Short-Term Variability of Respiration and Sleep During Unattended Nonlaboratory Polysomnography—The Sleep Heart Health Study

Stuart F. Quan, MD[α]; Michael E. Griswold, BS[β]; Conrad Iber, MD[χ]; F. Javier Nieto, MD[ε ι]; David M.Rapoport, MD[φ]; Susan Redline, MD[γ]; Mark Sanders, MD[η]; Terry Young, PhD[ι]

For the Sleep Heart Health Study (SHHS) Research Group

[α]Sleep and Arizona Respiratory Centers and Department of Medicine, University of Arizona College of Medicine, Tucson, AZ; [β]Center for Clinical Trials, Johns Hopkins School of Public Health, Baltimore, MD; [χ]Department of Medicine, University of Minnesota, Minneapolis, MN; [ε]Department of Epidemiology, Johns Hopkins School of Public Health, Baltimore, MD; [φ]Department of Medicine, New York University, New York, NY; [γ]Division of Clinical Epidemiology, Department of Pediatrics, Case Western Reserve University, Rainbow Babies and Children's Hospital, Cleveland, OH; [η]Department of Medicine, University of Pittsburgh, Pittsburgh, PA; [ι]Department of Preventative Medicine, University of Wisconsin, Madison, WI

Study Objectives: To determine the short-term variability of indices of disturbed respiration and sleep during 2 nights of unattended nonlaboratory polysomnography conducted several months apart.

Design: Participants were randomly selected using a block design with stratification on preliminary estimates of 2 criteria: respiratory disturbance index [RDI$_{3\%}$ (apnea or hypopnea events associated with ≥ 3% O$_2$ desaturation): <15 /hour total sleep time, ≥15 /hour total sleep time] and sleep efficiency (SEff: <85% and ≥85%). The RDI and sleep data from initial and repeated polysomnography were compared.

Setting: NA

Participants: A subset of 99 participants in the Sleep Heart Health Study who agreed to have a repeat polysomnogram within 4 months of their original study.

Interventions: NA

Measurements and Results: Acceptable repeat polysomnograms were obtained in 91 subjects (mean study interval: 77 ± 18 {sd} days; range: 31-112 days). There was no significant bias in RDI between study nights using several different RDI definitions including RDI$_{3\%}$ and RDI$_{4\%}$ (apnea or hypopnea events associated with ≥4% O$_2$ desaturation). Variability between studies estimated using intraclass correlations (ICC) ranged from 0.77 to 0.81. For subjects with a RDI$_{3\%}$ <15, variability increased as a function of increasing RDI, but for those with a RDI$_{3\%}$ ≥15, variability was constant. Body mass index, SEff, gender, or age did not directly predict RDI variability. Using RDI$_{4\%}$ cutpoints of ≤5, ≤10 and ≤15 events per hour of sleep demonstrated that 79.1%, 85.7%, and 87.9% of subjects, respectively, had the same classification of SDB status on both nights of study. There also was no significant bias in sleep staging, sleep efficiency, or arousal index between studies. However, variability was greater with ICC values ranging from 0.37 (% time in REM) to 0.76 (arousal index).

Conclusion: In the Sleep Heart Health Study, accurate estimates of the severity of sleep-disordered breathing and the quality of sleep were obtained from a single night of unattended nonlaboratory polysomnography. These findings may be applicable to other large epidemiologic studies provided that similar recording techniques and quality-assurance procedures are followed.

Keywords: Polysomnography, sleep-disordered breathing, hypopnea, apnea, variability, reproducibility

INTRODUCTION

THE MOST COMMONLY USED INDICATOR OF SLEEP DISORDERED BREATHING (SDB) SEVERITY IS THE RESPIRATORY DISTURBANCE INDEX (RDI), WHICH IS SYNONYMOUS WITH THE APNEA HYPOPNEA INDEX (AHI) IN MANY REPORTS. The RDI is the total number of apneas and hypopneas per hour of sleep. In clinical populations, the RDI is used to define the presence of SDB and thus identify individuals who require treatment.[1,2] In addition, for epidemiologic studies, the RDI is used to quantify SDB status and to estimate SDB prevalence.[1] Usually, the RDI is derived from information obtained after a single night of polysomnography. However, if the RDI exhibits considerable night-to-night variability, estimates of disease severity based on a single night study could be misleading.

Previous investigations of RDI variability are conflicting. Some have suggested that RDI derived from a single night of polysomnography is a stable estimate of SDB severity, with excellent reliability for categorizing individuals as having or not having SDB.[3-7] Other reports indicate that there may be considerable night-to-night variability in the RDI, leading to substantial risk of misclassification.[8-14] However, most studies have been small,[10,13] and almost all have utilized laboratory polysomnography.[3,5,8-13,15] Data from unattended nonlaboratory sleep monitoring in the setting of an epidemiologic study are limited.[16,17]

Disclosure Statement

Presented in part at the annual meeting of the Association of Professional Sleep Societies, June 20, 1998

Submitted for publication March 2002

Accepted for publication August 2002

Address correspondence to: Stuart F. Quan, M.D., Arizona Respiratory Center, University of Arizona College of Medicine, 1501 N. Campbell, Tucson, AZ 85724, Voice: 520-626-6115, Fax: 520-626-6970, Email: squan@resp-sci.arizona.edu

It is generally recognized that measurements of sleep architecture are subject to a "first-night effect." The first night of sleep in a laboratory in comparison to subsequent nights is characterized by more wakefulness, a longer initial sleep latency, greater amounts of stage 1 sleep, decreased REM sleep and more sleep fragmentation.[18,19] In contrast, there may be less first-night effect observed when studies are conducted in the home environment.[20-22] However, most of these latter studies were performed in small numbers of individuals, many of whom had insomnia.[21,22] There are few data pertaining to variability of sleep architecture indices from a general population sample.

