Sleep Breath
DOI 10.1007/s11325-008-0214-6

ORIGINAL ARTICLE

Assessment of the test–retest reliability of laboratory polysomnography

Daniel J. Levendowski · Nadene Zack · Srini Rao · Keith Wong · Michael Gendreau · Jay Kranzler · Timothy Zavora · Philip R. Westbrook

Received: 15 February 2008 / Revised: 15 July 2008 / Accepted: 28 July 2008

Abstract
Statement of the problem When conducting a treatment intervention study, it is assumed that a level of reliability can be obtained from the measurement tool such that the outcome can be reasonably assessed.
Purpose of study Investigate the reliability of laboratory polysomnography, the gold standard for assessment of treatment outcomes for obstructive sleep apnea, at a 1-month interval.
Materials and Methods In a clinical trial of 118 patients recruited to assess the effects of a pharmaceutical treatment intervention, a subset of 20 patients designated as placebo controls completed two polysomnography studies, one at baseline and one at least one month later.
Results The correlation between the overall Apnea/Hypopnea indices from the two polysomnography (PSG) studies was poor (r=0.44) and the results were biased, with a mean increase of seven events per hour on night 2. Twenty-five percent of the subjects had an increase greater than 20 events/hour on night 2 and only 45% of participants had a night-to-night difference of ≤5 events/hour. The correlation between overall apnea indexes for nights 1 and 2 (r=0.61) was improved, compared to the overall apnea/hypopnea indexes. The correlation in sleep efficiency across the two nights was relatively week (r=0.52) but significant. The correlations between nights 1 and 2 for the percentage of time supine (r=0.70) and the supine apnea–hypopnea index (AHI) (r=0.69) were similar and highly significant. The correlation for the non-supine AHI was only 0.25
Conclusions In this study, the reliability of a single-night PSG in measuring treatment outcome was compromised as a result of the large night-to-night variability of subjects' obstructive sleep apnea (OSA). Studies employing the AHI as an outcome need to be adequately powered with respect to the inherent night-to-night variability in the measurement. When assessing treatment intervention outcomes, there may be benefit from the acquisition and averaging of multiple nights of data in order to mitigate the inherent night-to-night variability of OSA and improve the accuracy of the outcome assessment.

Keywords Polysomnography · Repeated measures · Reliability · Treatment outcome · Case finding

D. J. Levendowski (✉) · T. Zavora · P. R. Westbrook
Advanced Brain Monitoring, Inc.,
2237 Faraday Avenue, Suite 100,
Carlsbad, CA 92008, USA
e-mail: Dan@b-alert.com

N. Zack · S. Rao · M. Gendreau · J. Kranzler
Cypress Biosciences, Inc.,
4350 Executive Drive, Suite 325,
San Diego, CA 92121, USA

K. Wong
Woolcock Institute of Medical Research, University of Sydney,
Sydney, Australia

Introduction

Laboratory polysomnography (PSG) is considered the "gold standard" for diagnosing [1] and assessing the benefit of a treatment intervention [2] for obstructive sleep apnea (OSA), a highly prevalent disease [3–5]. Practice parameters adopted by the American Academy of Sleep Medicine (AASM) for the assessment of treatment outcomes with mandibular repositioning devices indicate that PSG should be exclusively used [2], even though previous reports have raised questions as to the test-retest reliability of PSG.

Factors that may contribute to the variability in PSG results include intra- and inter-rater reliability in scoring [6–8], night-to-night variability in the percentage of the time supine [9–11], sleep latency/efficiency [12, 13] and/or consumption of alcohol [14]. Possibly due to the cost of PSG, there have been few published studies investigating the inter-trial differences in the PSG results on symptomatic patients.

The goal of this report is to evaluate the reliability of PSG in a small number of placebo control subjects included in a study to assess the effects of a pharmaceutical intervention for the treatment of OSA.

