Sleep Medicine 65 (2020) 161–169

Contents lists available at ScienceDirect

Sleep Medicine

journal homepage: www.elsevier.com/locate/sleep

Original Article

Quantifying statistical uncertainty in metrics of sleep disordered breathing

Robert J. Thomas [a, b], Shuqiang Chen [c], Uri T. Eden [d], Michael J. Prerau [a, e, f, *]

[a] Harvard Medical School, USA
[b] Pulmonary, Critical Care & Sleep, Department of Medicine, Beth Israel Deaconess Medical Center, USA
[c] Department of Mathematics and Statistics, Boston University, USA
[d] Boston University, USA
[e] Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, USA
[f] Department of Anesthesia, Critical Care, and Pain Medicine, Massachusetts General Hospital, USA

ARTICLE INFO

Article history:
Received 3 May 2019
Received in revised form
3 June 2019
Accepted 5 June 2019
Available online 13 June 2019

Keywords:
Sleep apnea
Apnea variability
Confidence interval
Apnea diagnosis
AHI
RDI

ABSTRACT

Background: The apnea-hypopnea index (AHI) (or one of its derivatives) is the primary clinical metric for characterizing sleep disordered breathing—the value of which with respect to a threshold determines severity of diagnosis and eligibility for treatment reimbursement. The index value, however, is taken as a perfect point estimate, with no measure of statistical uncertainty. Thus, current practice does not robustly account for variability in diagnosis/eligibility due to chance. In this paper, we quantify the statistical uncertainty associated with respiratory event indices for sleep disordered breathing and the effect of uncertainty on treatment eligibility.

Methods: We develop an empirical estimate of uncertainty using a non-parametric bootstrap on the interevent times, as well as a theoretical Poisson estimate reflecting the current formulation of the AHI. We then apply these methods to estimate AHI uncertainty for 2049 subjects (954/1095 M/F, age: mean 69 ± 9.1) from the Multi-Ethnic Study of Atherosclerosis (MESA).

Results and Conclusions: The mean 95% empirical confidence interval width was 11.500 ± 6.208 events per hour and the mean 95% theoretical Poisson confidence interval width was 5.998 ± 2.897 events per hour, suggesting that uncertainty is likely a major confounding factor within the current diagnostic framework. Of the 278 subjects in the symptomatic population (ESS>10), 27% (76/278) had uncertain diagnoses given the 95% empirical confidence interval. Of the 2049 subjects in the full population, 43% (880/2049) had uncertain diagnoses given the 95% empirical confidence interval. The inclusion of subjects with uncertain diagnoses increases the number of eligible patients by 21.3% for the symptomatic population and by 84.8% for the full population. The exclusion of subjects with uncertain diagnoses given the 95% empirical confidence interval decreases the number of eligible patients by 12.4% for the symptomatic population and by 34.8% for full population. Additional analyses suggest that it is practically infeasible to gain diagnostic statistical significance through additional testing for a broad range of borderline cases. Overall, these results suggest that AHI uncertainty is a vital additional piece of information that would greatly benefit clinical practice, and that the inclusion of uncertainty in epidemiological analysis might help improve the ability for researchers to robustly link AHI with co-morbidities and long-term outcomes.

1. Introduction

Sleep-disordered breathing impacts at least 10% of the adult population, with increasing prevalence in those with certain comorbid conditions (eg, atrial fibrillation, congestive heart failure, morbid obesity and advanced type II diabetes) where it may approach 50% [1–3]. The apnea-hypopnea index (AHI), or one of its derivatives such as the respiratory disturbance index (RDI), is the key statistic used to quantify the presence and severity of sleep apnea [4,5]. The AHI is used to assess the effects of therapy, such as continuous positive airway pressure (CPAP) devices, both in the sleep laboratory and within the home. The division of the AHI into

* Corresponding author. Brigham and Women's Hospital, Division of Sleep and Circadian Disorders, 221 Longwood Ave, Suite 231, Boston, MA 02115, USA.
E-mail address: mprerau@bwh.harvard.edu (M.J. Prerau).

https://doi.org/10.1016/j.sleep.2019.06.003
1389-9457/

obstructive and central components also determines the type of approved therapy, such as the far more expensive adaptive servo ventilators (ASV's) for central sleep apnea [6,7]. Additionally, reductions in the AHI (by 50%) are also used by the Food and Drug Administration to approve novel therapies for apnea, including Provent™, Winx™ [8], hypoglossal nerve stimulation [9] and phrenic nerve stimulation [10].

In the United States, clinical apnea severity and therapy decisions are determined using AHI thresholds: Normal: AHI < 5, Mild: $5 \leq$ AHI < 15, Moderate: $15 \leq$ AHI < 30 and Severe: AHI ≥ 30 [11]. The current standards, usually based on a single night of assessment, are to treat when the AHI ≥ 15 or, AHI ≥ 5 if symptomatic (eg, reported daytime sleepiness) or associated comorbidities (eg, atrial fibrillation). The threshold values are historical in origin, derived from work by Guilleminault and Dement in the mid 1970s, in which the data from 200 healthy subjects was collected for the purpose of serving as controls for future studies on dreaming or pharmacological intervention [12]. Given that the respiratory event rates of this cohort over the night tended to fall below 5 events/hour, the authors deemed 5 to be an adequate anomaly criterion for subsequent studies [13]. This ad-hoc convention based on normative data made its way into the first edition of The International Classification of Sleep Disorders (ICSD-1) and, through historical inertia, ended up as the criteria for reimbursement.

Use of any hard threshold on a clinical decision-making metric requires careful consideration of the statistical properties of that metric, the variability of the measured phenomena, and the implications for treatment in borderline cases. Thus, it is vital to understand AHI uncertainty to know if it is reasonable to intervene for an AHI of 5.1, while denying treatment for an AHI 4.9, or if indeed there is enough information to differentiate between the two conditions. Currently, no measure of uncertainty is standard for AHI, which is treated as a point estimate—a single value that perfectly describes an underlying property of the disorder. Consequently, it has been difficult to answer any of the aforementioned questions within a statistically principled framework.

The AHI characterizes apneas and hypopneas, which are governed by inherently dynamic processes, using an average rate. The AHI has substantial night-to-night variability, with large inter-individual differences [14–18] and rare negative reports [19]. One component of this variability is due to numerous intrinsic and extrinsic factors including sleep architecture, body position, sleep state, sleep stage, time of night effects, and fluid retention [14,15]. Moreover, during “split night” studies and therapy, sources of variability include adaptation to the mask and air pressure [20,21]. The interaction between these factors is complex and remains an open experimental question.