The Sleep Heart Health Study (SHHS) is a large multicenter cohort study that explores the link between SDB and cardiovascular and cerebrovascular mortality and morbidity in the general population.[23] From December 1995 through February 1998, unattended nocturnal polysomnograms (NPSG) were obtained from 6441 individuals at 10 geographic sites. This paper reports on a substudy performed on a sample of SHHS participants to determine the short-term variability of 2 nights of unattended nonlaboratory NPSG data collected several months apart with particular emphasis on reproducibility of the RDI.

METHODS

Study Design and Participant Selection

At 7 of the 10 field centers of the SHHS[23] (Framingham, South Dakota, and Phoenix did not participate), participants were recruited from individuals who previously had an unattended nonlaboratory NPSG as part of the study. Recruitment was performed using a random stratified block design to ensure that there would be broad ranges of sleep quality and RDI in the sample. Stratification was based on 2 criteria derived from preliminary analysis of a participant's first NPSG: sleep efficiency [Recorded Sleep Time/Sleep Period Time (*vide infra*), SEff <85% or $\geq$85% and respiratory disturbance index (RDI): <15 events/hour total sleep time or $\geq$15 events/hour total sleep time]. For stratification purposes, RDI was defined as the sum of apnea and hypopnea events each associated with a $\geq$3% desaturation per hour of sleep ($RDI_{3\%}$). Stratification was performed using a preliminary analysis of the first NPSG because final scoring at the SHHS Reading Center was not completed for several months after the study had been performed. Preliminary scoring consisted of visual editing and subsequent computerized scoring of sleep and respiration using Compumedics software (*vide infra*), and resulted in preliminary estimates of the $RDI_{3\%}$ that were generally within 2 events per hour of the final scored $RDI_{3\%}$.[24] The second NPSG was performed within 4 months of the initial study (mean study interval: 77 ± 18 (sd) days; range: 31-112 days).

The study design specified the recruitment of 25 participants in each block [low (<85%) SEff /low (<15) $RDI_{3\%}$, low SEff/high ($\geq$15) $RDI_{3\%}$, high ($\geq$85%) SEff/low $RDI_{3\%}$, high SEff/high $RDI_{3\%}$] for a total of 100 individuals to be restudied. Two factors determined the target total sample size. The first was the number required to minimize the 95% confidence interval for estimates of night-to-night variability. For this factor, precision reaches an asymptote at a sample size between 70 and 80. The second was our intention to compare variance estimates between subgroups such as those with a high SEff versus a low SEff. For such an analysis, precision improves when the subsample has a size of 50.

Each participating SHHS site was provided with a list generated by the Data Coordinating Center of individuals who met this substudy's eligibility criteria, identified by stratification block. Sequentially, subjects were approached with a goal of assigning 25 (study wide) to each block. The recruitment timeframe for this substudy occurred at the end of the overall SHHS enrollment period. Because there were fewer participants being recruited, the number of potential participants in 2 of the blocks was limited. Therefore, only 21 and 24 individuals were recruited for the high SEff/high $RDI_{3\%}$ and the low SEff/low $RDI_{3\%}$ blocks, respectively. The size of the other 2 blocks was increased to 27 subjects. This resulted in a total of 99 participants, which approached the intended sample size of 100.

Procedures

Polysomnography data for both the first and second NPSG were collected using an unattended monitor (Compumedics PS-2 series - Compumedics Pty. Ltd, Abbotsville, AU), as previously described in the SHHS.[23] The recording montage was identical in both studies and included: C_3/A_1 and C_4/A_2 electroencephalograms (EEG), right and left electrooculograms (EOG), a bipolar submental electromyogram (EMG), nasal/oral thermocouple (Protec, Woodenville, WA), thoracic and abdominal movement (recorded by inductive plethysmography bands), oximetry (finger pulse oximetry [Nonin, Minneapolis, MN], ECG (recorded by bipolar ECG lead), body position (using a mercury gauge sensor), and ambient light (by a light sensor secured to the recording garment).[23,24]

Briefly, all studies were performed in participants' homes or, for 3 participants, a motel room because their homes were inaccessible. Although no attempts were made to record participants in a uniform environment, NPSGs were not acquired during an acute illness or another event that might have disrupted their normal sleep pattern. Body weights were measured on all participants as part of the first-night data collection but obtained in only 36 participants on the second night. Sensors were placed and equipment was calibrated during an evening visit by a technician, certified by the Reading Center in the performance of SHHS studies. Combinations of tape, gauze, and water-soluble pastes and electrical conducting gels were used to secure sensors and electrodes. Participants wore a specially designed vest that had pockets and pouches used to secure wires and the head box. This vest allowed the participant some freedom of movement without becoming entangled in the lead wires. After hookup, signals were visualized and sensor positions were modified to improve signal quality when needed. Impedance values were checked, and EEG, EOG, and EMG electrodes were replaced if impedance values exceeded 5 KΩ. All data were downloaded from the monitor on the following day and were sent for scoring at the SHHS Polysomnography Reading Center (Cleveland, OH). The processing of the repeat NPSG was performed concurrently with initial NPSGs being scored as part of the main SHHS data collection. Scorers were aware that NPSGs from this study were being processed but were blinded from identifying repeat NPSGs from initial studies. The repeat NPSG always was assigned to the same scorer who had scored the first study.

