with multiple sclerosis (MS). A clinical assessment of MS, completion of subjective sleep and health questionnaires, and two consecutive complete overnight polysomnographies (PSGs) were conducted at the baseline evaluation of the patients. Subjects were followed for at least 3 months after the baseline

evaluation to assess if treatment for OSAH and other sleep disorders was successful. Here, we use data obtained from the baseline evaluation and first night PSG results. Further details of this study are described elsewhere (Kaminska *et al.*, 2011; Côté *et al.*, 2013). The Institutional Review Board of the participating institution approved the study. All subjects provided signed informed consent.

The baseline evaluation included a complete medical history and sleep history, physical exam, a collection of medication data, assessment of MS severity by the Expanded Disability Status Scale (EDSS, ranges between 0 and 10, greater disability is indicated by a higher score (Kurtzke, 1983), and standardized questionnaires. These included questionnaires for daytime sleepiness, subjective sleep quality via the Pittsburgh Sleep Quality Index (PSQI), Restless Legs Syndrome, narcolepsy symptoms, parasomnias including REM sleep behavior disorder (RBD), symptoms of depression (Center for Epidemiologic Studies-Depression (CES-D) questionnaire; (Radloff, 1977), total night pain, and health-related quality of life assessed with the Physical and Mental Component Summary (PCS and MCS) scores of the Short Form Health Survey (SF-36) (Ware *et al.*, 1993).

2.1 Objective Sleep Data

PSG recordings were scored manually by a certified polysomnographic technologist with expert physician review, using current American Academy of Sleep Medicine (AASM) scoring criteria except respiratory events which were scored using AASM Chicago criteria. OSAH was defined by an apnea-hypopnea index (AHI) of ≥ 15 events per hour of sleep (Kaminska *et al.*, 2011; Côté *et al.*, 2013).

From the multiple variables available from standard PSG analysis, we focused on the subset most commonly used by sleep clinicians and researchers Shrivastava *et al.* (2014): total recording time (TRT, time from lights off to lights on); total sleep time (TST), total wake time, sleep efficiency (TST/TRT), sleep latency (time to initial sleep onset), total sleep period (TSP, sleep time following sleep latency), wake time after sleep onset, number of awakenings and awakening index (per TSP), number of stage changes and stage change shift index (per TSP), number of stage 1 shifts and stage 1 shift index (per TST), percentage of time spent in stage 1, 2 and 3 and percentage of time in stage R, and micro-arousal index (per TST), with identification of micro-arousals as spontaneous, respiratory- or periodic limb movement-related. Respiratory measures included: respiratory events index (apneas, hypopneas, and respiratory-effort related arousals (RERA) per TST), respiratory event index during stage R index (apneas, hypopneas, RERA per total Stage R time), respiratory events index during stage N sleep, 4% oxygen desaturation index (4% ODI, per TST), mean SpO2 value, and minimum SpO2 value during sleep. Periodic limb movement variables were: periodic limb movement index during sleep (PLMS) and periodic limb movement associated with arousals (PLMA) index (Kaminska et al. 2012). The definitions of all sleep variables can be found at Kaminska *et al.* (2011).

2.2 Subjective Sleep Quality

Subjective sleep quality was measured via the PSQI. The PSQI generates scores corresponding to seven domains: subjective sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, use of sleep medications, and daytime dysfunction. Each component score ranges from 0 (no difficulty) to 3 (severe difficulty). The component scores are

summed to produce a global score (range of 0-21). A PSQI global score greater than five, is considered to be suggestive of clinically relevant sleep disturbance (Buysse *et al.*, 1989). We used the PSQI as a continuous variable in this study.

2.3 Statistical Analyses

We attempted several variable selection techniques, and data reduction approaches to deal with the many correlated sleep variables from the PSG. These techniques included PCA, all subset regression approaches for variable selection using different statistical criteria, LASSO method. We also used the set of sleep variables as explanatory variables for explaining our outcome (subjective sleep quality) adjusting for sex, age and BMI.

The primary objective of PCA is to reduce dimension while retaining the same amount of variability of the larger data set. PCA analysis attempts to explain the variance-covariance structure of a data set by a smaller number of uncorrelated linear combinations of the variables in the data set (Johnson and Wichern, 2007). The 'principal components' are then used as the input for further analyses.

