EDITORIAL

http://dx.doi.org/10.5665/sleep.2524

Entering the Era of "Big Data": Getting Our Metrics Right

Commentary on:
Malhotra et al. Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. SLEEP 2013;36:573-582.
Kuna et al. Agreement in computer-assisted manual scoring of polysomnograms across sleep centers. SLEEP 2013;36:583-589.
Magalang et al. Agreement in the scoring of respiratory events and sleep among international sleep centers. SLEEP 2013;36:591-596.

Susan Redline, MD, MPH[1]; Dennis Dean III, PhD[1]; Mark H. Sanders, MD[2]

[1]Department of Medicine, Harvard Medical School, Brigham and Women's Hospital and Beth Israel Deaconess Medical School, Boston, MA; [2]Wexford, PA

> "The era of 'Big Data' has arrived. There is an urgent need and increased opportunity for increased collaboration and coordination of access to, and analysis of, the many different data types that make up this revolution in biological information, including genomics, imaging and phenotypic data from electronic health records."[1]
>
> *Francis Collins, PhD, Director, National Institutes Health*

Moving into the "Big Data" era means moving away from reliance on small, underpowered clinical studies and analysis of limited data types, to an era where large, complex, and well-annotated datasets containing a full spectrum of clinical, physiological, imaging, and biological data can be reliably generated, easily accessed, and robustly analyzed. The potential of "Big Data" is especially pertinent to the field of Sleep Medicine. Notably, the richness of multi-channel physiological sleep and circadian data collected in clinical as well as in specialized research settings, if appropriately acquired, analyzed, and archived, and combined with other clinical and biological information, could significantly enhance efforts to identify likely phenotypes of cardiorespiratory, neurological, and psychiatric traits. The likely fundamental importance of sleep and circadian biology to all aspects of health, and their essential, integrative role in numerous physiological processes,[2] provide opportunities to accelerate discovery of many disease mechanisms and interventions and to improve patient outcomes.

Leveraging sleep data to support patient-oriented outcome studies, clinical trials, genetic epidemiological, and other studies, especially those that take advantage of "Big Data," requires the field to address several technical, logistical, organizational, and scientific challenges. One key need is to adopt methods that permit reliable acquisition and quantification of clinically and physiologically relevant sleep study metrics in large numbers of patients across multiple clinical and research settings. Clinically generated and clearly defined sleep metrics must be easily incorporated into electronic health records to support much needed outcomes research. For research, metrics should be appropriate for testing hypotheses about genetic and epigenetic etiologies (mechanisms) and for evaluation of intervention effects. Measurement error of sleep study findings should be minimized to limit random misclassification, which acts to attenuate detection of true associations. Error-minimizing methods are needed so that disease detection and classification do not vary according to where a sleep study is done, how the data are acquired, who scores it, or what technique they used to score it. The results of the sleep study will then be truly portable such that all who see the results will be presented with comparable information from which to make inferences. Achieving these research and clinical goals requires adoption of methods across centers that minimize noise in the data engendered by use of varying acquisition parameters and analytical methods as well as the inherent variability associated with manual identification and classification of events (scoring). Given economic imperatives, it is also necessary that measurements and analyses are made in a cost-efficient manner.

What have been the roadblocks to progress in minimizing error variance in sleep measures? Of course, the answer is multifaceted. Two related bottlenecks have been knowledge gaps regarding the links between physiological perturbations and clinical outcomes, and the complexities and vagaries in our measurement approaches. Ideally, the process of characterizing sleep traits would include (1) choosing variables that plausibly reflect underlying pathophysiological or genetic processes; (2) determining how to collect and quantify the signals reflecting those variables, (3) defining and identifying the properties of those variables that are abnormal; and (4) and validating a linkage between these events (in terms of magnitude of abnormality, frequency, or duration) and various health outcomes or intervention responses. Applying this process to data acquired from multiple channel sleep studies has been challenging. While polysomnography (PSG) can provide comprehensive information on dynamic changes in neurophysiological and cardiorespiratory parameters over a period of hours, its application across large samples of individuals, especially when facilitated by sampling geographically diverse populations, is complicated by a need to standardize scoring or quantitative automated output.

