Peer reviewed article

Inter-observer agreement on apnoea hypopnoea index using portable monitoring of respiratory parameters

Pierre-Olivier Bridevaux[a,b], *Jean-William Fitting*[a], *Jean-Marc Fellrath*[a], *John-David Aubert*[a]

[a] Division of Pulmonary Medicine, University Hospital Lausanne, Switzerland
[b] Division of Pulmonary Medicine, University Hospital of Geneva, Switzerland

Summary

Background: Although portable polygraphy or portable monitoring of respiratory parameters (PM) is commonly used to confirm obstructive sleep apnoea syndrome, agreement on apnoea hypopnoea index (AHI), the main measure of disease severity, has not been evaluated. The aim of this study was to assess the agreement on AHI among multiple observers as well as between individual observers and automated analysis.

Methods: A total of 88 ambulatory sleep recordings ("Embletta"®) were independently scored by 8 physicians (observers). Agreement on AHI, using intraclass correlation coefficient (ICC), was measured among observers. Bland Altman plots were built to compare individual observers with PM.

Results: Among observers, ICCs were 0.73 for agreement on AHI, 0.71 for hypopnoea index and 0.98 for desaturation index. Compared to visual analysis, automated analysis underestimated AHI by 5.1 events on average. When comparing individual observers with automated analysis, systematic bias varied from –1.2 to +16.5 events/h on AHI.

Conclusions: Among observers who used PM in a clinical setting, agreement on AHI was limited. When automated and individual visual analyses were compared, the systematic bias varied from almost zero to values sufficient to affect clinical diagnosis. Much of the discordance was due to different counts of hypopnoea, whereas agreement on apnoea and desaturation index was better. Efforts should be directed towards standardisation of visual analysis, improvement and quality control of ambulatory sleep studies.

Key words: home sleep study; portable monitoring of respiratory parameters; diagnostic test; sleep apnoea syndromes

Introduction

Obstructive sleep apnoea syndrome (OSAS) has been linked to road traffic accidents [1] and hypertension [2], and is suspected of constituting an independent risk factor for stroke [3]. It is estimated that 2–4% of the adult population suffer from OSAS, and that it is more frequent among older obese men [4]. The true prevalence of OSAS may be higher, as a result of probable underdiagnosis due to the lack of disease awareness among patients or physicians, as well as the limited availability of sleep laboratories. Consequently, interest in less expensive and more accessible portable monitoring of respiratory parameters (PM), also called portable polygraphy, has steadily grown during the last decade. However, the accuracy of PM has been insufficiently evaluated. Some studies compared the sensitivity and specificity of PM to polysomnography (PSG) in specialised centres, and concluded that PM was as reliable as PSG [5, 6]. No study to date has focused on the reliability of PM regarding apnoea and hypopnoea indexes in the everyday practice of sleep medicine.

Current recommendations strictly define apnoea as a cessation of oronasal flow of ≥10 s. On the other hand, the definition of hypopnoea may

Abbreviations:

AHI	Apnoea hypopnoea Index
AI	Apnoea index
BMI	Body-mass index
CPAP	Continuous positive airway pressure
EEG	Electroencephalogram
HI	Hypopnoea index
ICC	Intraclass coefficient
OSAS	Obstructive sleep apnoea syndrome
PM	Portable monitoring
PSG	Polysomnography

Disclaimer: none
Source of support: none

vary from centre to centre. Recent guidelines from the Academy of Sleep Medicine define hypopnoea as a reduction of 50% from baseline amplitude of the plethysmography signal or a clear reduction of this signal associated with either oxygen desaturation of ≥3% or arousal. The hypopnoea should last ≥10 s. [7]. Thus, in the absence of EEG signals, the PM automated software algorithm or observers may not count nasal flow reduction of less 50% if not associated with oxygen desaturation as a hypopnoeic event. These may be suspected only from other tracings such as pulse acceleration. This definition of hypopnoea leaves room for subjective interpretation. We postulated, however, that this would not lead to significant differences in AHI when reviewing the tracings of home sleep studies.

