Potential Underestimation of Sleep Apnea Severity by At-Home Kits: Rescoring In-Laboratory Polysomnography Without Sleep Staging

Matt T. Bianchi, MD, PhD1,2; Balaji Goparaju, MS1

1Neurology Department, Massachusetts General Hospital, Boston, MA; 2Division of Sleep Medicine, Harvard Medical School, Boston, MA

Study Objectives: Home sleep apnea testing (HSAT) is increasingly available for diagnosing obstructive sleep apnea (OSA). One key limitation of most HSAT involves the lack of sleep staging, such that the respiratory event index is calculated using the total recording time (TRT) rather than total sleep time (TST).

Methods: We performed a retrospective analysis of n = 838 diagnostic polysomnography (PSG) nights from our center; n = 444 with OSA (4% rule, apnea-hypopnea index (AHI) ≥ 5), and n = 394 with AHI < 5. We recalculated the AHI using time in bed (TIB) instead of TST, to assess the predicted underestimation risk of OSA severity.

Results: Of all the patients with OSA, 26.4% would be recategorized as having less severe or no OSA after recalculating the AHI using TIB rather than TST. Of the n = 275 with mild OSA, 18.5% would be recategorized as not having OSA. The risk of underestimation was higher in those with moderate or severe OSA. Of the n = 119 moderate OSA cases, 40.3% would be recategorized as mild, and of the n = 50 severe OSA cases, 36.0% would be recategorized as moderate. Age strongly correlated with the degree of underestimation of the AHI, because age was significantly correlated with time awake during PSG.

Conclusions: The risk of sleep apnea underestimation is predicted to be substantial in a tertiary sleep center population. Phenotyping errors included risk of falsely negative results (from mild to normal), as well as category errors: moderate or severe moving to mild or moderate severity, respectively. Clinicians should recognize this underestimation limitation, which directly affects diagnostic phenotyping and thus therapeutic decisions.

Commentary: A commentary on this article appears in this issue on page 531.

Keywords: diagnosis, prediction, underestimation


INTRODUCTION

The limitations of home sleep apnea testing (HSAT) kits are well described, and include lack of leg movements, lack of sleep staging, lack of body position tracking (in most devices), and lack of general validation in populations beyond straightforward high-probability obstructive sleep apnea (OSA) cases. One key limitation of HSAT for diagnosing sleep apnea is that the devices generally do not include electroencephalography (EEG) to score sleep-wake stages. In fact, only the ARES system (SleepMed, Palm Beach, Florida, United States) does this, through limited frontal channels. The remaining devices either use total recording time (TRT) as the denominator to calculate a respiratory event index, or actigraphy to estimate total sleep time (TST). Actigraphy does not substitute for EEG, because it is well known to overestimate sleep when movement is the only input due to quiet wake appearing similar to sleep. HSAT kits are known to underestimate the severity of sleep-disordered breathing. Underestimation is at least in part because the respiratory event rate is calculated using a denominator of the TRT and not TST, which is necessarily equal to or lower than the TRT. Using patient estimates of TST as an adjunctive value may not solve this problem, as many publications have documented heterogeneity and variability with both overestimation and underestimation of subjective TST (sTST) in clinical contexts.
severe cases falsely labeled as moderate might also drive therapy decisions. This might cause a physician to exclude surgical treatment options (that often are reserved for severe cases), or choose a treatment option that should be used on mild to moderate cases, such as oral appliance therapy. Therefore, the risks of underestimation are diverse. We undertook the current study in a large retrospective cohort to estimate such risks by rescoring in-laboratory PSG data as if they are HSAT kits.

**RESULTS**

Table 1 shows baseline characteristics of the two cohorts, with or without OSA defined solely by the AHI value (≥ 5 versus < 5) using the 4% desaturation rule for defining hypopnea. The expected differences associated with OSA were seen: the AHI > 5 cohort was older, more likely male, had a higher body mass index (BMI), and higher AHI. Importantly, the oTST and the recording time of PSG (TIB) did not significantly differ between the groups.