Sleep stages were scored according to Rechtshaffen and Kales criteria.[24] We calculated the recorded sleep time (RST) as the total number of hours of staged sleep during the study. The RST was measured as the total time spent asleep between sleep onset and either the final sleep epoch preceding wakefulness or the end of the recording. Values were calculated for the time spent in each of stages 1, 2, delta (3/4) and REM sleep and expressed as a percentage of RST. The sleep period time (SPT) was calculated as the time interval beginning with lights out (or sleep onset when "lights off" did not precede sleep onset) and ending with the last epoch of sleep prior to awakening or the end of the recording. The SEff was defined as the RST divided by the SPT. The SEff was not computed in those studies where the SPT could not be determined accurately because of ambiguous light transitions or concerns that the entire sleep period was not captured. Arousals were scored according guidelines published by an American Academy of Sleep Medicine taskforce[25] and reported as the number of arousals per hour of RST (AI).[26]

Apnea was defined as a complete or almost complete (<25% of the baseline) cessation of airflow and hypopnea was defined as a decrease below 70% of baseline on chest, abdominal, or thermocouple channels for at least 10 seconds duration.[24] As described previously, scoring of respiratory events using these definitions was highly reliable [intraclass correlation coefficients (ICC) >0.90].[26] All apneic and hypopneic events were manually identified, but tabulated using Compumedics scoring software. The consistency of manually scored SDB events among different scorers and for the same scorer over time was ensured by rigorous quality-assurance procedures at the SHHS Reading Center. As described in a previous analysis, Compumedics scoring software allowed us to use several definitions of RDI based on the magnitude of associated oxygen desaturation.[27] For this analysis, data is presented for the following definitions of RDI: 1) RDI_{TOT} = the total number of apneas plus hypopneas irrespective of any associated oxygen desaturation/RST; 2) $RDI_{3\%}$ = the total number of apneas plus hypopneas each associated with at least a 3% oxygen desaturation/RST; 3) $RDI_{4\%}$ = the total number of apneas plus hypopneas each associated with at least a 4% oxygen desaturation/RST; and 4) $RDI_{Hyp4\%}$ = the total number of apneas irrespective of any oxygen desaturation plus the total number of hypopneas associated with at least a 4% oxygen desaturation/RST. Previous reports from SHHS have used $RDI_{4\%}$ as the primary index of SDB severity.[28,29]

As described previously,[24] each data channel was assigned a quality code grade according to the duration and quality of signals collected, and each study was given an aggregate quality grade based on the overall interpretability and duration of artifact-free signals. Attempts were not made to score sleep stages or to quantify arousals in studies with excessive artifact in the EEG channels due to unsatisfactory electrode placements, environmental electrical interference, or equipment problems. These latter studies were scored “sleep/wake” only. Acceptable studies were those in which there was a minimum of 4 hours of interpretable data on a least one EEG channel, oximetry, and one respiratory channel (an inductance channel or airflow).[24]

Data Analysis

Basic distributional features of the 4 different definitions of RDI and descriptors of sleep are presented as boxplots. Bland-Altman plots were examined to determine relationships between the magnitude and the degree of variation in the RDI measurements.[30] Intra-subject reproducibility of both RDI and sleep descriptors was estimated using intraclass correlation coefficients (ICC). Mean intrasubject differences, sometimes referred to as the ‘bias’ from night 1 to night 2, were evaluated with confidence intervals and paired *t*-tests. To characterize the variability in RDI measurements while accounting for the intrasubject associations, negative-binomial and Gaussian models were fit using generalized estimating equations to the low and high $\text{RDI}_{3\%}$ stratification groups respectively. Finally, 3 different cutpoints to define the presence of SDB, $\geq$ 5, 10, and 15 events/RST for all 4 definitions of RDI, were used to examine the consistency of classifying participants as having SDB from the first to the second NPSG.

RESULTS

Table 1 shows the demographic characteristics of the 99 participants enrolled in the study. There were 55 men and 44 women ranging in age from 40 to 87 years. The mean body mass index (BMI) was 28.4 for men and 28.2 for women. Ethnic distribution was the following: 84% Caucasian, 6% African American, 6% Hispanic, 3% American Indian, and 1% other. After final scoring, second NPSGs from 8 participants (4 men and 4 women) were found to be unacceptable. All of these participants were Caucasian with a mean age of 73 years (range: 62-83 years) and a BMI of 30.2 (range: 23.3-40.6). Their data were excluded from further analysis. This failure rate is comparable to that observed for the overall SHHS cohort (5%-9%).

The boxplots in Figure 1 illustrate the distribution, including the mean and median values, quartiles, and variability for several different definitions of RDI between the first and second nights of polysomnography. Also shown in Figure 1 are the ICC and their 95% confidence intervals. As demonstrated by the ICC, the values of RDI from each night are highly correlated irrespective of the RDI definition. The ICC ranged from a low of 0.77 for RDI_{TOT} to a high of 0.81 for $\text{RDI}_{3\%}$. In addition, mean differences were not significantly different than 0 ($p > 0.05$) and ranged from 0.44 for $\text{RDI}_{3\%}$ to 0.99 for $\text{RDI}_{4\%}$.

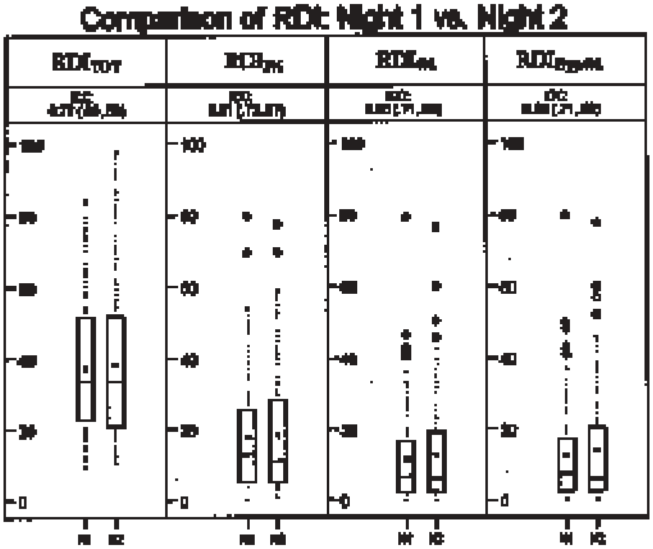

Figure 1—Comparison of RDI: night 1 vs night 2. Boxplots for 4 definitions of RDI (see text) show the mean (x), median (horizontal line), interquartile range (box), and 1.5x the interquartile range (whiskers) for both nights. Open circles (o) represent extreme outliers. Intraclass correlations (ICC) and their 95% confidence intervals are shown for each definition of RDI.