The goal of the present study was to assess the inter-observer reproducibility of reviewed data from PM, by measuring 1) agreement among observers on AHI, and 2) agreement between visual and automated analysis of AHI in a clinical setting.

Methods

Subjects, material and observers

Data were collected from home recordings of 11 subjects with suspected OSAS and referred to our centre. On average, the 11 patients included were middle-aged (mean age 54 years, SD 14), overweight (mean BMI 27 kg/m^2, SD 4) male snorers. Mean Epworth Sleepiness Scale score was 10 (SD 7) and mean neck circumference was 43 cm (SD 2).

Each subject underwent a full night study (minimum duration 6 hours) using portable monitoring of respiratory parameters "Embletta pds®" (Resmed Corporation, Reykjavik, Iceland). The system records nasal flow with a pressure transducer system, thoracic and abdominal movement through piezoelectric belts, oxygen saturation, pulse rate and body position. Criteria for apnoea and hypopnoea were manufacturer's stated default values [8].

Eight pulmonary physicians trained in reading and interpreting polygraphic records independently reviewed the sleep studies of the 11 patients. All worked in the same pulmonary clinic (Pulmonary Care Division, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland) and shared a common approach to sleep medicine. They were blinded to the patients' physical and historical data. After the automated analysis, tracings were reviewed on a computer screen. A qualitative general assessment of the tracings was made (0 = not interpretable, 1 = in some measure interpretable, 2 = partly interpretable, 3 = mostly interpretable, 4 = perfectly interpretable). Time devoted to reviewing the tracings was also recorded.

AHI definition

The total number of apnoea, hypopnoea, and oxygen desaturation were counted and divided by the total sleep time (hours) to calculate the AHI and desaturation index. As we postulated that hypopnoeas would represent the main source of AHI discordance, we analysed hypopnoeas and apnoeas separately.

Statistical analysis

Agreement among observers (physicians) was measured using the intraclass correlation coefficient (ICC) as described by Shrout and Fleiss [9]. The "case 2" method was applied, given that the same observers rate each patient and assuming that our observers can be considered a random sample of pulmonary physicians. The ICC represents concordance where 1 is perfect agreement and 0 is no agreement at all.

A second analysis, using Bland-Altman plots, focused on difference between the automated analysis and, i) the observers as individuals, ii) the observers as a group (mean AHI for the visual analysis). In this study, the Bland-Altman plots served to estimate the extent of the systematic bias between individual observers and the automated analysis [10].

All statistical analyses were performed with STATA™ version 9.1 (Stata Corporation, College Station Texas USA).

Results

Quality of tracings and scoring time

The median overall quality of tracings was 3 (range 1–4) meaning that most of the tracings were interpretable. The median time over all observers and patients was 23 min (interquartile range 9 min). The median review times for the fastest and slowest observers were 15 and 60 min respectively, revealing that some observers spent 4 times as long to review the tracings.

Individual AHI

The results of automated and visual analysis for AHI are shown in Table 1. For 8 out of the 11 patients, the range of score was wide enough to classify them in a different group of AHI severity as specified by the usual, however arbitrary, diagnostic cut-off (<10, 10 to 19.9, ≥20 events/h). Patient 2, for example, was assigned an AHI of 5.0 by one observer, and of 22.6 by another.

Agreement among observers

Table 2 presents agreement as measured by ICC between observers for AHI, AI, HI and desaturation index. Contrasting with an excellent agreement on desaturation, agreements on respiratory-flux derived indexes were substantially lower. For AI, the F values, derived from variance table analysis, show that observers do not significantly differ in their assessment.

Table 1

Individual apnoea hypopnoea index results.