It is, however, far more straight-forward to characterize the statistical uncertainty of an AHI estimate related to sampling error and the distributional properties of the statistic. In computing uncertainty, we assume that there is a ground truth endogenous average apnea rate for a patient for a given night, which we could observe accurately given an infinite amount of data. However, as there is a limited duration of time in which to observe a sleeping patient, the AHI we observe may differ from that endogenous rate due to the random nature of the events unfolding during sleep on any particular night. This is akin to not observing exactly 50% heads and 50% tails in a fixed set of tosses of a fair coin. Thus, it is possible for patients to appear to be above or below threshold due only to variability attributable to uncertainty in the statistic. By measuring AHI uncertainty using an appropriate confidence interval, it is possible to place a diagnosis within the proper context based on the data provided.

In this paper, we develop two estimators for AHI uncertainty. The first is a conservatively small estimate that reflects the theoretical uncertainty of the current AHI statistic, based only the number of respiratory events (N) and the total sleep time (TST). The second takes advantage of the timing information from a subject's scored events to provide a more accurate empirical estimate of the statistical uncertainty surrounding respiratory event rate. This empirical method uses a non-parametric bootstrap to provide an estimate of a distribution of the event rate. Using these methods, we are able to compute confidence intervals about a given AHI estimate from partial, single, or multiple nights, as well as determine the confidence with which a patient's AHI differs from a clinical threshold.

We then apply these methods to experimental data from a large polysomnographic (PSG) database, demonstrating the implications of AHI uncertainty for individual patients, as well as across a large population. In doing so, we explore the prevalence of patients for whom diagnosis is currently uncertain due to statistical variability, and the implications of uncertainty on treatment eligibility. Overall, this work highlights the importance of incorporating AHI uncertainty into clinical decision-making.

2. Materials and methods

2.1. Statistical uncertainty under the AHI is a theoretical Poisson model

The AHI is defined as

$$AHI = \frac{N}{TST} \tag{1}$$

where N is the number of events observed within a time period and TST is the total sleep time within that same period [4,5]. The AHI relies on only the number of respiratory events within an epoch and is not a function of their specific timing, duration, or interrelation. By averaging over the TST, the AHI implicitly ignores any temporal dependence structure between apnea events.

Point processes are mathematical constructs [22] that can be used to model random processes with events that occur over time, such as the arrival of trains at a station, the spiking of neurons, or the timing of earthquakes. A Poisson process is a specific type of point process that is described by a single rate parameter and has the assumption that the event occurrences are independent of each other. Thus, a description of the apnea process using only N and TST and ignoring all other covariates or dynamic changes in rate is equivalent to an assumption of Poisson structure. This perspective enables us to construct a theoretical confidence interval around an AHI using the known statistical properties of Poisson processes. A detailed formulation for exact and approximate Poisson estimates can be found in Appendix.

Similarly, other common clinical indices for diagnosis of sleep disordered breathing such as the central apnea index (CAI) and respiratory disturbance index (RDI) as well as other metrics such as the periodic limb movement index (PLMI) have the same functional form as the AHI but use different values for the event counts and total time. Consequently, all these same concepts and metrics apply equally to these and other equivalent clinical indices.

While the Poisson model may not fully capture the full temporal structure of apnea dynamics, it is useful because it is theoretically-based with well-known statistical properties, it reflects the assumptions in the current formulation of the AHI, its confidence interval easy to compute, and it provides a robust lower bound on

the uncertainty. Moreover, since the Poisson model only requires only *N* and *TST*, it is useful for analyses of data sets where timing information for individual respiratory events is unreliable or unavailable.

2.2. Event timing information provides a more accurate, empirical estimate of uncertainty

Sleep disordered breathing is a dynamic process in which the rate of respiratory occurrence can be highly influenced by covariates, such as loop gain [23–25], stage or body position dominance [26–28] and overall instability of sleep architecture [29]. Additionally, event rate likely has history dependence, which means that the likelihood of a respiratory event is influenced by the timing of previous events. Thus, the Poisson model, while a direct analog to the data and assumptions of the AHI, will likely produce a confidence interval that is smaller than the actual variability underlying endogenous apnea rate process.

Given the actual event times, however, we can compute an empirical confidence interval using a bootstrap scheme, which estimates uncertainty using repeated resampling with replacement to generate a distribution of event rates given the properties of the observed data. This approach uses the distribution of the interevent intervals themselves to provide a more realistic estimate of the statistical variability in AHI estimate given the properties of a given subject.

Given the scored respiratory event times, computed interevent intervals, and total sleep time from a given subject, we follow the following procedure:

1. For many iterations repeat:
   a. Sample with replacement from the interevent intervals repeatedly until the sum of the sampled intervals first equals or exceeds the total sleep time
   b. Compute and store the average event rate for that generated sample (number of events sampled in 1a./TST)
2. Estimate the confidence interval at a given critical value by computing the corresponding percentiles of the estimated event rates from all of the iterations

In this study, we use 3000 iterations and the 2.5th and 97.5th percentiles to estimate a 95% confidence interval on the data. To estimate the probability of an estimate being above/below a clinical threshold, we use the proportion of samples above/below the threshold.

2.3. Population analysis

To examine the effects of index uncertainty in large populations, we apply these methods to data from the Multi-Ethnic Study of Atherosclerosis (MESA) [30,31], obtained from www.sleepdata.org (the National Sleep Research Resource [32,33]). After excluding subjects with incomplete data records, we analyze 2049 subjects (954/1095 M/F, age: mean 69 ± 9.1). Subjects with an Epworth Sleepiness Scale (ESS) of greater than 10 are considered as being within the symptomatic population.

For the AHI analysis, we compute the number of counts related to AHI using all apneas and hypopneas associated with a ≥3% oxygen desaturation, accordingly. Subjects with an AHI ≥ threshold (5 symptomatic, 15 full population) are used to replicate the current clinical standard. For each subject, the 95% exact Poisson and empirical bootstrap confidence intervals are computed around the AHI statistic. Subjects with uncertain diagnoses are defined as having a 95% confidence interval containing the clinical threshold.

2.4. Estimating empirical intervals from Poisson intervals

We use the population data to characterize the relationship between the Poisson and empirical confidence intervals. In doing so, we are able to generalize the results from the population bootstrap analysis to estimate the empirical confidence interval for any given N and TST.

To these ends, we perform the following regression:

$$UB_{emp} = \beta_{upper} UB_{Poisson} + \varepsilon \tag{2}$$

$$LB_{emp} = \beta_{lower} LB_{Poisson} + \varepsilon \tag{3}$$

where UB_{emp} and LB_{emp} are the upper and lower empirical bounds from the bootstrap procedure, $UB_{Poisson}$ and $LB_{Poisson}$ are the exact upper and lower Poisson bounds, ε is the noise term, and the β parameters are estimated scaling factors.