Materials and methods

One hundred and eighteen patients suspected of having OSA and referred to one of three sleep laboratories were recruited to assess the effects of mirtazapine alone or taken with compound CD0012 or placebo as a treatment intervention for OSA. Inclusion criteria for this study were apnea-hypopnea index (AHI) between 10 and 40, age ≥21, body mass index (BMI) ≤40, not currently being treated for OSA with either Continuous Positive Airway Pressure (CPAP) or mandibular advancement device, non-smoker, baseline Epworth >10, and if female, post-menopausal.

A subset of 20 patients (17 males and three females) were designated as placebo controls, and completed two PSG studies, one at baseline (N1) and one at least one month later (N2) (inter-study interval mean 40±12 S.D. days).

The PSG studies were conducted at one of three sleep centers in Sydney, Australia (Royal Prince Alfred Hospital, Royal North Shore Hospital, and St. Vincent's Hospital). Each patient underwent both PSG studies at the same sleep center using the same equipment. The PSG was performed using the standard 18-channel montage. Airflow was measured with nasal pressure. Chest and abdominal respiratory effort was measured with chest and abdominal inductance plethysmography. Body position was measured by the sensors provided with the respective PSG equipment (i.e., Alice 4-Respironics Inc. and Compumedics E-series) used in each lab.

The PSG recordings from the study sites were scored at a central scoring facility, with the same technician scoring both studies from each patient in the order of acquisition. The technician was blinded as to which patients were assigned as placebo controls. The apnea index was defined

Table 1 Repeated measure results from two PSG studies administered at a 1-month interval

Subject number	Days between studies	Overall AHI		Apnea index		AHI supine		AHI Non-supine		% Time supine		Sleep efficiency (%)	
		N1	N2	N1	N2	N1	N2	N1	N2	N1	N2	N1	N2
11002	32	37	33	21	30	72	54	21	7	30	56	63	75
11007	34	23	24	10	6	89	40	13	16	14	32	87	84
11019	50	38	28	11	5	69	55	6	8	51	42	86	86
11025	41	33	35	5	12	10	58	44	26	31	29	79	87
11034	46	14	9	1	0	16	25	10	8	67	15	87	93
11038	34	43	35	3	13	40	40	45	39	43	39	73	83
12005	52	14	39	0	1	15	41	0	1	96	89	81	93
12006	42	29	83	0	19	71	90	9	79	30	11	64	80
13008	83	33	34	8	28	75	55	10	8	35	55	85	90
22001	35	28	33	7	1	35	39	21	22	52	65	83	85
22005	34	17	12	0	0	17	12	0	0	100	100	80	69
22008	37	40	74	24	50	114	120	37	74	3	0	69	75
22017	33	17	39	0	0	20	48	10	30	69	51	84	61
22014	35	18	21	5	9	32	39	3	4	52	48	92	90
22012	38	27	38	3	15	27	33	0	83	100	90	84	91
30006	29	32	53	6	21	48	54	13	34	53	95	73	76
32005	44	29	31	10	4	45	17	15	18	45	23	95	92
34005	29	19	13	1	0	19	14	18	7	63	88	96	93
34013	35	39	31	11	7	42	46	12	12	84	48	76	74
Mean, n=20	40	28	34	7	11	46	48	15	24	53	50	81	83
Std. Dev.	11.9	9.2	18.7	6.8	13.1	28.6	25.5	13.4	26.2	27.4	30.0	9.2	8.9
Correlation		0.44		0.61		0.69		0.25		0.70		0.52	
Sign. p		N.S.		<0.01		<0.001		NS		<0.001		<0.05	

as a 10-s cessation in airflow (measured by nasal pressure). A hypopnea was defined according to the American Academy of Sleep Medicine pre-2007 criteria requiring a minimum of a 10-s event with either: (a) a 50% reduction in airflow, (b) a clear amplitude reduction in airflow and a ≥3% reduction in SpO2, or (c) >30% reduction in airflow and a cortical arousal. Patients were allowed to consume their usual daily amount of alcohol prior to lights out.