	Apnoea hypopnoea index analysis				
	Automated	Visual			
Patient		Mean	Median	Min	Max
1	1.6	1.9	1.2	0.0	6.6
2‡	5.6	11.5	6.9	5.0	22.6
3‡	10.8	15.7	11.7	9.0	30.4
4‡	12.6	15.3	13.6	7.8	30.1
5‡	14.4	22.4	17.5	14.1	40.0
6‡	14.7	21.5	19.7	15.0	30.9
7‡	15.9	22.3	19.8	15.7	29.6
8‡	16.9	25.2	21.7	14.0	42.8
9‡	21.5	20.9	21.5	11.0	27.8
10	27.0	34.5	30.0	27.0	46.2
11	36.8	43.1	42.5	34.0	53.7

‡: Patients exposed to misclassification of apnoea hypopnoea index severity

Table 2

ICC measuring agreement between observers for apnoea hypopnoea index, hypopnoea index, apnoea Index and O_2 desaturation index.

Index	ICC	F tests* (associated p-value)
AHI, event·h-1	.73	49.72 (<0.01)
AI, event·h-1	.67	1.37 (0.24)
HI, event·h-1	.71	27.36 (<0.01)
Desaturation, event·h-1	.97	5.11 (<0.01)

AHI: apnoea hypopnoea index; AI: apnoea index; HI: hypopnoea index
* F-tests for variation among observers (assuming no interaction between observers and patients)

Agreement between automated and individual visual analysis

Figure 1 displays Bland-Altman plots for each observer, sorted by growing discordance on AHI with the automated analysis. When comparing individual observers to automated analysis, mean differences varied from –1.2 to +16.5 events per hour. This shows that some observers tended to adhere to the automated analysis while others largely disagreed. In contrast, agreements between individual and automated scores of oxygen desaturation were almost perfect (individual data not shown).

Figure 2 displays the Bland and Altman plots for all observers (mean of observers' reported indexes). Automated versus visual analysis systematically underestimated AHI, HI and AI by 5.1, 4.8 and 5.0 events on average, but the systematic difference for desaturation index was close to zero.

Discussion

In this study we found that for portable monitoring of respiratory parameters agreement among observers on AHI is limited. We also show that agreement between individual observers and automated analysis on AHI varies considerably from one observer to another.

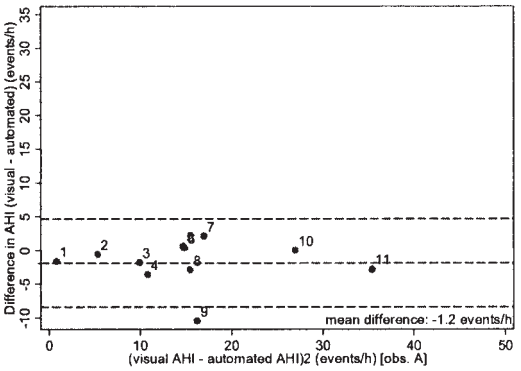
These results were unexpected, given the previous work comparing PSG and a portable device, which found a substantial correlation on AHI [6]. Several aspects of our study could explain these differences.

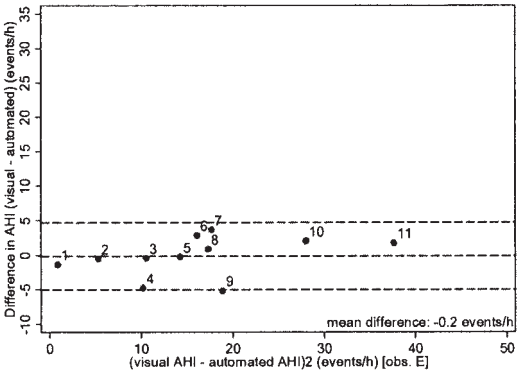
First, in our study, we had 8 observers – compared to 2 in other studies – reading each PM; this could lead to more variation in the interpretation of tracings.