Given *N*, *TST*, and β we can then compute the *estimated empirical confidence interval*, $[\beta_{upper} UB_{Poisson}, \beta_{lower} UB_{lower}]$, which maps the theoretical Poisson results to empicial bounds, so as to reflect variability observed in the population data. We also computed the correlation level between the empirical and Poisson bounds to determine the accuracy of the estimated empirical intervals.

2.5. Computing minimum testing time to significance

Additional testing is commonly proposed as an approach for resolving diagnoses for patients on threshold borderlines. The ability to construct a confidence interval on the AHI provides a quantitative means of exploring the practicality of this approach.

As the amount of data observed increases, the size of the statistical confidence interval decreases. Hence, the more sleep time observed, the smaller the uncertainty on the AHI will become. For a given endogenous apnea rate, it is therefore possible to figure out how much testing time is required by finding the N and TST for a given AHI at which the threshold falls outside the corresponding 95% estimated empirical confidence interval.

3. Results

3.1. Population confidence interval estimates

For each subject in the MESA population, we computed the 95% confidence intervals from empirical bootstrap and exact Poisson methods. The mean empirical confidence interval width was 11.500 ± 6.208 events per hour and the mean theoretical Poisson confidence interval width was 5.998 ± 2.897 events per hour. From a clinical standpoint, these are large values with respect to the diagnostic thresholds, suggesting that uncertainty is likely a major confounding factor within the current diagnostic framework.

We next computed the relationship between the theoretical and empirical upper and lower bounds using the regression analysis. The parameter estimates show that, within this population, the empirical lower bound is 93% of the Poisson lower bound ($R^2 = 0.9497$, $p < 2e\text{-}16$) and the empirical upper bound is 116% of the Poisson upper bound ($R^2 = 0.9909$, $p < 2e\text{-}16$), suggesting that the empirical bounds are consistently larger than the Poisson bounds. The goodness-of-fit metrics strongly suggest that the estimated empirical intervals based only on N and TST can be used to accurately predict the AHI uncertainty in the absence of subject event data. Thus, theoretical results derived from the Poisson model can be generalized to reliably reflect the empirically variability observed in this population.

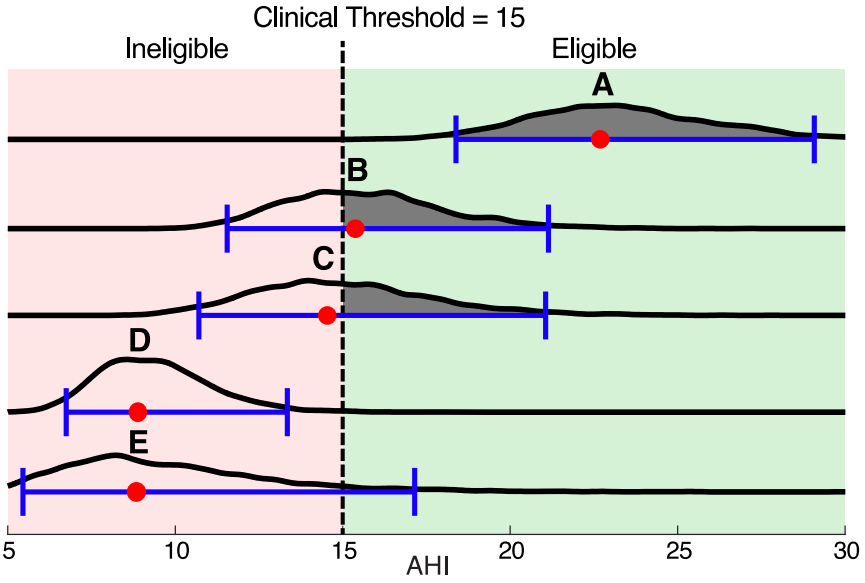
3.2. AHI uncertainty in individual patient diagnoses

To understand the implications of the AHI confidence interval on individual patient diagnosis and eligibility, we examined several individual subjects from the MESA population, shown in Fig. 1 and its associated table. For each of these asymptomatic subjects (ESS<10), we used the bootstrap procedure to compute the distribution (black curves) around the AHI (red dots) given the subject's observed event times. From these distributions, we then computed for each subject the 95% empirical confidence interval (blue bounds). Additionally, we computed probability of the AHI being above the diagnostic threshold of 15 (gray shaded region), which relates to the chance of the subject being declared eligible (above threshold) on subsequent retests.

Subject A has an AHI of 22.7 and is thus eligible for treatment, as the AHI is above the threshold of 15. The lower bound of the 95% empirical confidence interval is also above the threshold, and the probability of an above-threshold AHI is ~1.00. It is therefore highly unlikely that a subject with an endogenous AHI below the threshold would produce these data. Likewise, Subject D's AHI of 8.9 and upper 95% confidence interval are both below the clinical threshold. This subject would be denied treatment and would not likely have produced the observed below-threshold AHI by chance, with a 1% probability of retesting above the threshold.

While Subjects A and D produce unambiguous diagnoses, other subjects are not nearly as clear. Subject B has an AHI of 15.4 and would be diagnosed with moderate apnea and qualify for treatment. While this subject's AHI is only 0.4 away from the threshold, the 95% confidence interval width is 9.9 (events/hour) wide, with bounds spanning both sides the clinical threshold. Moreover, since the probability of an above-threshold AHI is 0.59 there is a 41% they would retest below-threshold on subsequent tests. Subject C has an observed AHI of 14.5 and would be diagnosed as having mild apnea and denied reimbursable treatment. However, while the subject's AHI is 0.5 away from the threshold, the 95% confidence interval width is 10.7, and there is a 47% chance that the subject would qualify for treatment on retest. For both of these subjects, the 95% confidence interval widths are roughly 20 times the distance from the observed AHI to the threshold, and the probabilities of being eligible are close to 0.5. The AHI distributions show us that it is not only unclear if their endogenous apnea rates differ from the threshold, but also if they differ from each other. Given these factors, there does not seem to be sufficient evidence in the data to deny treatment to one subject while granting it to the other.

The impact of AHI uncertainty is further accentuated in cases such as split night studies where the total time may be short. As a concrete example, Subjects E and D share a similar AHI, but D had 7.3 h of sleep while E had only 3.7 h. Subject E's short sleep time results in a confidence interval almost double the size of Subject D's, covering both Mild and Moderate categories, with a 7% chance of being above threshold on retest. These examples illustrate that extra attention must be taken to any diagnoses based on short time periods. The identical concepts are relevant when estimating severity of disease from a split night assessment. Thus it is clear that the duration of sleep recorded can markedly change the confidence intervals of severity, and impact clinical decision-making.