Submitted for publication February, 2013
Accepted for publication February, 2013
Address correspondence to: Susan Redline, MD, Harvard Medical School, Department of Medicine and Division of Sleep Medicine; Brigham and Women's Hospital and Beth Israel Deaconess Medical Center, 221 Longwood Avenue, Room 225, Boston, MA 02115; Tel: (617) 732-4013; Fax: (617) 732-4015; E-mail: sredline@partners.org

Factors that can adversely influence standardized scoring output may be present at virtually each level of the sleep study process. Assuming there is agreement on what sleep metrics provide discriminative or prognostic information, there are challenges in deciding how to: (1) acquire and visualize the appropriate signals; (2) operationalize definitions of scored annotations; (3) ensure that such scoring is reliable between and within technicians and across laboratories. In this issue of *SLEEP*, several of these areas are addressed in three interrelated articles that report the reliability of scoring PSG data. Topics covered by these articles include: (1) the across-laboratory agreement of scored data from PSGs that were collected at a single site and imported into different scoring platforms at 9 study sites for local analysis[3]; (2) the degree to which the reliability of PSG variables vary within and across study sites as well as the degree to which different definitions influence the reproducibility of hypopnea detection[4]; and (3) the reliability of an automated computer scoring algorithm compared to a manual scoring.[5] These articles also raise questions regarding which measurement approaches may be most applicable for multicenter sleep research and local versus centralized sleep scoring in supporting multicenter research.

In this commentary, we review some of the key issues related to signal acquisition and scoring reliability that we believe are relevant to understanding the reliability of PSG data and provide a framework to consider both when interpreting the results of reliability studies and in designing multi-center research in anticipation of moving into the "Big Data" era.

Signal Acquisition: To achieve the potential promised by "Big Data" initiatives, physiological data from many sites need to be harmonized so data from different sources can be easily combined and compared. To achieve harmony when data are collected using different equipment across sites, there is an initial need to understand how signal acquisition influences assessment of the physiological outcome. Although identical equipment across sites may not be necessary to obtain consistent results, it is critical that the technical aspects of data acquisition (sensor properties, acquisition system sampling algorithms, storage, etc.) are defined in order to assure the output provides appropriate physiological measurement (in terms of fidelity, physiological sensitivity, etc.). Harmonization of data acquisition procedures is easiest when there are fixed and objectively applied criteria for signal acquisition (and confirmatory calibration is feasible). This is exemplified in PSG recording of the electroencephalogram (EEG), electrooculogram (EOG) signals, and electrocardiogram (ECG). For these measures, the sensors (electrodes) are generally of a uniform nature and positioning, and sampling frequencies, filters, and signal amplitudes may be "set"/calibrated to community-wide standards that can be accommodated by commercially available acquisition devices regardless of brand.

It is a greater challenge to standardize collection of other signals including various pressures (e.g., "nasal pressure," positive airway pressure), thermistry, thoraco-abdominal excursion, oximetry, etc. Further challenges are introduced by brand differences in sensor properties. Although of lesser concern in studies performed at a single center where the same sensors/acquisition system are employed for all research subjects, it is a substantial consideration when data are collected at multiple sites using systems with different properties and algorithms. Moreover, even when data have been collected at a single center, extrapolation of the results to clinical practice may be limited by use of different systems in research and community-based clinical laboratories. It is also important to consider that the use of several new technologies, including those used in some home sleep test devices, fundamentally differ from traditional PSG sensors, potentially limiting harmonization.

Multicenter research data are most easily harmonized by proactively selecting study equipment, sensors, and recording montages that are similar if not identical across sites. This approach was employed by several large-scale multicenter or longitudinal studies[6-11] that used the same home-based signal acquisition system across sites, and studies were scored at a central sleep reading center. Another approach to multicenter studies may be to leverage existing equipment and expertise at various clinical facilities for research application, even when these facilities use different PSG systems. Such was the case for a recent multicenter pediatric sleep apnea trial where sleep studies were conducted in seven academic sleep centers using equipment that was non-homogeneous across the centers.[12] To ensure consistency in data collection, a common protocol was developed and all sites underwent site visits and training and were asked to modify sensors and montages to ensure inter-site comparability. Research data collection only began after each "bed" was certified to provide a standard output. With close monitoring of data quality, comparable data were obtained from each site over a 5-year period. This experience supports the feasibility of using varying equipment to provide data for central scoring using a single platform, as long as appropriate attention is paid to certification procedures and quality control.