Pearson correlation analysis and *t* tests were used to assess the night-to-night variability in the sleep efficiency, percentage of time supine, overall and positional AHI, and the apnea index. Bland–Altman plots were used to display the variances in the AHI. Student's *t* tests were used to identify significant differences between nights 1 and 2 in the total sleep time, sleep efficiency, minutes non-rapid eye movement (REM) and REM, percent time REM, sleep latency, REM latency, and minutes stage 1, 2 and slow wave sleep.

Results

Comparisons in the PSG results between N1 and N2 are presented in Table 1. Pearson correlation between the overall AHIs from two PSG studies conducted approximately 1 month apart was 0.44 (Fig. 1) and not significant. Figure 2 presents the corresponding Bland–Altman plot. All other correlations between N1 and N2 were significant except for the non-supine AHI (Table 1).

There were no significant differences between nights 1 and 2 in the total sleep time, sleep efficiency, minutes non-REM and REM, percent time REM, sleep latency, REM latency, and minutes stage 1, 2 and slow wave sleep across all subjects or within the group of five outliers who exhibited an AHI increase greater than 20 at retest (see Table 2).

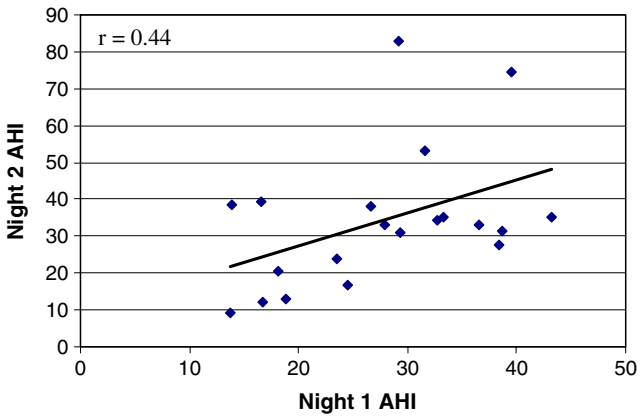

Fig. 1 Correlation plot of the overall AHI between nights 1 and 2

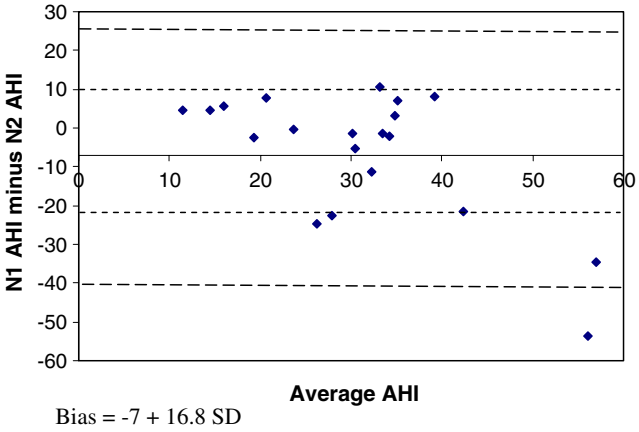

Fig. 2 Bland–Altman plot of the differences in AHI between nights 1 and 2

Discussion

When conducting a treatment intervention study, it is assumed that a level of reliability can be obtained from the measurement tool such that the outcome can be reasonably assessed. In this small study, either the reliability of PSG, the de facto gold standard for measuring OSA severity, was poor or the night-to-night variability of subjects' OSA is large. Because the study sample was small, it suggested that it may be prone to the effect of influential outliers. Figure 2 suggests that two patients are the main source of the overall lower AHI on the second night. With those two cases removed, however, the correlation only improved from 0.44 to 0.45. We found that 25% of the participants had an AHI increase of at least 20 events/hour on night 2 and only nine of the 20 subjects (45%) had a night-to-night AHI difference of ≤5 events/hour.