Name	N	TST	AHI	CI Width	Prob. AHI > 15
Subject A	159	7.0	22.7 {18.4, 28.9}	10.5	1.00
Subject B	108	7.0	15.4 {11.5, 21.4}	9.9	0.59
Subject C	106	7.3	14.5 {10.5, 21.2}	10.7	0.47
Subject D	65	7.3	8.9 {6.8, 13.5}	6.7	0.01
Subject E	33	3.7	8.8 {5.4, 17.1}	11.7	0.07

Fig. 1. AHI uncertainty for illustrative subjects from the MESA dataset. For each subject, we show the AHI (red dot), the full distribution of the AHI given the bootstrap procedure (black curves), the 95% confidence interval (blue bounds), and the portion of the distribution above the threshold (gray shaded area). While Subjects A and D will likely have endogenous apnea rates that differ from the clinical threshold of 15, Subjects B and C cannot be statistically differentiated from the threshold or from each other. Thus, there is not enough information to deny C treatment while allowing it as an option for B. This problem is exacerbated for patients with short sleep durations (e.g. split night studies), like Subject E, as the AHI confidence interval increases precipitously as TST reduces.

It is straightforward to apply these concepts to patient report data. Table 1 shows an example clinical report for an asymptomatic subject from the MESA dataset. In addition to the standard AHI values, we provide 95% empirical confidence intervals around the AHI and probability of the AHI being above threshold on retest. Similarly, we can compute the confidence interval for stage-dependent AHI empirical confidence intervals and above-threshold probabilities. In this particular case, the subject's AHI values do not qualify them for treatment, however the upper confidence intervals are far above-threshold and retest probabilities are high. This suggests that there is not enough data to conclusively deny eligibility.

Generalizing beyond specific patient examples, we can identify the conditions in which uncertain diagnoses will be likely. Fig. 2 shows a chart identifying values of N and TST for which the estimated empirical 95% confidence interval contains a clinical threshold (gray regions), with analogous Poisson interval regions shown as dotted lines. These regions represent values of N and TST for which a clinical diagnosis would be uncertain given the empirical variability estimated from the population regression analysis. Furthermore, these regions of AHI uncertainty provide objective definitions for intermediary states such as Mild/Moderate or Moderate/Severe apnea, given the current standard.

3.3. Online AHI uncertainty calculation tool

The interaction between N, TST, AHI uncertainty and diagnosis severity can be explored interactively using an online calculation tool that we have developed.

This tool can be found at http://sleepEEG.org/AHI.

3.4. AHI uncertainty in population diagnoses

Overall, the case studies in Fig. 1 and chart in Fig. 2 clearly illustrate the high degree of statistical uncertainty in the AHI with respect to clinical thresholds, as well the effect of sleep duration on AHI confidence interval size. Moreover, we observe the potential ambiguity in clinical diagnoses inherent in the current formulation of the AHI. What then is the effect of uncertain diagnoses within a large population?

The results of the AHI population analysis are summarized in Table 2. This analysis examines the 278 subjects in the symptomatic population (ESS>10), for which the AHI threshold is 5, and the 2049 subjects in the full population (no ESS restriction), for which the AHI threshold is 15. Table 2 shows results for the empirical method as well as the theoretical Poisson method for reference and as a lower bound.

We first identify subjects who would be considered eligible for treatment under the current criteria (AHI $\geq$ threshold), which does not consider uncertainty (Table 2, Currently Eligible). We then identify those subjects that have uncertain diagnoses (Uncertain Subjects Eligible), and identify the number of these subjects with AHIs above the threshold (Above Threshold) and the number of subjects with AHI below the threshold (Below Threshold). Using these values, we can explore the effect of different strategies for handling uncertain clinical diagnoses: granting eligibility to all uncertain cases (Uncertain Subjects Eligible) or denying eligibility to all uncertain cases (Uncertain Subjects Ineligible).

3.4.1. Identifying uncertain diagnoses

Of the 278 subjects in the symptomatic population (ESS>10), 27% (76/278) had uncertain diagnoses given the 95% empirical confidence interval. Of those uncertain subjects, 63% (48/76) are ineligible for treatment (below-threshold AHI) and 37% (28/76) would qualify (above-threshold AHI). Of the 2049 subjects in the asymptomatic population, 43% (880/2049) had uncertain diagnoses. Of those uncertain subjects, 71% (624/880) are currently ineligible for treatment and 29% (256/880) would qualify.

For the theoretical 95% Poisson confidence interval, 19% (52/278) of the symptomatic subjects had uncertain diagnoses, with 54% (28/52) ineligible for treatment and 46% (24/52) qualifying. For the asymptomatic population, 21% (435/2049) had uncertain diagnoses. Of those uncertain subjects, 58% (254/435) are currently ineligible for treatment and 42% (181/435) would qualify.

3.4.2. Including uncertain diagnoses in eligibility

Having identified the subjects with uncertain diagnoses, we can examine the effect of an inclusive eligibility clinical criterion in which all patients with uncertain diagnoses are eligible for treatment. To do so, we identify subjects for which the AHI upper 95% confidence interval is above or equal to the threshold. Patients A, B, C, and E from Fig. 1 would qualify under this criterion.

Given the 81% (225/278) eligible subjects in the symptomatic population (ESS>10), the inclusion of subjects with uncertain diagnoses (AHI 95% upper confidence bound $\geq$ 5) increases the eligibility by 21.3% (273 vs. 225) for the empirical bounds and by 12.4% (253 vs. 225) for the Poisson. For the 36% (736/2049) eligible subjects in full population (no ESS restriction), the inclusion of subjects with uncertain diagnoses increases the eligibility by 84.8% (1360 vs 736) for the empirical bounds and by 34.5% (990 vs 736) for the Poisson.

3.4.3. Excluding uncertain diagnoses

Similarly, we can examine the effect of an exclusive criterion in which all subjects with uncertain diagnoses are denied eligibility. We therefore identify those subjects for whom the lower AHI 95% confidence interval is above the threshold. Only Patient A from Fig. 1 would qualify under this criterion.

Given the eligible subjects in the symptomatic population (ESS>10), the exclusion of subjects with uncertain diagnoses (AHI 95% upper confidence bound $\geq$ 5) decreases the eligibility by 12.4% (197 vs. 225) for the empirical bounds and by 10.7% (201 vs. 225) for the Poisson. For the eligible subjects in full population (no ESS restriction), the exclusion of subjects with uncertain diagnoses decreases the eligibility by 34.8% (480 vs. 736) for the empirical bounds and by 24.6% (555 vs. 736) for the Poisson.