The subset regression methods procedure fits regression equations with one candidate regressor, two candidate regressors and so on. The resulting fits are evaluated according to some popular goodness-of-fit criterion and the best regression model is selected (Montgomery *et al.*, 2012). The Mallow's C_p, adjusted R^2 (Gujarati, 2009) and Bayesian Information Criterion (BIC) (Schwarz *et al.*, 1978) criteria was used in this study (Mallows, 1973). BIC is chosen here over Akaike Information Criterion (AIC) because it is generally more conservative than AIC and avoids overfitting.

LASSO is a widely used variable selection approach in regression (Tibshirani, 1996). It minimizes the residual sum of squares while constraining the sum of the absolute value of the regression coefficients to be less than a constant. By imposing this constraint, it yields some coefficients that are exactly 0 thereby excluding those variables from the regression. LASSO is as interpretable as best subset selection (Hastie *et al.*, 2009). In this study, we used the LARS function in R. Estimates from the step of the LASSO regression with the minimum approximate C_p statistic were used.

3 Results

Table 1 describes the characteristics of our study sample. Most patients had the relapsing/remitting form of MS (75.4%) and mild EDSS (44.2%). Mean age, and BMI of the patients were 46.9 years and 26.4 kg/m^2 respectively.

Mean, median and standard deviations (SD) of the PSG variables are presented in Table 2. Amongst the sleep quality variables increasing total sleep time, total sleep period and sleep efficiency indicate better sleep. Increasing wake time after sleep onset, wake count index and spontaneous arousal index indicate increasingly disturbed sleep. Sleep is deeper as the stage increases. Just over half (55%) of subjects had an AHI greater or equal to 15 and thus were diagnosed with OSAH.

Table 1: Characteristics of our study sample.

Types of MS	**Count (Percentages)**
Relapsing/Remitting	46 (75.4%)
Secondary Progressive	13 (21.3%)
Progressive relapsing	2 (3.2%)
Types of EDSS*	**Count (Percentages)**
Mild (0 - 2.5)	27 (44.2%)
Moderate (3 - 5.5)	19 (31.1%)
Severe (5.5 - 7)	15 (24.5%)
Sex	**Count (Percentages)**
Male	17 (27.8%)
Female	44 (72.1%)
Other subject characteristics	**Mean (Standard deviation)**
Age	46.9 (10.5)
BMI	26.4 (6.1)
PSQI*	8.1 (3.9)

* Ordinal variable.

Subjects had overall poor sleep quality as reflected by the mean PSQI value (= 8). We also see this in the binary situation: most study patients (77%) had PSQI greater than 5 indicating poor sleep.

3.1 Regression Analysis of the Sleep Variables as the Explanatory Variables of the PSQI

Correlations between all the sleep, respiratory and periodic limb movement variables were estimated. As anticipated, total sleep time (minutes), total sleep period (minutes) and sleep efficiency had a high positive correlation (> 0.6) among themselves. As well, respiratory variables, stage shift variables, and arousal-related variables showed a strong positive correlation between each other. These variables also had a strong negative correlation with the sleep efficiency-related variables. The periodic limb movement variables were correlated within themselves (> 0.6), but they did not show any high correlation with other variables (near 0).

The multicollinearity in the data can cause severe problems in data analysis. To avoid this, separate regression models were fitted to explain subjective PSQI (see Table 3). Total sleep time (minutes) was the only significant variable found from the univariable regressions. After adjustment for sex, age and BMI we found that for every one minute increase in the total sleep time, the expected PSQI decreased by 0.015 units.

3.2 Principal Component Analysis of the Sleep Variables

We used PCA to reduce the data dimensionality by obtaining uncorrelated principal components to carry out further analysis. Descriptive results suggested that the first eight components were sufficient to capture about 90% of the cumulative proportion of the total variation in the data set. (See Supplementary Material Table 5).

Table 2: Descriptive measures of the sleep variables.