Scoring Reliability: While inter-laboratory differences in sensors and signal acquisition algorithms are a concern, as recently noted[13,14] one of the greatest barriers to standardization is the visual and largely subjective nature of sleep stage and event identification that characterizes current PSG. Although there are published scoring guidelines,[15] identification and categorization of specific sleep attributes (e.g., sleep stage, hypopneas, obstructive vs. "central" pattern) are done by visual inspection, usually without the aid of tools that quantify the signals to ensure that the defining criteria are met. This forces reliance on the scorer's visual perception and judgment. Both will be influenced by training, experience, and focus/distraction, all of which may vary across scorers and within scorers over time. The issue may be compounded in multicenter research when scoring standardization has not been achieved across different signal acquisition, processing equipment and scoring software across the sites.

The American Academy of Sleep Medicine (AASM) has recognized the importance of standardized, reliable PSG scoring for clinical applications.[16] Sleep laboratory accreditation now requires demonstration of scorer reliability. To facilitate this, the AASM has developed an interscorer reliability (ISR) program. The results of this program with respect to visual sleep stage scoring were recently reported.[13] Over 2,500 technicians from accredited laboratories scored a set of studies using a common visual display of test epochs. Overall, they showed 83% agreement for epoch-by-epoch staging; however, more disagreement was observed for identification of

stages N1 and N3 and for identification of sleep stage transitions. This exercise attests to the abilities of the trained scorers, unblinded to the test condition, to accurately recognize sleep stages using one visual display. However, while the ISR program is a needed, welcome, and in its context a successful innovation, it may overestimate reliability because: assessments were made during explicit scoring evaluations (which may not represent typical work conditions); scorers used a common visualization display; and statistical assessment did not adjust for "chance" agreement.

The article by Magalang and colleagues[3] examined interscorer reliability of nine scorers, each located at different sites. Scorers were asked to use locally available PSG software to score the same test set of 15 PSGs collected at a single site and converted to European Data Format (EDF) files. Thus, the reported reliability is likely a reflection of both interscorer as well as across-site factors, including software differences. Scorers had at least 5 years' experience and participated in a minimal protocol-specific training exercise (reviewed a slide set and manual). The highest interscorer reliability index was observed for the oxygen desaturation index (ODI), a metric that was (as is the usual case) automatically generated and therefore does not require scorer interaction (intraclass coefficient, ICC: 0.97). The apnea-hypopnea index (AHI) scored using the 2007 recommended AASM definition of hypopnea also was highly reliable (ICC: 0.95), likely benefiting from the ease of identifying events when linked to clear desaturations. Lower levels of reliability were observed for other metrics, including respiratory disturbance subtypes (central events, obstructive events, and mixed events), some of which varied by as much as 5-fold in magnitude across sites, as well as sleep stage agreement, which showed an ICC of only 0.21 for stage N1, which is considerably lower than the reliability reported by Rosenberg (63% level of agreement). In addition, reliability metrics were lower than what has been reported for centrally scored data representing a wide range of signal quality from studies collected in an unattended setting.[7,17]

Kuna and colleagues[4] in this issue of *SLEEP* evaluated intrasite as well as across-site scoring reliability; they also examined how reliability varied with hypopnea definitions.[4] In their study, 70 research PSGs selected to meet high quality standards from a single site were exported in EDF and were manually scored by 2 technicians at each of 5 sites (using various computer-assisted scoring systems). The scorers were at AASM accredited laboratories and had at least 4 years' experience but did not undergo common training for this protocol. Intrasite scorer reliability was determined by comparing the output of the two scorers at each site; across-site reliability was assessed by evaluating the level of agreement across sites for the average scored output of each site's paired scorers. The highest level of agreement was for an automatically generated metric, the average SpO_2 (ICC: 0.99). In this sample of studies with generally low AHI levels (mean 7.4), very high reliability also was observed for the overall AHI (ICC: 0.98). However, large interscorer variability was observed for hypopnea identification, which varied in magnitude by 2-fold from the scorer with the lowest compared to highest scored number. An important contribution was made by Kuna and colleagues[4] in showing that scoring done locally across centers yielded reliability in hypopnea scoring that was dependent on the criteria employed to define these events. Lower agreement was observed for the AASM alternate hypopnea definition (ICC: 0.73) and the Chicago Criteria (ICC: 0.53)[18] compared to the 2007 AASM recommended definition (ICC: 0.84). As in the study by Magalang et al.,[3] relatively low levels of agreement were observed for respiratory event subtypes (e.g., central apneas), arousals, and some sleep stage variables, with the ICC for N3 only 0.40.