Shown in Figure 2 are Bland-Altman plots[30] for the 4 different definitions of RDI used in this analysis. As depicted by the 95% population intervals, for participants in the low-RDI stratification group, night-to-night variability increased as a function of increasing RDI for all definitions. For participants in the high-RDI group, variability appeared constant. Therefore, in assessing the effects of participant characteristics on RDI, we employed two separate models; a negative binomial model linking the mean and the variance in the low-RDI group, and a Gaussian model specifying mean and variance orthogonality in the high-RDI group (A detailed description of the modeling analysis is given in the Appendix). Regardless of RDI definition, or model used, the mean difference did not significantly vary from 0, and 95% confidence intervals encompassed the 0 difference line, thus reaffirming the absence of any bias from the first to second NPSG. Using our modeling strategy, we found that variability was small but increasing with age and RDI level for subjects in the low-RDI group, while variability was larger, but constant and unaffected by covariates for subjects, in the high-RDI group. Larger values of BMI were observed in the high-RDI group and smaller values of BMI were observed in the low-RDI group. Thus, BMI affected variability by moving subjects from the low-RDI group (in which variability was small but increasing) to the high-RDI group (in which variability was larger but constant). The appendix contains details of our modeling procedures for interested readers. For example, the absolute differences between the first night RDI_{TOT} and the second night RDI_{TOT} were 2.8 and 7.3 for the low-RDI and high-RDI groups, respectively, and did not appear to vary with covariates. Thus, on average, a subject’s second night RDI_{TOT} value should be within approximately 3 events/RST of their first night’s RDI_{TOT} value in the low-RDI group, and within approximately 7 events/RST of their first night’s RDI_{TOT} value in the high-RDI-group. Sleep efficiency, gender, and time interval between studies were not found to affect RDI variability. There were insufficient data on weight change to make assessments of its affect on RDI variation.

Table 2 shows the consistency of classification from the first to the second NPSG using 3 different cutpoints to define the presence of SDB for the 4 definitions of RDI employed in this study. With an $\text{RDI}_{4\%}$, which has been utilized in 2 previous major SHHS publications,[28,29] the same classification was observed in 79.1% of participants on both nights when the presence of SDB was based on an $\text{RDI}_{4\%} \geq 5$. This increased to 85.7% and 87.9% when SDB was defined as an $\text{RDI}_{4\%} \geq 10$ and $\text{RDI}_{4\%} \geq 15$ respectively. A similar degree of consistency was observed using the other 3 definitions of RDI. However, consistent with our previous report,[27] for RDI_{TOT} virtually all participants had an RDI exceeding the cutpoints ≥ 5 and ≥ 10, making this analysis nonmeaningful.

The quality of NPSG data on both nights was adequate to accurately score sleep stages and arousals in 91 participants. As previously described, SPT was not calculated when there were ambiguous light transitions or concerns that the entire sleep period was not captured. Thus, SEff was available in only 49 participants. In contrast to RDI, there was less reproducibility between studies in parameters describing

Table 1—Demographics of Study Participants

		Mean	SD	Min	Max
Men (n=55)	Age	66	12	43	87
	BMI	28.4	4.0	19.4	40.6
Women (n=44)	Age	62	12	40	85
	BMI	28.2	4.9	18.2	39.6

Table 2—Changes in RDI Classification from Night 1 to Night 2

RDI Threshold	No Change [n (%)]	Increased Above Threshold [n (%)]	Decreased Below Threshold [n (%)]
RDI_{TOT}			
≥ 5	*	*	*
≥10	*	*	*
≥15	84 (92.3)	6 (6.6)	1 (1.1)
$RDI_{3\%}$			
≥ 5	74 (81.3)	9 (9.9)	8 (8.8)
≥10	74 (81.3)	5 (5.5)	12 (13.2)
≥15	75 (82.4)	7 (7.7)	9 (9.9)
$RDI_{4\%}$			
≥ 5	72 (79.1)	8 (8.8)	11 (12.1)
≥10	78 (85.7)	5 (5.5)	8 (8.8)
≥15	80 (87.9)	8 (8.8)	3 (3.3)
$RDI_{Hyp4\%}$			
≥ 5	73 (80.2)	6 (6.6)	12 (13.2)
≥10	76 (83.5)	7 (7.7)	8 (8.8)
≥15	79 (86.8)	9 (9.9)	2 (2.2)

* For cutpoints of > 5 and > 10, all participants except 1 had values exceeding the threshold. Therefore, these analyses are not displayed.

sleep and sleep quality. As illustrated in Figure 3, the ICC values for the percentage of time spent in stages 1, 2, and REM sleep as well as sleep efficiency were notably lower than those observed for RDI. Nevertheless, the mean differences between studies were quite small, and there was no significant bias. The ICC values for stage 3/4 sleep and the AI were intermediate between those noted for RDI and other sleep stages and in a range considered to be good.

DISCUSSION

In this study, we observed that there was a high level of agreement between a RDI measured during an unattended NPSG and a subsequent NPSG performed several months later. This finding was consistent among several definitions of RDI. Among those with a low RDI, variability between studies increased as a function of increasing RDI and age. Variability also was altered by BMI, as mediated by its effect on increasing RDI. However, it was not affected by sleep architecture, sleep quality, or gender. In contrast, a greater level of variability was observed in parameters describing sleep and sleep quality.