Sleep Variables	**Median**	**Mean**	**Standard Deviation**
Total sleep time (min)	314.0	309.8	68.5
Total sleep period(min)	389.0	384.7	50.8
Sleep efficiency	79	76.1	14.4
Wake up after sleep onset	58	74.8	50.5
Stage 1 sleep index (Total sleep time)	6.7	8.1	4.9
Stage shift index (Total sleep period)	22.0	23.2	6.5
Wake count index (TSP)	3.4	3.8	1.7
S1 shifts and wakes per TSP	9	9.9	4.7
Stage 1 percentage	8.4	10.9	7.6
Stage 2 percentage	49.5	49.4	9.7
Stage 3 percentage	23.9	25.3	11.2
Stage R percentage	14.9	14.2	6.3
Stage 0 during total recording time minutes	85.0	97.0	56.7
Spontaneous arousal index	17.1	18.2	9.0
Respiratory Variables	**Median**	**Mean**	**Standard Deviation**
Respiratory events index (apneas, hypopneas, RERA)	16.0	19.9	15.2
Respiratory events REM index (apneas, hypopneas, RERA)	18.6	22.8	19.4
Respiratory events NREM index (apneas, hypopneas, RERA)	15.4	19.0	15.2
Desaturation events sleep only index	0.5	3.1	6.8
SpO2 average value	95.7	95.34	1.6
SpO2 minimum value	88.9	87.7	6.6
PLM Measures	**Median**	**Mean**	**Standard Deviation**
PLM total asleep index	6.9	21.6	33.4
PLMA index	1.6	3.5	5.1

The results of the multiple regression analysis of the continuous PSQI on the PC scores before and after adjustment for age, sex and BMI are given in Table 4. Regression coefficients and their corresponding confidence intervals are reported. None of the principal components were statistically significant in estimating the continuous PSQI.

3.3 All Subset Regression Approaches

This study used the best subsets approach using Mallow's C_p, adjusted R^2, and BIC as goodness of fit criteria. 2^{22} number of possible models were fitted. The best model found using the lowest C_p (-4.9) among all the fitted models included total sleep time (minute), wake count index (TSP) and average SpO2. The best model using the highest adjusted R^2 (0.19) among all the fitted models included sleep efficiency, wake count index (total sleep period), spontaneous arousal index, respiratory events index (apneas, hypopneas, RERA), respiratory events NREM index (apneas, hypopneas, RERA), average SpO2, PLM total sleep index, PLMA index and wake up after sleep onset. Finally, the best model selected using the lowest BIC 3.6 among all the fitted models included only the total sleep time (minutes).

3.4 Least Absolute Shrinkage and Selection Operator (LASSO)

The first step had the lowest C_p (1.96). This step included only one non-zero coefficient for total sleep time (minutes). The coefficient was -0.004 which is very small. This model had an adjusted R^2 of 0.028 indicating a poor fit.

Table 3: Estimating continuous PSQI by selected covariates by regression analyses.