3.5. The effect of additional testing on AHI confidence

A common solution proposed for dealing with ambiguous sleep study data is to provide additional testing. In principle, this appears to be a very reasonable approach. The apnea rate that we observe

Table 1
Example clinical report with added confidence interval for AHIs. In this case, the additional information from the confidence interval suggest that there is not enough information to sufficiently motivate the denial of eligibility for treatment.

						95% Confidence Interval	Prob. AHI≥15
Total Sleep Time:	5.78 h	Total Events:	81	AHI:	14.03	[10.16,20.82]	0.39
NREM Time:	4.88 h	NREM Events:	72	NREM AHI:	14.21	[9.6, 21.96]	0.40
REM Time:	0.90 h	REM Events:	9	REM AHI:	13.08	[10.55, 25.78]	0.59

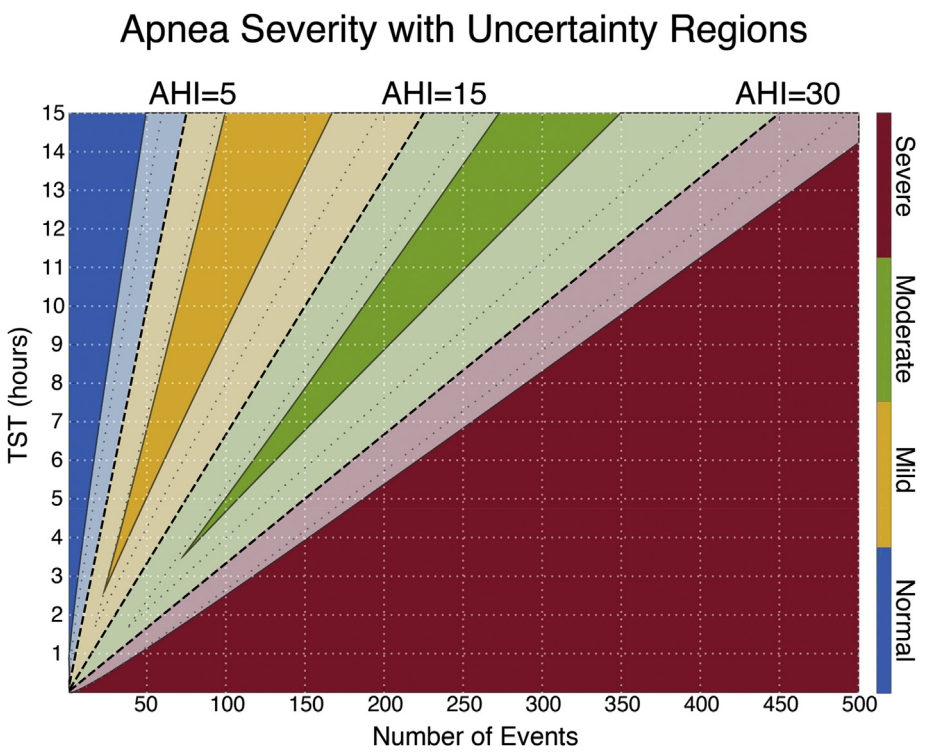


Fig. 2. Visualizing regions of uncertainty around clinical thresholds. This graphic illustrates the clinical thresholds for Mild (AHI≥5), Moderate (AHI≥15), and Severe (AHI≥30) apnea (dashed lines) as a function of number of events (N) and total sleep time (TST). The shaded regions show the values of (N, TST) at which the 95% bootstrap (gray regions) and exact Poisson (dotted lines) confidence bounds encompass a clinical threshold. These regions indicate conditions for which the count data alone does not provide enough information to determine the severity of a patient's apnea. Additionally, they provide an objective definition for intermediary categories of apnea severity (eg, Mild/Moderate, Moderate/Severe).

Table 2
Population analysis of apnea-hypopnea index (AHI) statistical uncertainty using data from the MESA study. For the 278 subjects in the symptomatic (ESS>10) population and the 2049 subjects in the full population, this table summarizes the number of subjects in the following classes: Currently Eligible – subjects qualifying for treatment under current standards (AHI ≥ Threshold), Uncertain Diagnoses – subjects for whom the threshold is contained within the upper and lower 95% confidence bounds on the AHI (Lower Bound < Threshold < Upper Bound)—separated into percent uncertain subjects above and below threshold, Uncertain Subjects Eligible – subjects qualifying if all uncertain subjects are eligible (Upper Bound ≥ Threshold), Uncertain Subjects Ineligible – subjects qualifying if all uncertain subjects are excluded (Lower Bound > Threshold). Population proportion percentages are reported with the exact binomial 95% confidence interval in curly braces.

3% Desaturation	Empirical		Poisson	
Symptomatic (ESS>10) Population: Threshold - AHI ≥ 5				
Currently Eligible	225/278	80.94% {75.82% 85.38%}	225/278	80.94% {75.82% 85.38%}
Uncertain Diagnoses	76/278	27.34% {22.19% 32.98%}	52/278	18.71% {14.30% 23.79%}
Below Threshold	48/76	63.16% {51.31% 73.94%}	28/52	53.85% {39.47% 67.77%}
Above Threshold	28/76	36.84% {26.06% 48.69%}	24/52	46.15% {32.23% 60.53%}
Uncertain Subjects Eligible	273/278	98.20% {95.85% 99.41%}	253/278	91.01% {87.01% 94.10%}
Uncertain Subjects Ineligible	197/278	70.86% {65.14% 76.14%}	201/278	72.30% {66.64% 77.48%}
Full Population: Threshold - AHI ≥ 15				
Currently Eligible	736/2049	35.92% {33.84% 38.04%}	736/2049	35.92% {33.84% 38.04%}
Uncertain Diagnoses	880/2049	42.95% {40.79% 45.12%}	435/2049	21.23% {19.48% 23.07%}
Below Threshold	624/880	70.91% {67.78% 73.89%}	254/435	58.39% {53.60% 63.07%}
Above Threshold	256/880	29.09% {26.11% 32.22%}	181/435	41.61% {36.93% 46.40%}
Uncertain Subjects Eligible	1360/2049	66.37% {64.28% 68.42%}	990/2049	48.32% {46.13% 50.51%}
Uncertain Subjects Ineligible	480/2049	23.43% {21.61% 25.32%}	555/2049	27.09% {25.17% 29.07%}

over a finite amount of time is an estimate of the true, endogenous rate we might observe with infinite data. As the amount of data observed increases, the size of the statistical confidence interval decreases. So the more sleep time observed, the smaller the uncertainty of the AHI will become. But to what degree will additional testing affect the size of the confidence interval, and how much data is needed to know that our estimate is significantly above or below threshold?

The ability to construct a confidence interval on the AHI provides a quantitative means of exploring the practicality of this approach. For a given endogenous apnea rate, it is therefore possible to figure out how much testing time is required such that the threshold falls outside the empirical 95% confidence interval. This analysis therefore identifies the required time to test to statistical significance for any given endogenous rate.