We found that despite an interval between NPSGs of up to 4 months, there was no consistent bias in RDI between the 2 nights of study and the short-term variability was low. An underlying assumption of most epidemiologic studies of sleep, including the SHHS, is that representative estimates of respiratory and sleep variables can be obtained from a single NPSG recording. This estimate is then assumed to accurately represent the nature of sleep in the individual for all nights during this interval of time. However, signals recorded in an NPSG are subject to both true biologic differences and measurement or "scoring" variability from one night to the next. In either case, large amounts of systematic bias or variability can lead to inaccuracies if there is reliance on data from a single night of study. In addressing this concern for the SHHS, we have found ICC values ranging from a low of 0.77 for RDI_{TOT} to a high of 0.81 for $RDI_{3\%}$. Fleiss classifies ICC values greater than 0.75 as indicating excellent reproducibility.[31] Another interpretation for the ICC is given by Rosner,[32] who notes that √ICC is the correlation between a single measurement on a subject and that subject's "true measurement" (the average of an infinite number of measurements on the subject). Consequently, the estimated correlation between RDI_{TOT} measured on the first night and the "true" RDI_{TOT} measurement is √0.77 = 0.88. Our ICC values thus suggest that 1 night of measurement should suffice for drawing conclusions in the SHHS. Nevertheless, this level of reproducibility is less than previous observations over similar time spans for blood pressure[33,34] and spirometry,[35] which are commonly done in epidemiologic settings. However, it exceeds the repeatability reported for ultrasonic measurements of arterial stiffness, a more complex test that also has been performed in an epidemiologic study.[33] Our data are consistent with a previous study recording only respiratory variables performed in an unattended setting[16] and a small sample of children with unattended monitoring using a recording technique nearly identical to that used in the current study.[17] However, they extend these findings by demonstrating that they can be obtained using a full NPSG montage in SHHS where there are a large number of adults with a broad spectrum of sleep quality and SDB.

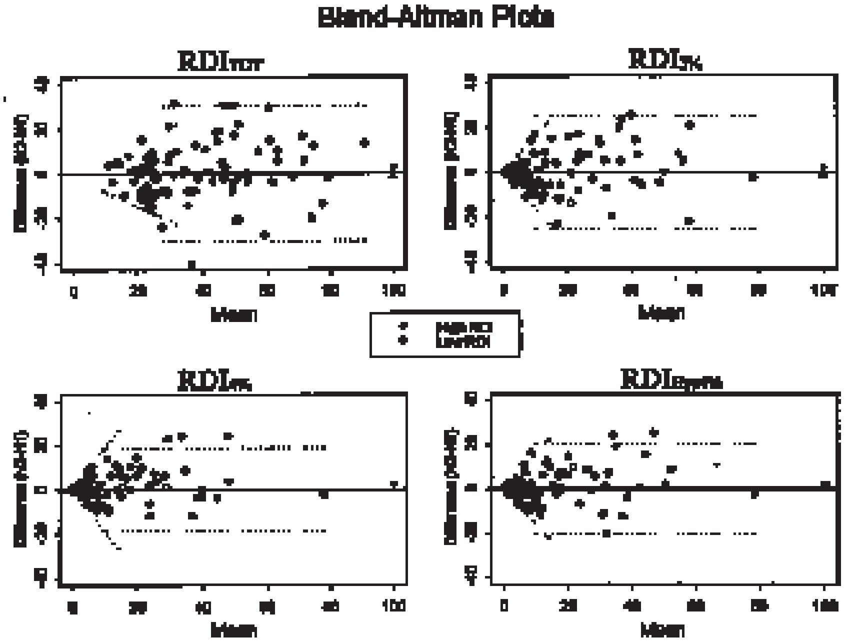

Figure 2—Bland Altman plots for RDI. Plots for 4 definitions of RDI (see text) show the mean bias, 95% population intervals and the 95% confidence intervals. Closed diamonds represent subjects with a preliminary $RDI_{3\%}$ <15 events/hour of sleep and open diamonds represent subjects with a preliminary $RDI_{3\%}$ ≥15 events/hour of sleep. The 95% population intervals are depicted as dashed lines. The 95% confidence intervals are shown as a vertical line on the right of each plot.

We assessed short-term RDI variability by determining consistency of SDB classification using 3 commonly employed cutpoints.[7,14,27] Using an $RDI_{4\%}$ ≥5 to define the presence of SDB, we found that 79.1% of participants received the same classification on both nights of study, which is slightly higher than the 68% to 70% rate of consistency noted in smaller samples studied in a sleep laboratory.[6,11-13] Furthermore, our consistent classification rates of 85.7% and 87.9% for $RDI_{4\%}$ cutpoints of ≥10 and ≥15, respectively, as well as similar classification rates for the other RDI definitions used in this study, are either comparable or slightly better than previously reported.[7,14,16,36] The level of reliability we observed, notwithstanding sampling procedures that were intended to maximize heterogeneity within the study population, may be due to the absence of first night effect on sleep quality in an unattended setting. This hypothesis is supported by data obtained in a sleep-laboratory environment demonstrating an association between better sleep on the second night of polysomnography and a higher RDI. [8 9,36] In addition, the stringent quality-control procedures used to obtain and process polysomnographic data minimized variability related to technical factors.[24]

Despite the level of reproducibility found in this study, we acknowledge that some misclassification may be unavoidable. Because of our rigorous technical standards and the absence of any finding of a significant difference in RDI between study nights, this potential error should be random and not result in any major systematic bias in the assessment of SDB. Nonetheless, the repeatability of a measurement is a major factor in demonstrating whether it is associated with a given outcome. As reproducibility diminishes, there is a greater risk for either underestimating the strength of a relationship or incorrectly accepting the null hypothesis. For the purposes of the SHHS, we believe that this risk is small given the relatively high degree of reproducibility of RDI measurements and the size of the SHHS cohort. Therefore, our results provide evidence to support the utility of NPSG data obtained from a single night to yield reliable indices of the RDI in the SHHS.

Despite our relatively high degree of consistent SDB classification, caution is advised in using these data to determine the adequacy of a single unattended NPSG to make the diagnosis of SDB for patient care purposes. As previously emphasized, these data were obtained using a standardized research protocol using quality-control procedures that may not be available in the clinical arena. Appropriate clinical judgement, in conjunction with information derived from a history and physical examination focused on sleep disorders, should be exercised in such circumstances.

Our data demonstrate that RDI definition was not a significant factor in determining variability between study nights. We analyzed the effect of several commonly used definitions of RDI including $RDI_{4\%}$ which was used in previous major SHHS analyses.[28,29] Although we have shown that RDI definition impacts the prevalence of SDB in our cohort,[27] the data from the current study indicate that a single NPSG can be used to ascertain prevalence rates irrespective of the definition.