Inter-rater reliability was eliminated as a factor because the scorer was kept consistent for each patient. The correlations between nights 1 and 2 for the percentage of

Table 2 Repeated measures sleep architecture results from two PSG studies

Mean + SD	N1	N2	N2 minus N1
Total sleep time	364±74	375±81	11±49.4
Sleep efficiency	79±9.6	82±10.8	3±8.9
Non-REM—min	292±72	300±75	8±46.4
REM—min	68±24	72±33	3±26.0
REM—% time	19±5.1	18±7.4	0±7.7
Sleep latency	20±19.9	17±17.6	−2±17.1
REM latency	89±46.6	96±69.5	−7±57.2
Stage 1—min	32±47.4	42±60.5	10±22.0
Stage 2—min	212±53.8	205±62.0	−7±50.9
SWS—min	61±35.7	69±29.1	8±28.4

time supine (r=0.70) and the Supine AHI (r=0.69) suggest that the influence of the supine position was not the primary factor contributing to the high night-to-night variability. Changes in sleep architecture did not appear to be a significant source of between-test variability. We found that only three of the 20 cases showed a difference in the percent time REM greater than 10 with two cases in the outlier group (i.e., AHI increase >20 upon retest). There was a bias toward increased sleep efficiency on night 2 across subjects: three of the six subjects with substantial sleep efficiency changes (i.e., >10%) were in the outlier group. There was a bias toward a decrease in REM latency across subjects: three of the five cases with an increase in REM latency greater than 50 min were outliers. Two non-outliers, however, showed a decrease in REM latency greater than 100 min. All five outliers exhibited either a substantial change in the percent time REM or sleep efficiency but not both. The only apparent explanation for the poor reliability was the surprisingly poor correlation observed in the non-supine AHI (r=0.25).

One of the limitations of this study is that we were unable to control more thoroughly other potential sources of the night to night variability. As it is usual practice to allow patients to consume their usual nightly amount of alcohol prior to lights out in the sleep centers conducting the PSG, it is uncertain what contribution alcohol consumption had on these findings. In hindsight, it would have been helpful to document the amount of alcohol, which was consumed prior to the start of each study. In reviewing the available data, there were no important changes in medication between the test and retest. The impact of placebo on the retest results is uncertain and may have contributed to the substantial bias toward increasing AHI values.

The literature on repeated-measures PSG on symptomatic patients is somewhat limited and tended to focus on patients with mild OSA. Dean and Chaudhary [12] investigated PSG variability in nine patients who appeared negative during the first PSG study and positive in the second PSG. Chediak et al. [11] reported that 12 of 37 (32%) of their cases exhibited a difference in AHI ≥10 in two sequential nights of PSG. Le Bon et al. [13] studied 243 subjects during sequential nights of PSG and evaluated the benefit of improved sensitivity and specificity as a result of having multiple nights of data. More recently, Carlile and Carlile [15] found that 48% of patients with an AHI <5 had AHI values ≥5 on a repeat study. In all four studies, it was apparent that night-to-night variability was significant, and two studies confirmed our findings that AHI values are biased toward an increase at retest with PSG.

As a result of the patients enrolled in this study having more severe OSA, had either the PSG test or retest been used to make a diagnosis using a clinical threshold of 10, only one of 20 subjects would have been classified differently. With a clinical AHI threshold of 15, three of the cases (14%) would have been classified differently. The PSG results obtained in this study, however, were used to validate treatment efficacy whereby a 50% reduction in the overall AHI was the applicable clinical threshold. Under this scenario, variability in the PSG results becomes much more important. We found that 12 of the 20 subjects (60%) showed greater than a 25% change in overall AHI between the two studies and there was a bias toward increasing AHI values on night 2 (mean 27% + 16 SE).

This study suggests a conundrum when assessing OSA treatment outcomes. Human subjects appear to have night-to-night AHI variability, which is not readily controlled by laboratory procedures. To avoid confounding findings, either multi-night studies are required to average out the variability or more sophisticated measures need to be applied to the test results to measure outcomes.