Fig. 3A illustrates the reduction in confidence interval size as a function of hours of observation for endogenous event rates of 3 (upper panel) and 3.9 (lower panel), as well as the testing time required until the threshold is no longer encompassed by the 95% confidence interval. The closer to threshold the endogenous rate is, the more testing is required. Thus, while it would take 9.70 h of testing to distinguish someone with an endogenous apnea rate of 3

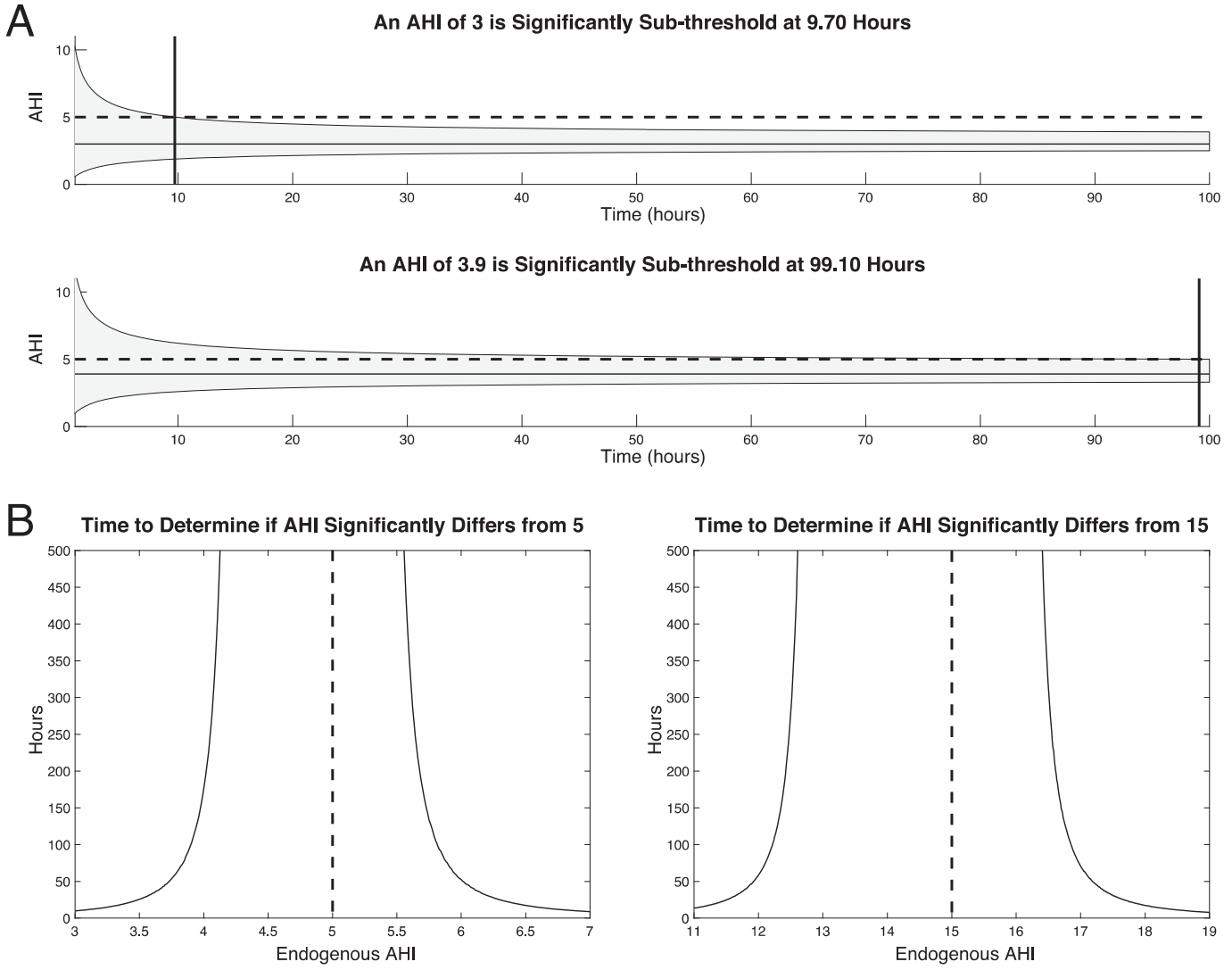


Fig. 3. Testing time required to differentiate a given endogenous AHI from a threshold. (A) The AHI 95% confidence interval (gray shaded region) decreases in size given more data observed over a longer total sleep time. The point of time (vertical line) at which the confidence interval differs from the threshold (dashed horizontal line) determines the testing time required to differentiate a given endogenous rate from the threshold with 95% confidence. The larger the difference between the endogenous rate and the threshold, the shorter testing time required, such that it would take 99.10 h of testing to distinguish an AHI of 3.9 from a threshold of 5 (bottom panel), but only 9.70 h for an AHI of 3 (top panel). (B) Generalizing, it is possible to show the testing time required to differentiate a range of endogenous rates from thresholds of 5 (left panel) and 15 (right panel). As the endogenous rates approach the threshold, the time required diverges towards infinity.

from a threshold of 5, it would require 99.10 h of testing for someone with an endogenous rate of 3.9.

Generalizing this analysis, Fig. 3B shows the required time to test for endogenous apnea rates surrounding thresholds of 5 (left panel) and 15 (right panel). These results illustrate the rapidity at which the required testing time diverges to infinity as the endogenous rate approaches the threshold. The divergence is so rapid that testing to significance for endogenous rates of 4.25 (for a threshold of 5) or 13 (for a threshold of 15) will require over 1000 h of data. Thus, the amount of additional testing required to establish statistical significance for common borderline cases falls well outside the realm of practical feasibility.

Conversely, we can also ask the question of which endogenous apnea rate values can be captured in a single night of sleep by identifying values in which the required time to testing is under 8 h. For a threshold of 5, endogenous apnea rates of under 2.8 and over 7.1 can be resolved in 8 h or less. For a threshold of 15, endogenous apnea rates under 10.4 and over 18.9 can be resolved in 8 h or less. This is equivalent to determining the boundaries of the uncertain area (gray region) on Fig. 2 for TST $= 8$. Given the steep exponential increase in time, adding an extra night (16 h) provides only a small benefit, pushing the range to 3.3 and 6.5 for a threshold of 5, and 11.1 and 18.1 for a threshold of 15.