Sleep Variables	**Co-efficient**	**95% CI**	**Co-efficient (adj.)**	**95% CI (adj.)**
Total sleep time (min)	-0.02*	[-0.03 -0.001]	-0.01*	[-0.03 -0.001]
Total sleep period (min)	-0.01	[-0.031 0.007]	-0.01	[-0.03 0.008]
Sleep efficiency	-0.06	[-0.13 0.006]	-0.06	[-0.13 0.007]
Wake up after sleep onset	0.02	[-0.003 0.035]	0.02	[-0.004 0.037]
Stage 1 sleep index (total sleep time)	0.15	[-0.05 0.35]	0.12	[-0.10 0.346]
Stage shift index (total sleep period)	0.09	[-0.06 0.24]	0.08	[-0.08 0.23]
Wake count index (TSP)	0.43	[-0.13 1.00]	0.37	[-0.22 0.96]
S1 shifts and wakes per TSP	0.16	[0.05 0.36]	0.12	[-0.10 0.35]
Stage 1 percentage	0.11	[-0.01 0.24]	0.10	[-0.05 0.25]
Stage 2 percentage	0.03	[-0.07 0.13]	0.02	[-0.09 0.12]
Stage 3 percentage	-0.07	[-0.16 0.01]	-0.06	[-0.17 0.04]
Stage R percentage	0.00	[-0.16 0.15]	0.00	[-0.16 0.16]
Stage 0 during total recording time minutes	0.01	[-0.003 0.03]	0.01	[-0.003 0.032]
Spontaneous arousal index	-0.04	[-0.153 0.071]	-0.02	[-0.14 0.11]
Respiratory Variables	**Co-efficient**	**95% CI**	**Co-efficient (adj.)**	**95% CI (adj.)**
Respiratory events index (apneas, hypopneas, RERA)	0.01	[-0.053 0.079]	0.006	[-0.07 0.073]
Respiratory events REM index (apneas, hypopneas, RERA)	0.02	[-0.031 0.071]	0.01	[-0.04 0.071]
Respiratory events NREM index (apneas, hypopneas, RERA)	0.01	[-0.054 0.078]	-0.001913	[-0.07 0.07]
Desaturation events sleep only index	-0.02	[-0.177 0.12]	-0.054618	[-0.21 0.10]
SpO2 average value	-0.04	[-0.65 0.57]	0.120	[-0.58 0.83]
SpO2 minimum value	0.07	[-0.07 0.22]	0.11	[-0.046 0.27]
PLM Measures	**Co-efficient**	**95% CI**	**Co-efficient (adj.)**	**95% CI (adj.)**
PLM total asleep index	0.01	[-0.02 0.040]	0.01052	[-0.02 0.042]
PLMA index	0.13	[-0.06 0.33]	0.12	[-0.07 0.33]

Models were adjusted for age, sex, and BMI. Values with * were statistically significant at 0.05.

Table 4: Multiple regression of the continuous PSQI on the PC scores.

	Estimate	Conf.int	Estimate (adj)	Conf.int
Score 1	0.27	[-0.09 0.63]	0.23	[-0.21 0.67]
Score 2	0.35	[-0.14 0.84]	0.29	[-0.14 0.92]
Score 3	0.17	[-0.56 0.92]	0.12	[-0.74 0.98]
Score 4	-0.05	[-0.81 0.70]	-0.07	[-0.87 0.74]
Score 5	0.13	[-0.65 0.93]	0.15	[-0.68 0.99]
Score 6	-0.04	[-0.91 0.83]	0.08	[-0.94 1.09]
Score 7	-0.78	[-1.77 0.19]	-0.81	[-1.85 0.23]
Score 8	-0.90	[-1.96 0.17]	-0.84	[-2.00 0.33]

3.5 Summary

To compare the results from all methods utilized, we summarized our findings for explaining the subjective PSQI by the objective sleep measures reporting important variables. These variables are obtained in the regression models using the objective sleep variables and other selected variables with different data reduction techniques. Linear regression analyses indicated that the only important predictor of sleep quality was total sleep time. LASSO and best subsets regression using the BIC as the criterion provided the same result. Best subsets regression using Mallow's C_p, and the adjusted R^2 both suggested several additional variables.

Important variables obtained in different data reduction strategies

1. Linear Regression (adjusted for sex, age and BMI) (Continuous): Total sleep time (minutes)
2. Mallow's C_p: total sleep time (minute), Wake count index (TSP), SpO2 average value
3. Adjusted R^2: sleep efficiency, wake count index (total sleep period), spontaneous arousal index, respiratory events index (apneas, hypopneas, RERA), respiratory events NREM index (apneas, hypopneas, RERA), SpO2 average value, PLM total sleep index, PLMA index, wake up after sleep
4. BIC: total sleep time (minutes)
5. LASSO: total sleep time (minutes)

4 Discussion

We compared different data reduction strategies in the context of a sleep study in MS patients. We investigated whether and which features of objective sleep quality assessed by overnight PSG explained the Pittsburgh Sleep Quality Index (PSQI) score which is a subjective measure of sleep quality. We applied data reduction techniques to a set of clinically relevant sleep quality, respiratory, and periodic limb movement variables measured during PSG, as well as other relevant characteristics of the subjects (age, sex and BMI). Very few of these variables were found to be important in explaining subjective PSQI.