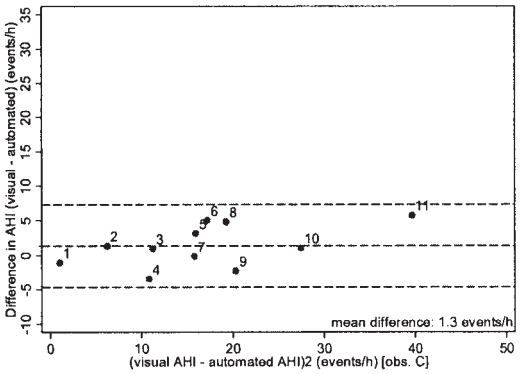
Second, much of the discordance on AHI was due to differences in the count of hypopnoeic events. In contrast to PSG, where hypopnoea can be detected from a nasal flow reduction of less than 50% when associated with oxygen desaturation or a micro-arousal, PM does not record EEG tracings. Thus, detection of nasal flow reduction without associated micro-arousal cannot be directly diagnosed as hypopnoea. Moreover, lack of agreement on the indirect evidences of micro-arousal with PM (such as heart rate changes) may widen the differences in interpretation when considering hypopnoea. However, adding EEG traces to PM probably would not entirely resolve the issue of discordance on hypopnoeas, as a previous study using supervised polysomnography showed limited agreement on micro-arousal [11]. In addition, although data on arousals are produced by EEG monitoring, the superiority of PSG over PM is not proven in terms of a decision to treat subjects with suspected OSAS.

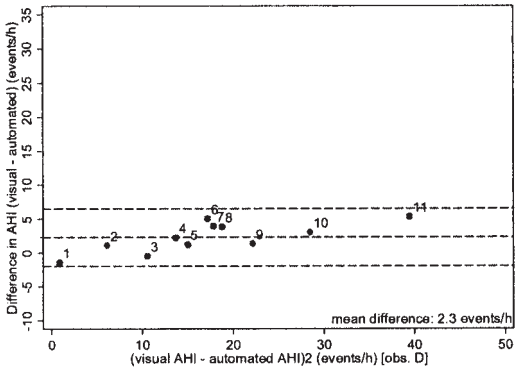
Figure 1

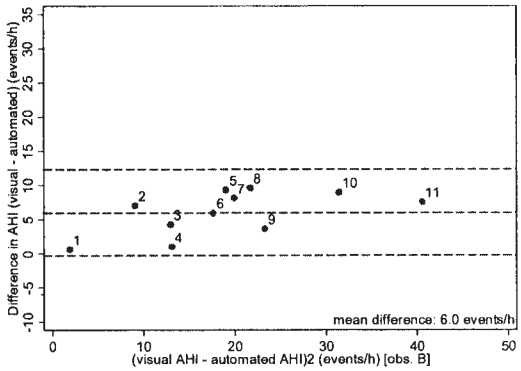
Bland Altman plots illustrating systematic differences between visual and automated scores on apnoea Lines on each plot represent the mean difference (visual – automated analysis) and +/- 1.96 SD of the difference. Label 1-11: patients; AHI: apnoea hypopnoea index.