4. Discussion

The AHI and its derivatives drive sleep apnea care and the approval of new diagnostic and treatment devices, with specific AHI values determining a patient's access to therapy, as well as to the type of therapy available. Studies, however, have shown that individual AHI estimates are not strongly linked to clinical and research outcomes [34], and our results suggest that uncertainty is likely a major contributing factor in this disparity. Within the MESA population, the impact of AHI uncertainty is clearly illustrated, with up to 43% of the general population having uncertain diagnoses. For symptomatic subjects, given that ~63% of subjects with uncertain

diagnoses are below threshold, it is likely that many patients are currently being denied treatment eligibility due to variability attributable purely to chance.

Our analyses show that the patients most impacted clinically by uncertainty will be those with low TST. Thus, extra consideration is important in split night studies, where the time interval is small, as a sub-threshold AHI can easily have uncertainty that reaches above the threshold (see Patient E in Fig. 1). This problem is exacerbated in pediatrics, where the threshold for abnormality (AHI≥1) involves small counts over potentially short time intervals. For example, observing 3 events over 8 h yields an AHI of 0.37 but an estimated 95% empirical confidence ranging from 0.07 to 1.3, which is over the pediatric threshold. Even the lack of observation of any events does not guarantee a below-threshold AHI if the time period is small enough—a pediatric patient with 0 events over 4 h has an AHI of 0 but an estimated 95% empirical confidence interval ranging from 0 to 1.05.

The link between N and TST and confidence interval width also means that we cannot deal with uncertainty simply by proposing a global alteration of threshold (eg, lowering the AHI threshold from 5 to 4). This is because different patients with the same AHI could have drastically different confidence intervals if their N and TST differ (see Patient D vs Patient E in Fig. 1). Therefore, there is no one-size-fits-all threshold adjustment that can be made to account for uncertainty.

Overall, statistical significance is fundamentally a statement about the evidence available in the data, not about the underlying mechanisms of the pathology. Lack of significance should be a sign that more data is needed, not that therapy is unlikely to be effective. Unfortunately, our results have shown that, for a wide range of borderline cases, it is not feasible to reduce uncertainty for determining statistical significance by testing further. In the future, wearable devices may be able to provide months or years of data in borderline cases if clinically desirable, but empiric therapy for symptomatic individuals may be more cost-effective.

If we cannot experimentally reduce uncertainty by additional testing for many of the uncertain cases, we can at least quantify the degree of AHI uncertainty by providing clinicians with patient-specific uncertainty in all reports, analyses, and assessment of any future eligibility criteria. Then, rather than a hard AHI threshold, providers would have greater ability to consider other quantifiable clinical factors in determining treatment eligibility for statistically uncertain cases, especially for therapeutic trials in borderline cases—both above and below the threshold. While overdiagnosis could potentially be an issue in general population screening, the risk of overdiagnosis in symptomatic patients presenting to sleep centers is much less likely. Furthermore, more inclusive diagnostic criteria could potentially reduce overall long-term costs associated with co-morbidities.

In addition to providing more information for current diagnoses, uncertainty estimates are essential in evaluating any future strategies and criteria proposed for eligibility. Further improvement may be achieved by characterizing sleep disordered breathing as more than a simple average event rate. By explicitly modeling respiratory events as a dynamic process affected by numerous intrinsic and extrinsic factors, rather than a static average, future work can provide a far more information-rich framework for phenotyping, clinical assessment and for treatment development.

Our analysis focused on obstructive sleep apnea, as central sleep apnea was very rare in the MESA cohort. However, the principles described here can be applied, in a more complex fashion to diagnostic and coverage decisions for central sleep apnea therapy. The central apnea-hypopnea index (CAHI) has, in addition to a threshold on the central event rate, an additional threshold on the ratio of central to total events (above 50%), which is required for current treatment coverage decisions (eg, supplemental oxygen, adaptive ventilation). Thus, the uncertainty in the CAHI is likely greater than that of the AHI, given the additional variability in estimating the event ratio. Future work can help to quantify the degree and implications of this variability.

5. Conclusions

The ability to quantify uncertainty for metrics of sleep disordered breathing is vital to understanding patient diagnosis and as well as the effects of treatment. The uncertainty inherent in the current system is vast compared to the clinical thresholds due to limitations of the data feasibly collectable in a clinical setting. By observing the effect of uncertainty on eligibility in the MESA dataset, we can identify that there are many uncertain cases, suggesting that it is likely that patients are currently being incorrectly diagnosed based on variability attributable to chance. It is therefore insufficient to rely on the relationship between a single number and threshold alone, and strategies must be developed to incorporate uncertainty, as well other available data, into clinical decision-making.

Financial Disclosure

Dr. Thomas reports the following: (1) Patent, license and royalties from MyCardio, LLC, for an ECG-based method to phenotype sleep quality and sleep apnea; (2) Grant support, license and intellectual property (patented) from DeVilbiss Healthcare; (3) Guidepoint Global, ClearView Healthcare Partners, and GLG Councils - consulting for general sleep medicine; and (4) Intellectual Property (patent, unlicensed) for a device using CO_2 for central/complex sleep apnea.

Acknowledgements

This work was supported by the National Institutes of Health, National Institute of Neurological Disorders and Stroke R01 NS-096177 (M.J.P.) and Simons Foundation 542971 (U.T.E.).

Conflict of interest

The ICMJE Uniform Disclosure Form for Potential Conflicts of Interest associated with this article can be viewed by clicking on the following link: https://doi.org/10.1016/j.sleep.2019.06.003.

Appendix

Constructing Exact and Approximate Bounds on the AHI

For Poisson processes, the variance of the event count within any time interval is equal to the rate of the events. This means that AHI uncertainty will increase as the number of events increases and as the duration of the epoch decreases. We can construct bounds on a given AHI using the formula for the confidence interval of a Poisson process

$$CI_\alpha = \left[q_{\chi^2}\left(\frac{\alpha}{2},\ 2N\right)\ [2\ TST]^{-1},\ \ q_{\chi^2}(1-\alpha/2,\ 2(N+1))[2\ TST]^{-1}\right] \tag{A1}$$

where q_{χ^2} is the quantile function for the chi squared distribution, α is the significance level for the confidence interval (eg, 0.05 for 95% confidence) and N and TST are the number of events at total sleep time, respectively [35]. One-sided confidence intervals can also be constructed using this same, chi squared framework.

Alternatively, there is an approximate solution [36] for the 95% AHI confidence interval

$$CI_{95} = \exp\left\{\log\left(\frac{N}{TST}\right) \pm \frac{1.96}{\sqrt{N}}\right\} \quad \text{(A2)}$$

which is can be easily computed without requiring specialized statistical software and is accurate for practical purposes as long as $TST \geq 2$ hours and $N \geq 5$. In this paper we will exclusively use the exact confidence interval on all AHI computations.