In this study— age, RDI, and BMI differentially affected variability according to level of RDI. There are few previous studies examining factors associated with RDI night-to-night variability.[11,15,36] Our data are consistent with these previous reports, which also failed to demonstrate a consistent impact of age, gender, or sleep efficiency on nightly RDI variability. However, they do suggest that the impact of these covariates may differ according to RDI level. Accordingly, future studies in this area may wish to analyze their data similarly. It is possible that differences in body position between the 2 nights of recording explain some of the variability noted in our study. Unfortunately, the accuracy of our body position data collected in the unattended setting could not be independently verified, and thus we did not consider it suitable for analysis. However, previous studies did not find that body position was a major determinant of night-to-night variability.[11,15] Furthermore, although we did not study participants when they were having an acute upper respiratory tract infection, no specific instructions were given regarding alcohol consumption or sedative-hypnotic use. A greater usage of such agents on either the first or second night might result in higher RDI values, changes in sleep architecture, and greater between-study variability. However, no consistent biases were observed between studies, and thus we believe any effect from differential alcohol or sedative/hypnotic use is minimal. We did observe that in those participants with a RDI < 15, some increase in variability between studies occurred as a function of increasing RDI. This finding may be explained in part by the relationship between increasing RDI and age noted in this subgroup. Our observation is not consistent with a previous study.[15] However, as demonstrated on the Bland-Altman plots, the magnitude of this variability in those with a relatively low RDI is small. Thus, it is unlikely that there is any meaningful imprecision in estimating RDI among those with relatively little SDB.

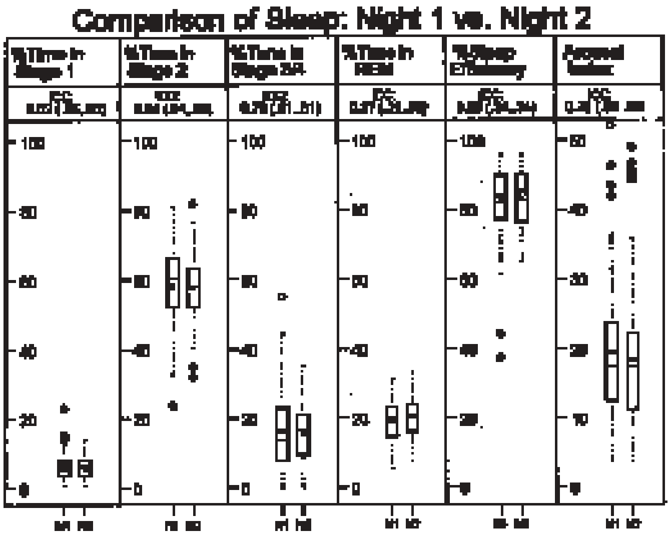

Figure 3—Comparison of Sleep: Night 1 vs. Night 2. Boxplots for % time in stages 1, 2, 3/4, REM; % sleep efficiency; and arousal index show the mean (x), median (horizontal line), interquartile range (box), and 1.5x the interquartile range (whiskers) for both nights. Open circles (o) represent extreme outliers. Intra-class correlations (ICC) and their 95% confidence intervals are shown for each measure.

Consistent with the absence of any meaningful difference in RDI, we did not observe any significant bias in measures of sleep architecture and quality. This finding is at variance with several previous studies demonstrating a first-night effect characterized by more disrupted and less consolidated sleep during an initial night in a sleep laboratory.[18,19,37] However, it is compatible with data obtained by others in the home environment where a systematic first-night effect was minimal or not found.[20-22] Irrespective of whether these previous studies were performed in the sleep laboratory or in the home, data were accrued on successive nights of monitoring. In contrast, the interval between NPSGs in this study was several weeks or months. Therefore, while our data suggest the absence of a first-night effect, we acknowledge that the time between studies may have been sufficiently long to negate any adaptation on the second night of study, perhaps resulting in 2 NPSGs with a first-night effect.

Although we did not find any overall bias in measures of sleep architecture and quality, intrasubject variability represented by the ICC was modest (stages 3/4 and AI) to moderate (stages 1, 2, REM, and SEff), and greater than for RDI. Our findings are consistent with previous studies that showed not only considerable variability in these measurements from night to night, but also that the amount of variation was dependent on the sleep parameter measured.[22,34, 38] Our observation that stages 3/4 and AI had the least amount of variability is similar to some previous reports,[22,34,39] but at variance with others.[38,40] Except for one study performed in insomnia patients,[22] these previous reports utilized data from laboratory polysomnography.[36,38-40] Thus, our data suggest that stages 3/4 and AI may be the most reliable indices of sleep quality in epidemiologic studies when only a single unattended NPSG can be obtained.

The explanation for the amount and pattern of variability between studies in sleep architecture and sleep quality observed in our study is not clear. However, we believe it is not related to scoring variability. Each pair of NPSGs in this study was analyzed by the same scorer and intrascorer reliability for sleep stages in the SHHS Reading Center has been shown to be excellent.[26]

In conclusion, we have demonstrated using unattended nonlaboratory NPSG that there is no consistent bias in RDI, or measures of sleep architecture and quality, between studies separated by up to 4 months. Short-term variability between studies was low for several definitions of RDI, and there was consistent classification of participants as having SDB using several common RDI cutpoints. However, variability between studies for measures of sleep architecture and quality was greater than for RDI. These data indicate that, in the SHHS, accurate estimates of the severity of SDB and the quality of sleep can be obtained from a single night of unattended polysomnography. These observations may be helpful in the interpretation of other epidemiologic studies, provided that similar recording techniques and quality-assurance procedures are used.

ACKNOWLEDGMENTS

This work was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53940 (University of Washington), U01HL53941 (Boston University), U01HL53938 (University of Arizona), U01HL53916 (University of California, Davis), U01HL53934 (University of Minnesota), U01HL53931 (New York University), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL63463 (Case Western Reserve University), and U01HL63429 (Missouri Breaks Research).