Thus, the results of the studies by both Kuna et al.[4] and Magalang et al.[3] in this issue of *SLEEP* support the feasibility of localized scoring of signals collected using the same system and analyzed using disparate software by an experienced scoring staff when the primary study metric is the overall AHI. However, the lower reliability for other measures suggest that this approach may be suboptimal when study outcomes include event subtypes or stage distributions, where significant misclassification may be introduced by use of disparate software and a distributed scoring team. These studies also do not address the possible "drift" over time in scoring, a critical concern in clinical and epidemiological research. The poorer reliability for hypopneas detected using current scoring criteria compared to 2007 standard criteria by some laboratories also point to the need for training and quality control to ensure that technicians are appropriately trained to consistently identify events using discernible amplitude criteria, arousal identification, and lesser degrees of desaturation.

The variable levels of reliability for some indices, even by experienced technicians working in accredited laboratories, also raises the need to consider more objective approaches to generate PSG metrics. In the current issue, Malhotra and colleagues hypothesized that application of a computerized scoring algorithm would lead to acceptable scoring reliability.[5] Using the average of the manually scored indices from the 70 studies reported by Kuna et al.[4] in this issue of *SLEEP*, the investigators reported that a proprietary scoring program achieved scoring reliability comparable to what was observed between human scorers. Notably, the ICC for a summary of the manual versus automated scored AHI was remarkably high (0.97). Reliability coefficients also were moderate to high for most other PSG parameters, and similar to the prior studies, were relatively low for N3 (0.47) and lowest for REM arousal index (0.39). Systematic misclassification of some PSG parameters also was observed (e.g., stage N3% [59% vs 30.4% for automated vs. manually scored; REM AHI [22.7 vs 36.2]). Although this report supports the potential utility of automatic scoring to provide objective and reproducible results, there needs to be future consideration of not only reliability but validity. This begs the question of what is the gold standard to which automated scoring should be held.

When considering the generalizability of the findings reported in this issue, it is important to note that the data reported by both Kuna et al.[4] and Malhotra et al.[5] were from a single research study using home-based PSGs (level II) with a low average AHI (7.4). Scoring reliability was not assessed from PSGs integrated into the normal work flow of scorers, and thus, the "testing" conditions may have altered scoring practices. Issues related to pediatric scoring, which can be quite complex due to subtle changes in breathing, high arousability, low desaturation events, and high levels of EEG slow wave activity

remain to be addressed, as does reliability of level III home sleep testing systems. Although the impact of changing criteria for amplitude and desaturation were appropriately considered for hypopnea identification, future efforts need to consider possible additional variability associated with the co-occurrence of limb movements and cortical or subcortical arousals.

Magalang et al.[3] and Kuna et al.[4] have done the clinical research field a service by objectively testing the reliability of a scoring paradigm that in concept would be more cost-efficient for multicenter research than the central reading station paradigm that has been used in several large studies of the last two decades. On the positive side, these studies suggest that the AHI as defined using the AASM 2007 standard definition can be visually scored by experienced technicians with high interscorer reliability across laboratories using different scoring software. Automatically derived outputs, such as levels of oxygen desaturation, may be even more reliably measured. As discussed by these investigators, local scoring is logistically simpler than central scoring and may expedite the initiation of cross study protocols, including those occurring internationally. These investigators also appropriately identify the challenges in exporting data into EDF format to allow centralized scoring using a single platform and the complexities associated with electronic transmission of data to a central reading center. The latter concerns are ones that have generated important discussions among academics, clinicians, and representatives from industry. Fortunately, new open source software has been developed that markedly eases the ability to normalize and de-identify EDF header information (circumventing inconsistencies in local EDF export routines and removing protected health information). It also supports signal analysis and should facilitate the transfer of such files across sites (http:\\sleep.partners.org\EDF). Secure file transfer protocols are now used routinely and are amenable to transfer of even large PSG files.

In further considering the role of local laboratories in multicenter research, it is important to recognize that the current articles show that even in accredited sleep laboratories, scoring reliability is less than desirable for respiratory event subtypes, some sleep stage determinations, and stage-specific AHI. Of particular concern is the low reported reliability for N3, a feature of sleep that has been associated with numerous physiological functions, including hypertension incidence.[19] Even modest levels of measurement error (shown by ICCs below 0.60), can attenuate the strength of the observed relationship among variables of interest; more severe measurement errors may dramatically reduce the likelihood of observing associations of interest.[20] For example, when the true correlation between a sleep variable and a health outcome is strong as noted by an r = 0.70, but when the variables are measured with moderate error (measurement r = 0.40), the observed correlation will only be r = 0.28. In contrast, if the same variables were more reliably measured (measurement r = 0.80), the observed correlation doubles (r = 0.56). This bias to the null could have devastating effects on otherwise well-designed research protocols.