Conclusions

This study confirms previous findings that night-to-night variability in symptomatic untreated patients with an AHI >5 contributes to poor between-trial reliability. Studies employing the AHI as an outcome need to be adequately powered with respect to the inherent night-to-night variability in the measurement. When assessing outcomes in clinical research or performing case findings, there may be benefit from the acquisition and averaging of multiple nights of data in order to reduce the inherent night-to-night variability of OSA and improve the accuracy of the outcome assessment.

References

1. Flemons WW, Littner MR, Rowley JA, Gay P, Anderson WM, Hudgel DW, McEnvoy RD, Loube DI (2003) Home diagnosis of sleep apnea: a systematic review of the literature. An evidence review cosponsored by the American Academy of Sleep Medicine, the American College of Chest Physicians, and the American Thoracic Society. Chest 124(4):1535–1542
2. Kushida C, Morgenthaler T, Littner M, Alessi CA, Bailey D, Coleman JA Friedman L, Hirschkowitz M, Kapen S, Kramer M, Lee-Chiong T, Owens J, Pancer JP (2006) Practice parameters for the treatment of snoring and obstructive sleep apnea with oral appliances: an update for 2005. Sleep 29(2):240–243
3. Dement WC, Vaughan CC (1999) In: House R (ed) The promise of sleep: a pioneer in sleep medicine explores the vital connection between health, happiness, and a good night's sleep. Delacorte, New York
4. Healthy People 2010. 2000: US Department of Health and Human Services
5. Kripke DF, Ancoli-Israel S, Klauber MR, Wingard DL, Mason WJ, Mullaney DJ (1997) Prevalence of sleep-disordered breathing in ages 40–64 years: a population-based survey. Sleep 20(1):65–76

6. Norman RG, Pal I, Stewart C, Walsleben JA, Rapoport DM (2000) Interobserver agreement among sleep scorers from different centers in a large dataset. Sleep 23(7):901–908
7. Whitney CW, Gottlieb DJ, Redline S, Norman RG, Dodge RR, Shahar E, Surovec S, Nieto FJ (1998) Reliability of scoring respiratory disturbance indices and sleep staging. Sleep 21 (7):749–757
8. Collop NA (2002) Scoring variability between polysomnography technologists in different sleep laboratories. Sleep Med 3(1):43–47
9. Cartwright RD (1984) Effect of sleep position on sleep apnea severity. Sleep 7(2):110–114
10. Kavey NB, Blitzer A, Gidro-Frank S, Korstanje K (1985) Sleeping position and sleep apnea syndrome. Am J Otolaryngol 6(5):373–377
11. Chediak AD, Acevedo-Crespo JC, Seiden DJ, Kim HH, Kiel MH (1996) Nightly variability in the indices of sleep-disordered breathing in men being evaluated for impotence with consecutive night polysomnograms. Sleep 19(7):589–592
12. Dean RJ, Chaudhary BA (1992) Negative polysomnogram in patients with obstructive sleep apnea syndrome. Chest 101 (1):105–108
13. Le Bon OG, Hoffmann G, Tecco J, Staner L, Noseda A, Pelc I, Linkowski P (2000) Mild to moderate sleep respiratory events: one negative night may not be enough. Chest 118(2):353–359
14. Le Bon O, Verbanck P, Hoffmann G, Murphy JR, Staner L, De Groote D, Mampunza S, Dulk AD, Vacher C, Kornreich C, Pelc I (1997) Sleep in detoxified alcoholics: impairment of most standard sleep parameters and increased risk for sleep apnea, but not for myoclonias—a controlled study. J Stud Alcohol 58(1):30–36
15. Carlile J, Carlile N (2008) Repeat study of 149 patients suspected of having sleep apnea but with an AHI < 5. Sleep 31:A153