Data reduction techniques have been used in this context before. Several studies have used principal components analysis to reduce data dimensionality in sleep apnea studies (Tangugsorn *et al.*, 1999). A sleep study (Myers and Downs, 2009) had been conducted using LASSO to select the minimum predictors' subset to predict cognitive fatigue levels. In

all these studies, only one data reduction technique was used. We aimed to compare the performance of several data reduction techniques.

In a set of highly correlated variables it is possible to retain the same amount of information from fewer numbers of variables. In order to obtain a smaller set of variables in the high dimensional data, various data reduction strategies have been discussed in this study. Our data reduction strategies included univariable regressions of the sleep variables adjusted for the most important confounders, PCA, all subset regression approaches for variable selection (based on Mallow's C_p, Adjusted R^2 and BIC) and LASSO. LASSO and all subsets regression using BIC as the criterion found only one important predictor (total sleep time). PCA found no important predictors. All subsets regression using Mallow's C_p or the adjusted R^2 suggested several variables (including total sleep time) were important predictors of subjective sleep quality.

Each method has its own strengths and limitations. We found eight principal components contained as much information as the full data set. However, none were found as important predictors of subjective sleep quality. Moreover, interpretation of the principal components was difficult. The loadings were not very informative so that we could not interpret the components according to some specific characteristics of the original variables.

The all subset regression method and LASSO overcame this interpretability issue, as these methods directly indicated the most important variables for explaining subjective PSQI. Using different criteria with all subset regression resulted in different covariates being chosen as predictors for PSQI. For this data, the LASSO method was the best data reduction strategy. It forced some coefficients of the variables to be exactly zero, eliminated them as predictors, and the results were interpretable.

The main limitation of this work is the small number of observations relative to the total number of variables considered. Because of the small sample size, the study results may be limited to the sample under study and may not be generalized to the population. Moreover, some of the variables may have been associated with high variability due to the small sample. Particularly for those variables associated with smaller coefficient values, the high variability might have caused the non-significance in the regression analyses. Also, this study does not adhere to “event per variable" recommendations to avoid overfitting. In addition, the rationale for this study was only around the correlation among the sleep variables which could be further assessed using more rigorous multicollinearity testing tools.

On the other hand, the strength of the work presented here is that we used various statistical tools to fulfill our objectives. We explored the data set in several different ways and found largely similar results in all cases, that is, few objective measures of sleep quality were important in explaining the PSQI. This may be because MS patients may have poor judgment and memory loss, i.e. their subjective perceptions of sleep can be unreliable. In addition, there may be other factors that can contribute to poor sleep which were not considered in this study such as pain, spasticity, anxiety and depression, and environmental variables. Alternatively, PSQI assesses sleep quality during the past month while the PSG is measured on one specific night. Also, PSQI may be most useful as a dichotomous rather than categorical variable. We considered only global PSQI scores which include some self-reported questions on habitual total sleep time. This may account for the association we found between total sleep time and PSQI.

5 Conclusion

In the current work, the LASSO method was the easiest to implement; number of variables could be reduced and the results were easily interpretable. Future work could include simulation to understand the reliability of the results, potentially with a reasonably larger sample size.

The conclusion of this study is similar to the recommendations provided in other epidemiologic studies (Greenland, 2008; Walter and Tiemeier, 2009). Based on our study findings and the ease of interpretation, we recommend using the LASSO method in future studies of the relationship of PSG parameters and clinical symptoms of poor sleep quality.

Declarations

Acknowledgment

In this work, AH defined the analysis plan, carried out the statistical analyses, and drafted the manuscript. AB helped to define the analysis plan, interpreted the results and edited and revised the manuscript. JK and DT conceived of the original study, acquired the data, helped to interpret the results, and edited the manuscript. We are grateful to the patients who participated in the study. All authors have read and approved the final version.

Funding

This work was supported by an operating grant from the Multiple Sclerosis Society of Canada.

Conflict of Interest

The authors declare that they have no competing interests.

Ethical Approval

This study was approved by the Montreal Neurological Institute and Hospital Research Ethics Board.