Is the high reported between-lab reliability for the ODI and AHI sufficient justification to use local scoring approaches for multicenter research? If the sole research outcome is the overall AHI, the answer may be yes, especially if rigorous training of scorers is conducted and ongoing performance is monitored. However, an exclusive focus on the overall AHI will limit the research study's ability to evaluate more specific relationships and sleep disorder phenotypes, including those which may be genetically determined or linked to health outcomes. In contrast, central scoring using a single software platform can achieve high levels of reliability for both sleep and breathing-related parameters, including the AHI derived using the current AASM hypopnea definition.[17] However, as described before,[9] high levels of scoring reliability for the range of PSG parameters requires quality control procedures that are likely more intensive than are typical in today's clinical settings and are much more intense than those described in the reports in this issue. Whether higher levels of reliability than those reported in the current studies can be achieved with local scoring through more intense scorer training, scorer monitoring, and feedback; use of a single scoring platform (as was the case in the large multicenter studies that employed central scoring) possibly accessed through a web portal; or with use of a wider range of test PSGs are important areas to consider in future research. If the wealth of data embedded within the PSG is to be used optimally either clinically or for research, additional effort is needed to improve reliability in more routine settings other than found in central sleep reading centers. Furthermore, given the richness of the varied, dynamic physiological information recorded over hours within the PSG, there are important opportunities to validate and disseminate objective, quantitative analyses software programs, such those that compute power spectra of the EEG and ECG, providing potentially richer information in regard to sleep architecture and heart rate variability than is generated by either binary classification of sleep stages or by summary measures of heart rate and sleep stage distributions. An advantage of central scoring of data may be that this approach can facilitate the archival of harmonized, de-identified sets of physiological signals, which also could be linked to national databases of genetic, imaging, and clinical data—a goal consistent with "Big Data" initiatives.

The articles by Magalang et al.,[3] Kuna et al.,[4] and Malhotra et al.[5] make important contributions to understanding scoring reliability and the promise for improvement in the setting of multicenter research. In addition, it is essential to note that while reliability of reporting is critical, scoring reliability does not address whether the most physiologically relevant variables are being collected. This remains one other critical direction in which to proceed as we move forward in multicenter sleep research. Ultimately, the decision of how to structure a multicenter study, and whether to use local or central scoring resources, needs to balance costs and resources and study needs and objectives (e.g., scope of outcomes).

Whether the reliability of PSG scoring in research and clinical settings can be improved through use of automated systems remains to be seen but clearly warrants exploration. Automated scoring is a promising method to both improve reliability as well to reduce cost and facilitate exploration of better disease detection and outcomes. However, as noted above, more research is needed to determine how such systems should be validated as well as how they should be operationalized within and across investigational studies in adult and pediatric populations. Additional potential advantages of validated automated systems include the ability to reprocess original signals using

different settings to test whether different definitions of events provide better or greater insights into associations and causal relation to outcomes of interest; and objectivity. These features are attractive as we move into the "Big Data" era. However, if automated scoring is to be used in research, the algorithms used to define and measure variables should be disclosed; the comparability of collecting systems (in their aggregate, including inputting sensors and signal processing and properties of the acquisition systems) across the field sites should be documented; methods for standardizing different systems across sites should exist; and if automated systems of different brands are used within the same study, comparability data should be presented.

Despite these challenges, there are enormous opportunities for Sleep Medicine to make important contributions in Big Data initiatives. Embracing the rich multidimensional data that comprise the foundation for Sleep Medicine, while working to improve the consistency in collection and scoring, and developing well-defined data archives accessible to the community that can be used for a wide range of purposes, including quantitative signal analysis, will pave the way for sleep data to be integrated into the powerful databases of genomic, imaging and health outcome data that promise to transform science and health.

CITATION

Redline S; Dean D; Sanders MH. Entering the era of "big data": getting our metrics right. *SLEEP* 2013;36(4):465-469.