The Sleep Heart Health Study (SHHS) acknowledges the Atherosclerosis Risk in Communities Study (ARIC), the Cardiovascular Health Study (CHS), the Framingham Heart Study (FHS), the Cornell/Mt. Sinai Worksite and Hypertension Studies, the Strong Heart Study (SHS), the Tucson Epidemiologic Study of Airways Obstructive Diseases (TES) and the Tucson Health and Environment Study (H&E) for allowing their cohort members to be part of the SHHS and for permitting data acquired by them to be used in the study. The SHHS is particularly grateful to the members of these cohorts who agreed to participate in SHHS as well. The SHHS further recognizes all of the investigators and staff who have contributed to its success. A list of SHHS investigators, staff, and their participating institutions is available on the SHHS website, www.jhsph.edu/shhs.

The opinions expressed in the paper are those of the authors and do not necessarily reflect the views of the IHS.

REFERENCES

1. Redline S, Sanders M. Hypopnea, a floating metric: implications for prevalence, morbidity estimates, and case finding. Sleep 1997; 20:1209-17.
2. Berry DTR, Webb WB, Block AJ. Sleep apnea syndrome: A critical review of the apnea index as a diagnostic criterion. Chest 1984; 84:529-31.
3. Mendelson WB. Use of the sleep laboratory in suspected sleep apnea syndrome: Is one night enough? Cleve Clin J Med 1994; 61:299-303.
4. Lord S, Sawyer B, O'Connell D. Night to night variability of disturbed breathing during sleep in an elderly community sample. Sleep 1991; 14:252-8.
5. Wittig RM, Romaker A, Zorick FJ, Roehrs TA, Conway WA, Roth T. Night-to-night consistency of apneas during sleep. Am Rev Respir Dis 1984; 129:244-6.
6. Lee K, Giblin E. Reliability of a one-night diagnostic study for sleep apnea. Sleep Res 1982; 11:155.
7. Masaquel A, Stepnowsky C, Estline E, Mason WJ, Ancoli-Israel S. Night-to-night variablity in sleep disordered breathing in elderly African-Americans recorded at home. Sleep Res 1997; 26:675.
8. Le Bon O, Hoffmann G, Tecco J, et al. Mild to moderate sleep respiratory events: One negative night may not be enough. Chest 2000; 118:353-9.
9. Dean RJ, Chaudhary BA. Negative polysomnogram in patients with obstructive sleep apnea syndrome. Chest 1992; 101:105-8.
10. Meyer TJ, Eveloff SE, Kline LR, Millman RP. One negative polysomnogram does not exclude obstructive sleep apnea. Chest 1993; 103:756-60.
11. Chediak AD, Acevedo-Crespo JC, Seiden DJ, Kim HH, Kiel MH. Nightly variability in the indices of sleep-disordered breathing in men being evaluated for impotence with consecutive night polysomnograms. Sleep 1996; 19:589-92.
12. Mosko SS, Dickel MJ, Ashurst J. Night-to-night variability in sleep apnea and sleep-related periodic leg movements in the elderly. Sleep 1988; 11:340-8.
13. Aber WR, Block AJ, Hellard DW, Webb WB. Consistency of respiratory measurements from night to night during the sleep of elderly men. Chest 1989; 96:747-51.
14. Kramer M. Obstructive sleep apnea: One night is not enough. Sleep Res 1988; 17:205.
15. Bliwise DL, Benkert RE, Ingham RH. Factors associated with nightly variability in sleep-disordered breathing in the elderly. Chest 1991:973-6.
16. Redline S, Tosteson T, Boucher MA, Millman RP. Measurement of sleep-related breathing disturbances in epidemiologic studies. Assessment of the validity and reproducibility of a portable monitoring device. Chest 1991; 100:1281-6.
17. Goodwin JL, Enright PL, Kaemingk KL, et al. Feasibility of using unattended polysomnography in children for research—Report of the Tucson Children's Assessment of Sleep Apnea Study (TuCASA). Sleep 2001; 24:937-44.
18. Agnew HW, Webb WB, Williams RL. The first night effect: An EEG study of sleep. Psychophysiology 1966; 2:263-6.
19. Mendel J, Hawkins DR. Sleep laboratory adaptation in normal subjects and depressed patients ("first night effect"). Electroencephalogr Clin Neurophysiol 1967; 222:556-8.
20. Sharpley AL, Soloman RA, Cowen PJ. Evaluation of first night effects using ambulatory monitoring and automatic sleep stage analysis. Sleep 1988; 11:273-6.
21. Coates TJ, George JM, Killen JD, Marchini E, Hamilton S, Thorensen CE. First night effects in good sleepers and sleep maintenance insomniacs when recorded at home. Sleep 1981; 4:293-8.
22. Edinger JD, Marsh GR, McCall WV, Erwin CW, Lininger AW. Sleep variability across consecutive nights of home monitoring in older mixed DIMS patients. Sleep 1991; 14:13-7.
23. Quan SF, Howard BV, Iber C, et al. The Sleep Heart Health Study: Design, Rationale, and Methods. Sleep 1997; 20:1077-1085.
24. Redline S, Sanders M, Lind B, et al. Methods for obtaining and analyzing unattended polysomnography data for a multicenter study. Sleep 1998; 21:759-67.
25. EEG arousals: scoring rules and examples: a preliminary report from the Sleep Disorders Atlas Task Force of the American Sleep Disorders Association. Sleep 1992; 15:173-84.
26. Whitney CW, Gottlieb DJ, Redline S, et al. Reliability of scoring respiratory disturbance indices and sleep staging. Sleep 1998; 21:749-57.
27. Redline S, Kapur VK, Sanders MH, et al. Effects of varying approaches for identifying respiratory disturbances on sleep apnea assessment. Am J Respir Crit Care Med 2000; 161:369-74.
28. Shahar E, Whitney CW, Redline S, et al. Sleep-disordered breathing and cardiovascular disease: cross-sectional results of the Sleep Heart Health Study. Am J Respir Crit Care Med 2001; 163:19-25.
29. Nieto FJ, Young TB, Lind BK, et al. Association of sleep-disordered breathing, sleep apnea, and hypertension in a large community-based study. Sleep Heart Health Study. JAMA 2000; 283:1829-36.
30. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1:307-10.
31. Fleiss, JL, The design and analysis of clinical experiments. New York, NY: Wiley, 1986.
32. Rosner B, Fundamentals of Biostatistics. Pacific Grove, CA: Duxbury, 2000.
33. Arnett DK, Chambless LE, Kim H, Evans GW, Riley W. Variability in ultrasonic measurements of arterial stiffness in the Atherosclerosis Risk in Communities Study. Ultrasound Med Biol 1999; 25:175-80.
34. Klungel OH, de Boer A, Paes AH, Nagelkerke NJ, Seidell JC, Bakker A. Estimating the prevalence of hypertension corrected for the effect of within-person variability in blood pressure. J Clin Epidemiol 2000; 53:1158-63
35. Randell JT, Salonen RO, Pekkarinen H, Tukiainen H. Short-term