Abbreviations Used

Abbreviation	Full Name
AASM	American Academy Of Sleep Medicine
BIC	Bayesian Information Criterion
BMI	Body Mass Index
CES-D	Center For Epidemiologic Studies-Depression
EDSS	Expanded Disability Status Scale
LASSO	Least Absolute Shrinkage And Selection Operator
MCS	Mental Component Summary
MS	Multiple Sclerosis
4% ODI	4% Oxygen Desaturation Index
OSAH	Obstructive Sleep Apnea-Hypopnea
PCA	Principal Component Analysis
PCS	Physical Component Summary
PLMA	Periodic Limb Movement Associated With Arousals
PLMS	Periodic Limb Movement Index During Sleep
PSG	Polysomnography
PSQI	Pittsburgh Sleep Quality Index
RBD	REM Behavior Disorder
REM	Rapid Eye Movement
RERA	Respiratory-Effort Related Arousals
SF-36	Short Form (36) Health Survey
SSE	Residual Sum Of Squares
TRT	Total Recording Time
TSP	Total Sleep Period
TST	Total Sleep Time

References

Agarwal R, Gotman J (2002). "Digital tools in polysomnography." *Journal of Clinical Neurophysiology*, **19**(2), 136–143. doi:10.1097/00004691-200203000-00004.

Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ (1989). "The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research." *Psychiatry Research*, **28**(2), 193–213. doi:10.1016/0165-1781(89)90047-4.

Compston A, Coles A (2008). "Multiple sclerosis." *The Lancet*, **372**(9648), 1502–1517. doi:10.1016/S0140-6736(08)61620-7.

Côté I, Trojan D, Kaminska M, Cardoso M, Benedetti A, Weiss D, Robinson A, Bar-Or A, Lapierre Y, Kimoff R (2013). "Impact of sleep disorder treatment on fatigue in multiple sclerosis." *Multiple Sclerosis Journal*, **19**(4), 480–489. doi:10.1177/1352458512455958.

Cox DR, Snell EJ (1974). "The choice of variables in observational studies." *Applied Statistics*, pp. 51–59. doi:10.2307/2347053.

Greenland S (2008). "Invited commentary: variable selection versus shrinkage in the control of multiple confounders." *American Journal of Epidemiology*, **167**(5), 523–529. doi:10.1093/aje/kwm355.

Gujarati DN (2009). *Basic econometrics*. Tata McGraw-Hill Education.

Hastie T, Tibshirani R, Friedman J (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer.

Hocking R, Leslie R (1967). "Selection of the best subset in regression analysis." *Technometrics*, **9**(4), 531–540. doi:10.2307/1266192.

Hocking RR (1976). "A Biometrics invited paper. The analysis and selection of variables in linear regression." *Biometrics*, **32**(1), 1–49. doi:10.2307/2529336.

Hocking RR (1983). "Developments in Linear Regression Methodology: 1959–l982." *Technometrics*, **25**(3), 219–230. doi:10.1080/00401706.1983.10487871.

Johnson A, Wichern W (2007). *Applied multivariate statistical analysis*. Prentice Hall.

Kaminska M, Kimoff R, Schwartzman K, Trojan D (2011). "Sleep disorders and fatigue in multiple sclerosis: evidence for association and interaction." *Journal of the Neurological Sciences*, **302**(1), 7–13. doi:10.1016/j.jns.2010.12.008.

Kurtzke JF (1983). "Rating neurologic impairment in multiple sclerosis an expanded disability status scale (EDSS)." *Neurology*, **33**(11), 1444–1452. doi:10.1212/wnl.33.11.1444.

Mallows CL (1973). "Some comments on C_p." *Technometrics*, **15**(4), 661–675. doi:10.1080/00401706.1973.10489103.

Miller AJ (1984). "Selection of subsets of regression variables." *Journal of the Royal Statistical Society. Series A (General)*, pp. 389–425. doi:10.2307/2981576.

Montgomery DC, Peck EA, Vining GG (2012). *Introduction to linear regression analysis*, volume 821. John Wiley & Sons.

Myers LJ, Downs JH (2009). "Parsimonious identification of physiological indices for monitoring cognitive fatigue." In *International Conference on Foundations of Augmented Cognition*, pp. 495–503. Springer. doi:10.1007/978-3-642-02812-0_58.