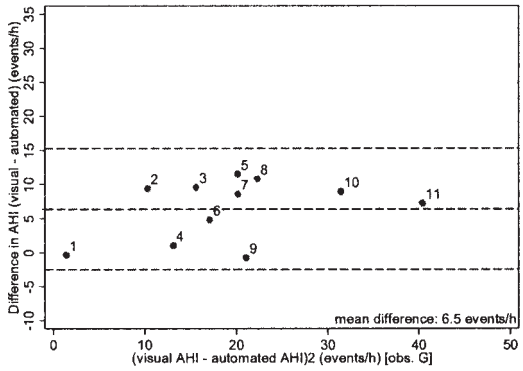


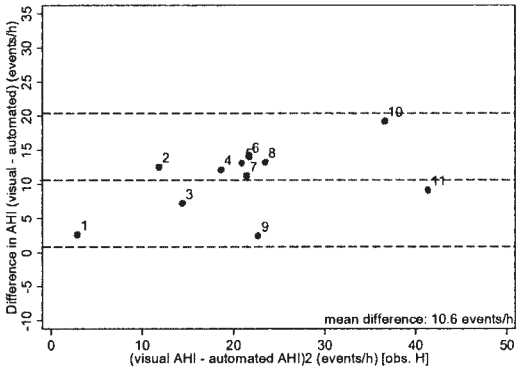


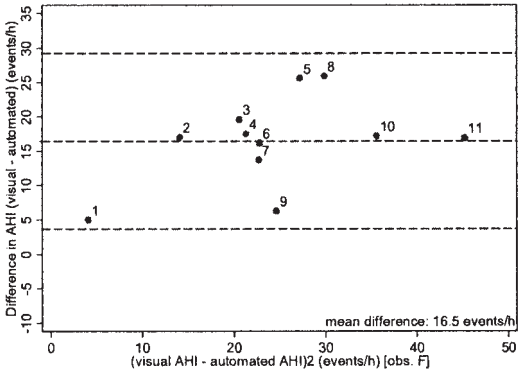












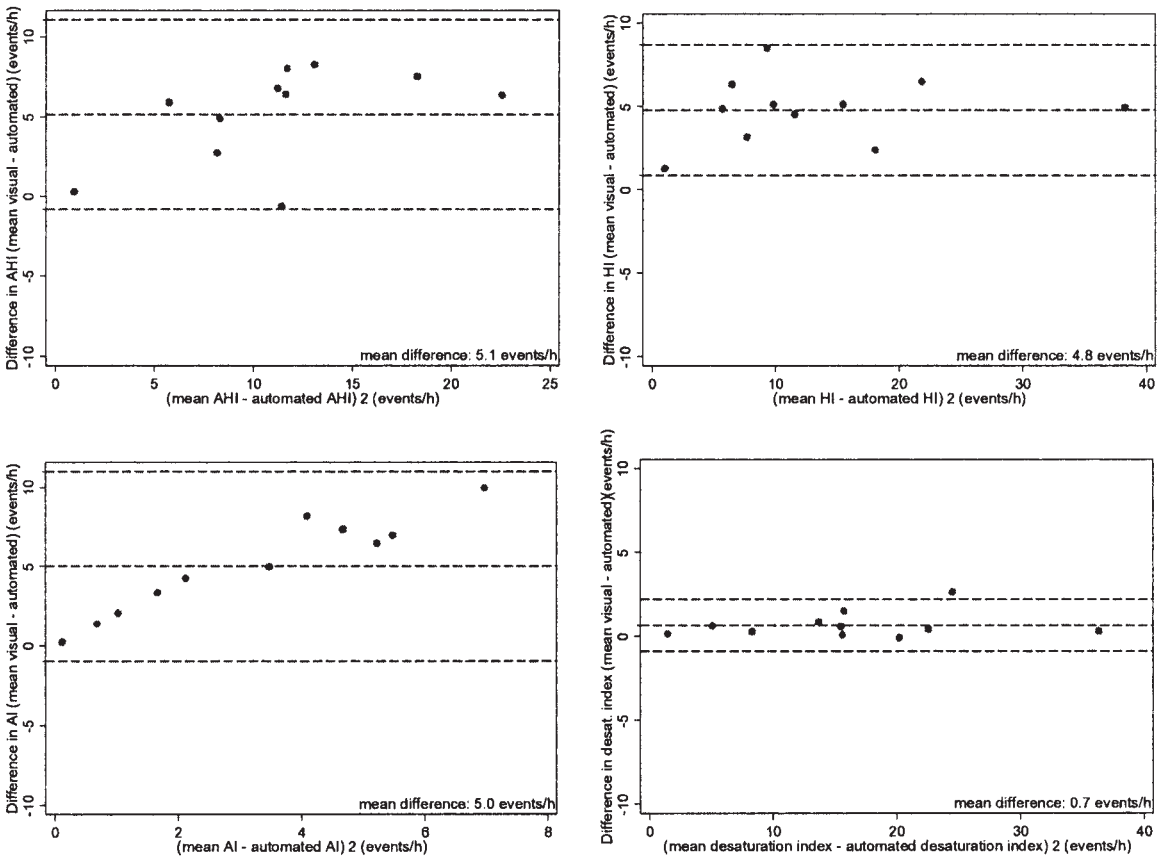


Figure 2

Bland Altman plots of differences in apnoea hypopnoea Index, hypopnoea index, apnoea index and O_2 desaturation index for all observers. Full circles: mean of all patients' results. Lines on each plot represent the mean difference (visual – automated analysis) and +/- 1.96 SD of the difference; AHI: apnoea hypopnoea index; AI: apnoea index; HI: hypopnoea index.