variations in oscillatory and spirometric lung function parameters of non-asthmatic adults. Clin Physiol 1999; 4:329-337.
36. Bliwise DL, Carey E, Dement WC. Nightly variation in sleep-related respiratory disturbance in older adults. Exp Aging Res 1983; 9:77-81.
37. Schmidt HS, Kaelbling R. A differential laboratory adaptation of sleep parameters. Biol Psychiatr 1971; 3:33-45.
38. Clausen J, Sersen EA, Lidsky A. Variability of sleep measures in normal subjects. Psychophysiology 1974; 11:509-16.
39. Loredo JS, Clausen JL, Ancoli-Israel S, Dimsdale JE. Night-to-night arousal variability and interscorer reliability of arousal measurements. Sleep 1999; 22:916-20.
40. Larsen LH, Moe KE, Vitiello MV, Prinz PN. A note on the night-to-night stability of stages 3 + 4 sleep in healthy older adults: a comparison of visual and spectral evaluations of stages 3 + 4 sleep. Sleep 1995; 18:7-10.

APPENDIX — Modeling Details

We present the details of our modeling strategies to permit cross-study comparisons of parameter estimates. Employing generalized estimating equation (GEE) models allows use of the information in the untransformed, (raw scale), RDI values, while accounting for both the observed mean-variance relationships and the intrasubject associations. Since the variability appeared to increase in the low-RDI stratification group but not the high-RDI group, separate models for the two groups were constructed.

For the high-RDI group, RDI levels did not appear to be modified by covariates and variability appeared constant as RDI levels increased. Hence, the final model for the high-RDI group was specified as: $E(RDI_i) = \beta_0 + \beta_1 I_{night=2}$, $var(RDI_i) = \sigma^2 R$, where $I_{night=2}$ is an indicator for the second night and R is a 2x2 matrix with ones on the diagonal and correlation coefficient ρ on the off-diagonals. The parameter estimates of this model are summarized in Table A1.

High-RDI Models

Variable	β_0	β_1	σ	ρ
RDI_{TOT}	51.5 (46.2, 56.8)	0.45 (-4.3, 5.2)	18.8	0.69
$RDI_{3\%}$	32.8 (28.1, 37.4)	0.33 (-3.6, 4.2)	16.7	0.74
$RDI_{4\%}$	23.1 (18.5, 27.6)	1.25 (-1.7, 4.2)	15.9	0.85
$RDI_{Hyp4\%}$	24.1 (19.9, 28.4)	1.52 (-1.6, 4.7)	16.2	0.87

Table A1—High-RDI model regression parameter estimates, (95% CI), dispersion estimates and working correlation estimates.

Low-RDI Models

Variable	β_0	β_1	β_2	ϕ	ρ
RDI_{TOT}	11.9 (-1.3, 25.1)	0.91 (-1.8, 3.6)	0.24 (0.03, 0.44)	0.119	0.60
$RDI_{3\%}$	-1.37 (-7.8, 5.1)	0.69 (-1.0, 2.3)	0.13 (0.02, 0.23)	0.382	0.33
$RDI_{4\%}$	-1.35 (-5.1, 2.4)	0.95 (-0.3, 2.2)	0.07 (0.01, 0.13)	0.563	0.34
$RDI_{Hyp4\%}$	-1.74 (-6.4, 2.9)	0.91 (-0.27, 2.1)	0.09 (0.01, 0.16)	0.464	0.41

Table A2—Low-RDI model regression parameter estimates, (95% CI), dispersion estimates and working correlation estimates.

For the low-RDI group, the RDI values were found to increase with increasing age and the variability appeared to increase as the RDI level increased. Thus, Poisson, over-dispersed Poisson, and negative binomial GEE models were fit and examined to evaluate the degree of these relationships. The negative binomial model was found to fit the data closest, and the final model for the low-RDI group was specified as: $E(RDI_i) = \mu_i = \beta_0 + \beta_1 I_{night=2} + \beta_2 Age_i$, $var(RDI_i) = \mu_i(1+\phi\mu_i)R$, where $I_{night=2}$ and R are as previously specified and ϕ is the negative binomial distribution dispersion parameter. Different GEE mean link structures were also inspected, and conclusions were not altered by the choice of link; since we have focused on the difference between nightly RDI measurements, the identity link was chosen for model presentation. The parameter estimates of this model are summarized in Table A2.

To ascertain which patient characteristics related to belonging in the high-RDI group (Night 1 RDI3% ≥ 15), versus the low-RDI group, (Night 1 $RDI_{3\%} < 15$), a logistic regression model was examined. BMI was determined to be the only characteristic associated with group definition, and the final model formed was: $logit\{Pr(RDI_{3\%} \geq 15)\} = \beta_0 + \beta_1 BMI$. The parameter estimates (95% CI) for β_0 and β_1 were -6.94 (-10.5,-3.4) and 0.24 (0.11, 0.36), respectively. Thus, the odds for belonging to the high-RDI group increased by approximately e 2.4 = 11 times for a 10-unit increase in BMI.