DISCLOSURE STATEMENT

Susan Redline has received grant support from ResMed Inc. and ResMed Foundation and use of equipment from ResMed Inc. and Philips-Respironics for research studies. She is the Director of the Brigham and Women's Hospital Sleep Reading Center which has received NIH grant support to oversee central sleep analyses for the Sleep Heart Health Study, the Outcomes of Sleep Disorders in Older Men Study, the Study of Osteoporotic Fractures In Women Sleep Study, Honolulu Asian American Sleep Apnea Study, the Childhood Adenotonsillectomy Trial, the Heart Biomarkers in Apnea Treatment Study, the Hispanic Community Health Study, Jackson Heart Study, Multiethnic Study of Atherosclerosis, the Nulliparous Pregnancy Outcome Monitoring Study Monitoring Mothers To Be, and the Starr County Health Study. She is a member of the Board of Directors for the AASM. She also has a pending NIH grant to develop a national sleep research resource comprised of signals collected from many research studies. SR is in the same academic department/division as Dr. Atul Malhotra but had not provided input in the design, execution, analysis or interpretation of his study. Mark H. Sanders is a Scientific Consultant to Philips-Respironics and receives compensation; he is not serving as an advisor on the products reported in the articles by Malhotra et al., Kuna et al. and Magalang et al. that are the subject of this Commentary. He is a Field Editor for Sleep Medicine and receives compensation; Until recently he was Editor-in-Chief - Pulmonary, Critical Care, and Sleep Medicine and Editor - Sleep Medicine for UpToDate for which he received compensation; He is a Deputy Editor for SLEEP for which he doesn't receive compensation. Dr. Dean has indicated no financial conflicts of interest.

REFERENCES

1. Francis Collins, PhD, Director, National Institutes Health: http://nihrecord.od.nih.gov/newsletters/2013/02_01_2013/story5.htm
2. Luyster FS, Strollo PJ Jr., Zee PC, Walsh JK. Sleep: A health imperative. Sleep 2012;35:727-34.
3. Magalang UJ, Chen NH, Cistulli PA, et al. Agreement in the scoring of respiratory events and sleep among international sleep centers. Sleep 2013;36:591-6.
4. Kuna ST, Benca R, Kushida CA, et al. Agreement in computer-assisted manual scoring of polysomnograms across sleep centers. Sleep 2013;36:583-9.
5. Malhotra A, Younes M, Kuna ST, et al. Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. Sleep 2013;36:573-82.
6. Foley DJ, Masaki K, White L, Larkin EK, Monjan A, Redline S. Sleep-disordered breathing and cognitive impairment in elderly Japanese-American men. Sleep 2003;26:596-9.
7. Mehra R, Stone KL, Blackwell T, et al. Prevalence and correlates of sleep-disordered breathing in older men: the MrOS Sleep Study. J Am Gerontol Soc 2007;55:1356-64.
8. Redline S, Aylor J, Clark K, Graham G, Richardson S. Predictors of longitudinal change in the respiratory disturbance index. Am Resp Crit Care Med 1998;157:61.
9. Redline S, Sanders MH, Lind BK, et al. Methods for obtaining and analyzing unattended polysomnography data for a multicenter study. Sleep Heart Health Research Group. Sleep 1998;21:759-67.
10. Sorlie PD, Aviles-Santa LM, Wassertheil-Smoller S, et al. Design and implementation of the Hispanic Community Health Study/Study of Latinos. Ann Epidemiol 2010;20:629-41.
11. Yaffe K, Laffan AM, Harrison SL, et al. Sleep-disordered breathing, hypoxia, and risk of mild cognitive impairment and dementia in older women. JAMA 2011;306:613-9.
12. Redline S, Amin R, Beebe D, et al. The Childhood Adenotonsillectomy Trial (CHAT): rationale, design, and challenges of a randomized controlled trial evaluating a standard surgical procedure in a pediatric population. Sleep 2011;34:1509-17.
13. Rosenberg RS, Van Hout S. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. J Clin Sleep Med 2013;9:81-7.
14. Penzel T, Zhang X, Fietze I. Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules. J Clin Sleep Med 2013;9:89-91.
15. Redline S, Budhiraja R, Kapur V, et al. Reliability and validity of respiratory event measurement and scoring. J Clin Sleep Med 2007;3:169-200.
16. Silber MH, Ancoli-Israel S, Bonnet MH, et al. The visual scoring of sleep in adults. J Clin Sleep Med 2007;3:121-31.
17. Whitney CW, Gottlieb DJ, Redline S, et al. Reliability of scoring respiratory disturbance indices and sleep staging. Sleep 1998;21:749-58.
18. Flemons WW, Buysse D, Redline S, et al. 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;22:533-70.
19. Fung MM, Peters K, Redline S, et al. Decreased slow wave sleep increases risk of developing hypertension in elderly men. Hypertension 2011;58:596-603.
20. Charles EP. The correction for attenuation due to measurement error: clarifying concepts and creating confidence sets. Psychol Methods 2005;10:206-26.