Third, in real life, physicians interpret the home sleep studies with a pre-test probability of OSAS that can influence their final decision on AHI in either direction. This probability is based on history and clinical examination. In our study, observers were blinded to the subjects' clinical characteristics, and thus were not biased when reviewing the tracings to change their measure of AHI. The exact role of pre-test probability as an aid in the clinical decision-making process of prescribing CPAP treatment has so far not been prospectively studied.

In our view inter-observer variability on AHI with PM is clinically important, mainly because it involves the risk for the patient with OSAS of being misclassified as "simple snorer" and therefore not eligible for CPAP therapy. This risk is most marked for patients with an AHI in the range of 10–20 events per hour, since few added or suppressed respiratory events may dramatically affect the final diagnosis. Although the decision to treat is not made exclusively on AHI, the latter constitutes an important outcome in the event of future therapy or follow-up sleep studies.

Also, as AHI has been associated with motor vehicle accidents [12], there has also been discussion as to whether to screen professional drivers for high AHI as a key measure of OSAS. As PSG is expensive, time-consuming and of limited access, PM will probably be employed in this setting. The design of our study, focused on patients and not on randomly selected subjects, limits extrapolation to other populations such as professional drivers. However, given the low sensitivity of PM for the diagnosis of OSAS, and the variability of interpretation of the tracings, inter-observer agreement on apnoea and hypopnoea indexes in this specific population should be evaluated before implementation of PM as screening tools for OSAS among professional drivers. Indeed, as professional drivers' characteristics (wider range of age or weight, lower prevalence of OSAS symptoms) differ considerably from a clinical sample, PM may be considered sufficiently reliable in this setting.

Our study suffers from some limitations. First, it did not evaluate inter-observer agreement on the decision to treat a patient with suspected OSAS. Hence conclusions on the impact of different interpretations regarding AHI and the choice of final treatment cannot be drawn. A study on the whole clinical process with treatment options as an outcome has thus far not been designed. It could integrate the clinical pre-test probability of OSAS as well as the interpretation of PM traces, and would allow different conclusions on the changes in treatment induced by the varying interpretation of PM tracings.

Second, ICC is a ratio of the variability of different ratings on the same subject to the total variation across all ratings and subjects. Thus, ICC depends heavily on the variability of test results in the population studied and should be interpreted with caution when applied to different populations. In contrast to heterogeneous populations (mixing healthy and diseased populations

with large between-subject variance), homogeneous populations such as a sample of patients exhibit common characteristics (e.g. elevated IHA with small between-subject variance). Thus, in a random sample drawn from the general population with a low prevalence of abnormal AHI, the observers will probably show better agreement on AHI than in a clinical subgroup of middle-aged, overweight men with a rather high mean AHI, as measured in our sample. Hence our results may not be generalisable to a population with a low prevalence of OSAS, such as a group of subjects randomly selected and including young, non-snoring, non-obese persons.

In summary, we found that inter-observer agreement on AHI, derived from PM, is limited in a clinical setting. Second, we showed that systematic differences on AHI counts varied considerably when comparing individual and automated analysis. This discordance is of clinical importance and may result in under- or over-diagnosis of OSAS. To improve the inter-observer agreement on AHI, efforts should be made to standardise the interpretation of ambulatory sleep studies and thus to allow more reliable diagnosis of patients suspected of having OSAS. This goal could be achieved by various means, among which quality controls supervised by certified sleep centres would certainly enhance inter-observer agreement. For patients with an AHI score between 10 and 20 a second review of the tracing could be recommended, particularly if a decision not to treat is made.

We are indebted to Christiane Ruffieux, Biostatistician, Institute of Social and Preventive Medicine, University of Lausanne, for statistical assistance.