References

[1] Tsai M, Khayat R. Sleep apnea in heart failure. Curr Treat Options Cardiovasc Med 2018;20(4):33.
[2] Muraki I, Wada H, Tanigawa T. Sleep apnea and type 2 diabetes. J Diabetes Investig 2018;9:991–7.
[3] Kwon Y, Koene RJ, Johnson AR, et al. Sleep, sleep apnea and atrial fibrillation: questions and answers. Sleep Med Rev 2018;39:134–42.
[4] Punjabi NM. COUNTERPOINT: is the apnea-hypopnea index the best way to quantify the severity of sleep-disordered breathing? No. Chest 2016;149(1):16–9.
[5] Rapoport DM. POINT: is the apnea-hypopnea index the best way to quantify the severity of sleep-disordered breathing? Yes. Chest 2016;149(1):14–6.
[6] Javaheri S, Brown LK, Randerath WJ. Clinical applications of adaptive servo-ventilation devices: part 2. Chest 2014;146(3):858–68.
[7] Javaheri S, Brown LK, Randerath WJ. Positive airway pressure therapy with adaptive servoventilation: part 1: operational algorithms. Chest 2014;146(2):514–23.
[8] Colrain IM, Black J, Siegel LC, et al. A multicenter evaluation of oral pressure therapy for the treatment of obstructive sleep apnea. Sleep Med 2013;14(9):830–7.
[9] Strollo Jr PJ, Soose RJ, Maurer JT, et al. Upper-airway stimulation for obstructive sleep apnea. N Engl J Med 2014;370(2):139–49.
[10] Costanzo MR, Ponikowski P, Javaheri S, et al. Sustained 12 Month benefit of phrenic nerve stimulation for central sleep apnea. Am J Cardiol 2018;121(11):1400–8.
[11] Kushida CA, Chediak A, Berry RB, et al. Clinical guidelines for the manual titration of positive airway pressure in patients with obstructive sleep apnea. J Clin Sleep Med 2008;4(2):157–71.
[12] Guilleminault C, Dement WC. Sleep apnea syndromes and related sleep disorders. In: Sleep disorders : diagnosis and treatment/edited by Robert L Williams, Ismet Karacan. New York: Wiley; 1978. p. 9–28. 1978.
[13] Guilleminault C. Obstructive sleep apnea. The clinical syndrome and historical perspective. Med Clin N Am 1985;69(6):1187–203.
[14] Stoberl AS, Schwarz EI, Haile SR, et al. Night-to-night variability of obstructive sleep apnea. J Sleep Res 2017;26(6):782–8.
[15] White LH, Lyons OD, Yadollahi A, et al. Night-to-night variability in obstructive sleep apnea severity: relationship to overnight rostral fluid shift. J Clin Sleep Med 2015;11(2):149–56.
[16] Aber WR, Block AJ, Hellard DW, et al. Consistency of respiratory measurements from night to night during the sleep of elderly men. Chest 1989;96(4):747–51.
[17] Aarab G, Lobbezoo F, Hamburger HL, et al. Variability in the apnea-hypopnea index and its consequences for diagnosis and therapy evaluation. Respiration 2009;77(1):32–7.
[18] Chediak AD, Acevedo-Crespo JC, Seiden DJ, et al. Nightly variability in the indices of sleep-disordered breathing in men being evaluated for impotence with consecutive night polysomnograms. Sleep 1996;19(7):589–92.
[19] Davidson TM, Gehrman P, Ferreyra H. Lack of night-to-night variability of sleep-disordered breathing measured during home monitoring. Ear Nose Throat J 2003;82(2):135–8.
[20] Kishi A, Van Dongen HP, Natelson BH, et al. Sleep continuity is positively correlated with sleep duration in laboratory nighttime sleep recordings. PLoS One 2017;12(4):e0175504.
[21] Somiah M, Taxin Z, Keating J, et al. Sleep quality, short-term and long-term CPAP adherence. J Clin Sleep Med 2012;8(5):489–500.
[22] Snyder DL, Miller MI. Random point processes in time and space. In: Springer science & business media; 2012.
[23] Edwards BA, Andara C, Landry S, et al. Upper-airway collapsibility and loop gain predict the response to oral appliance therapy in patients with obstructive sleep apnea. Am J Respir Crit Care Med 2016;194(11):1413–22.
[24] Joosten SA, Leong P, Landry SA, et al. Loop gain predicts the response to upper airway surgery in patients with obstructive sleep apnea. Sleep 2017;40(7).
[25] Stanchina M, Robinson K, Corrao W, et al. Clinical use of loop gain measures to determine continuous positive airway pressure efficacy in patients with complex sleep apnea. A pilot study. Ann Am Thorac Soc 2015;12(9):1351–7.
[26] Barnes H, Edwards BA, Joosten SA, et al. Positional modification techniques for supine obstructive sleep apnea: a systematic review and meta-analysis. Sleep Med Rev 2017;36:107–15.
[27] Joosten SA, O'Donoghue FJ, Rochford PD, et al. Night-to-night repeatability of supine-related obstructive sleep apnea. Ann Am Thorac Soc 2014;11(5):761–9.
[28] Yalciner G, Bababemez MA, Gul F. Association of sleep time in supine position with apnea-hypopnea index as evidenced by successive polysomnography. Sleep Breath 2017;21(2):289–94.
[29] Thomas RJ, Mietus JE, Peng CK, et al. An electrocardiogram-based technique to assess cardiopulmonary coupling during sleep. Sleep 2005;28(9):1151–61.
[30] Chen X, Wang R, Zee P, et al. Racial/ethnic differences in sleep disturbances: the multi-ethnic study of Atherosclerosis (MESA). Sleep 2015;38(6):877–88.
[31] Yeboah J, Redline S, Johnson C, et al. Association between sleep apnea, snoring, incident cardiovascular events and all-cause mortality in an adult population: MESA. Atherosclerosis 2011;219(2):963–8.
[32] Dean 2nd DA, Goldberger AL, Mueller R, et al. Scaling up scientific discovery in sleep medicine: the national sleep research Resource. Sleep 2016;39(5):1151–64.
[33] Zhang GQ, Cui L, Mueller R, et al. The national sleep research Resource: towards a sleep data commons. J Am Med Inform Assoc 2018;25(10):1351–8.
[34] Vgontzas AN. Excessive daytime sleepiness in sleep apnea: it is not just apnea hypopnea index. Sleep Med 2008;9(7):712–4.
[35] Garwood F. Fiducial limits for the Poisson distribution. In: Biometrika, vol. 28. [Oxford University Press, Biometrika Trust]; 1936. p. 437–42.
[36] Schwertman NC, Martinez RA. Approximate Poisson confidence limits. In: Communications in statistics - theory and methods, vol. 23. Taylor & Francis; 1994. p. 1507–29.