Myers RHRH (1990). *Classical and modern regression with applications*. PWS-Kent.

Radloff LS (1977). "The CES-D scale: A self-report depression scale for research in the general population." *Applied Psychological Measurement*, **1**(3), 385–401. doi:10.1177/014662167700100306.

Schwarz G, *et al.* (1978). "Estimating the dimension of a model." *The Annals of Statistics*, **6**(2), 461–464. doi:10.1214/aos/1176344136.

Shrivastava D, Jung S, Saadat M, Sirohi R, Crewson K (2014). "How to interpret the results of a sleep study." *Journal of Community Hospital Internal Medicine Perspectives*, **4**(5), 24983. doi:10.3402/jchimp.v4.24983.

Silva GE, Goodwin JL, Sherrill DL, Arnold JL, Bootzin RR, Smith T, Walsleben JA, Baldwin CM, Quan SF (2007). "Relationship between reported and measured sleep times: the sleep heart health study (SHHS)." *Journal of Clinical Sleep Medicine: JCSM: Official Publication of the American Academy of Sleep Medicine*, **3**(6), 622.

Tangugsorn V, Krogstad O, Espeland L, Lyberg T (1999). "Obstructive sleep apnea: a principal component analysis." *The International Journal of Adult Orthodontics and Orthognathic Surgery*, **14**(3), 215–228.

Tibshirani R (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Walter S, Tiemeier H (2009). "Variable selection: current practice in epidemiological studies." *European Journal of Epidemiology*, **24**(12), 733. doi:10.1007/s10654-009-9411-2.

Ware J, Snow K, Kosinski M, Gandek B (1993). *SF-36 Health Survey: Manual and Interpretation Guide*. Nimrod Press, Boston, MA.

A Supplementary Materials

Table 5: Principal Component Analysis of the sleep variables.

Components	Standard Deviation	Proportion of Variation	Cumulative Proportion
1	2.75	0.34	0.34
2	2.00	0.18	0.53
3	1.33	0.08	0.61
4	1.30	0.08	0.68
5	1.24	0.07	0.75
6	1.13	0.06	0.81
7	1.00	0.05	0.86
8	0.92	0.04	0.90
9	0.81	0.03	0.93
10	0.69	0.02	0.95
11	0.58	0.02	0.96
12	0.49	0.01	0.97
13	0.42	0.01	0.98
14	0.40	0.01	0.99
15	0.37	0.01	1.00
16	0.25	0.00	1.00
17	0.17	0.00	1.00
18	0.09	0.00	1.00
19	0.07	0.00	1.00
20	0.04	0.00	1.00
21	0.01	0.00	1.00
22	0.00	0.00	1.00

A.1 Mallow's C_p

Mallow's C_p is a statistic used to determine the mean squared prediction error of a fitted value. If we select P regressors from a set of $K > P$, the C_p statistic for that particular set of regressors is defined as: $C_p = \frac{SSE_p}{S^2} - N + 2P$ where SSE_p is the residual sum of squares for the model with P regressors, S^2 is the residual mean square after regression on the complete set of K regressors, and N is the sample size.

A.2 Adjusted R^2

Adjusted R^2 adjusts for the number of explanatory terms in a model. It gives the proportion of variation explained by the regression. The adjusted R^2 increases only if the new term improves the model more than would be expected by chance. It is used to adjust for the overfitting of the model. Adjusted R^2 is a better goodness of fit than the usual R^2 statistic when we keep adding variables in the model since adding variables will always increase the usual R^2. Adjusted R^2 is defined as

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} = R^2 - (1 - R^2)\frac{p}{n-p-1},$$

where p is the total number of regressors in the linear model (without the constant term), and n is the sample size.

A.3 Bayesian Information Criterion

The Bayesian information criterion (BIC) or Schwartz Bayesian Criterion (SBC) is a model selection tool. While comparing different models we calculate for each model,

$$BIC = n\ln(SSE) - n\ln(n) + \ln(n)p,$$

where SSE is the residual sum of squares of that model, p is the number of parameters in the model, and n is the sample size. Finally we call the one the best fitted model with the smallest BIC.