Correspondence:
Dr Pierre-Olivier Bridevaux, MSc
Service de pneumologie
Hôpitaux Universitaires de Genève
CH-1211 Genève
Pierre-Olivier.Bridevaux@hcuge.ch

References

1 Teran-Santos J, Jimenez-Gomez A, Cordero-Guevara J. The association between sleep apnea and the risk of traffic accidents. Cooperative Group Burgos-Santander. N Engl J Med. 1999;340(11):847–51.
2 Nieto FJ, Young TB, Lind BK, Shahar E, Samet JM, Redline S, et al. Association of sleep-disordered breathing, sleep apnea, and hypertension in a large community-based study. Sleep Heart Health Study. Jama. 2000;283(14):1829–36.
3 Yaggi HK, Concato J, Kernan WN, Lichtman JH, Brass LM, Mohsenin V. Obstructive sleep apnea as a risk factor for stroke and death. N Engl J Med. 2005;353(19):2034–41.
4 Young T, Palta M, Dempsey J, Skatrud J, Weber S, Badr S. The occurrence of sleep-disordered breathing among middle-aged adults. N Engl J Med. 1993;328(17):1230–5.
5 Flemons WW, Littner MR, Rowley JA, Gay P, Anderson WM, Hudgel DW, et al. 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. 2003;124(4):1543–79.
6 Golpe R, Jimenez A, Carpizo R. Home sleep studies in the assessment of sleep apnea/hypopnea syndrome. Chest. 2002; 122(4):1156–61.
7 Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research. The Report of an American Academy of Sleep Medicine Task Force. Sleep. 1999;22(5):667–89.
8 Dingli K, Coleman EL, Vennelle M, Finch SP, Wraith PK, Mackay TW, et al. Evaluation of a portable device for diagnosing the sleep apnoea/hypopnoea syndrome. Eur Respir J. 2003;21(2):253–9.
9 Shrout PEF, Joseph L. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin. 1979;2(Mar):420–8.
10 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307–10.
11 Drinnan MJ, Murray A, Griffiths CJ, Gibson GJ. Interobserver variability in recognizing arousal in respiratory sleep disorders. Am J Respir Crit Care Med. 1998;158(2):358–62.
12 George CF, Smiley A. Sleep apnea & automobile crashes. Sleep. 1999;22(6):790–5.

Established in 1871

Formerly: Schweizerische Medizinische Wochenschrift

Swiss Medical Weekly

Official journal of the Swiss Society of Infectious diseases, the Swiss Society of Internal Medicine and the Swiss Respiratory Society

The many reasons why you should choose SMW to publish your research

What Swiss Medical Weekly has to offer:

- SMW's impact factor has been steadily rising. The 2005 impact factor is 1.226.
- Open access to the publication via the Internet, therefore wide audience and impact
- Rapid listing in Medline
- LinkOut-button from PubMed with link to the full text website http://www.smw.ch (direct link from each SMW record in PubMed)

- No-nonsense submission – you submit a single copy of your manuscript by e-mail attachment
- Peer review based on a broad spectrum of international academic referees
- Assistance of our professional statistician for every article with statistical analyses

- Fast peer review, by e-mail exchange with the referees
- Prompt decisions based on weekly conferences of the Editorial Board
- Prompt notification on the status of your manuscript by e-mail
- Professional English copy editing
- No page charges and attractive colour offprints at no extra cost

We evaluate manuscripts of broad clinical interest from all specialities, including experimental medicine and clinical investigation.

We look forward to receiving your paper!

Guidelines for authors:
http://www.smw.ch/set_authors.html

EMH FMH Schwabe
Editores Medicorum Helveticorum

All manuscripts should be sent in electronic form, to:

EMH Swiss Medical Publishers Ltd.
SMW Editorial Secretariat
Farnsburgerstrasse 8
CH-4132 Muttenz

Manuscripts:	submission@smw.ch
Letters to the editor:	letters@smw.ch
Editorial Board:	red@smw.ch
Internet:	http://www.smw.ch