For the group without OSA, further category underestimation is not possible, though the normal-range AHI values decrease further when using TIB instead of oTST; this group had similar mean time awake after sleep onset compared to the OSA group (50 minutes versus 54 minutes; not shown). For the group with OSA, the recalculated AHI based on TIB rather than oTST resulted in severity category reclassification in 26.4% of patients (Figure 1). Most of these were “one-level” reclassifications (eg, mild cases to no OSA, and moderate cases to mild). There were similar levels of reclassification when the cohort with OSA was divided into subcategories based on self-reported insomnia symptoms (data not shown). Age was significantly correlated with the difference between the AHI and the recalculated AHI based on TIB (Spearman R = 0.36, P < .0001), which was accounted for by a strong correlation of age with increased time awake during PSG (R = 0.39, P < .0001). Neither sex nor BMI nor TIB were correlated with the AHI difference upon recalculation.

One potential solution to the issue of underestimation based on TIB is to use stST in hopes that the value is closer to the oTST, and any nights with stST estimate of 30 minutes or less. We recalculated the AHI values based on the time in bed (TIB) for each subject with OSA using the criteria of AHI ≥ 5 (n = 444). TIB excludes disconnections/out of bed time (eg, using the restroom). For recalculation of the AHI in the non-OSA group based on the stST as the denominator, we excluded n = 33 with stST > TIB (this would be obvious from HSAT evaluation clinically). Nonparametric methods were used because of generally nonnormal distribution of variables; significance was defined by a value of P < .05.

**METHODS**

We analyzed a retrospective sample of 838 diagnostic PSG tests performed in our clinical sleep laboratory, irrespective of indication, although most referrals to our center are for evaluation of OSA. The time frame was 2010 to 2012, which is mostly before recent insurance changes affected in-home versus in-laboratory testing decisions. The Institutional Review Board approved the retrospective analysis of our clinical sleep laboratory database without requiring consent. PSG was performed according to AASM standards and scored by experienced registered technologists. We use the 4% desaturation rule for defining hypopnea; we also rescoring the cohort according to the 3% desaturation rule. Because our standard reports only consider 4%, this required retrospective rescoring, and some studies could not be rescored for technical reasons (n = 398 still available that met criteria of AHI > 5 using a 3% desaturation definition). Subjective sleep duration is queried after each PSG, using the question “How long do you think you slept?” We previously reported internal inconsistency when sleep duration is reported this way, versus indirectly (by clock time), but we did not separately analyze that factor here. The cohort included n = 394 without OSA based on the criteria of AHI < 5, and n = 444 with OSA based on the criteria of AHI ≥ 5, using the 4% desaturation rule for defining hypopnea. We prespecified exclusions of age younger than 18 years, missing stST estimation, nights with less than 60 minutes of objective TST (oTST), and any nights with stST estimate of 30 minutes or less. We recollected the AHI values based on the time in bed (TIB) for each subject with OSA using the criteria of AHI ≥ 5 (n = 444). TIB excludes disconnections/out of bed time (eg, using the restroom). For recalculation of the AHI in the non-OSA group based on the stST as the denominator, we excluded n = 33 with stST > TIB (this would be obvious from HSAT evaluation clinically). Nonparametric methods were used because of generally nonnormal distribution of variables; significance was defined by a value of P < .05.

**Table 1**—Clinical characteristics.

<table>
<thead>
<tr>
<th></th>
<th>AHI &lt; 5</th>
<th>AHI ≥ 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>394</td>
<td>444</td>
</tr>
<tr>
<td>Age</td>
<td>(18–82)</td>
<td>(20–88)**</td>
</tr>
<tr>
<td>% male</td>
<td>43.7</td>
<td>54.5*</td>
</tr>
<tr>
<td>BMI</td>
<td>(24–33)</td>
<td>(27–36)**</td>
</tr>
<tr>
<td>ESS</td>
<td>(7–12)</td>
<td>(4–11)</td>
</tr>
<tr>
<td>TST (min)</td>
<td>379 (339–422)</td>
<td>377 (330–414)</td>
</tr>
<tr>
<td>TIB (min)</td>
<td>441 (408–474)</td>
<td>439 (410–465)</td>
</tr>
<tr>
<td>stST (min)</td>
<td>360 (240–405)</td>
<td>330 (240–390)</td>
</tr>
<tr>
<td>AHI (events/h)</td>
<td>1.6 (0.3–3.1)</td>
<td>12.0 (7.7–19.3)**</td>
</tr>
<tr>
<td>No insomnia</td>
<td>33.0</td>
<td>39.2</td>
</tr>
</tbody>
</table>

Values are median and interquartile range except for age (mean and range), sex (%). No insomnia (%) refers to the absence of insomnia symptoms on our pre-PSG intake forms (that is, most patients reported insomnia symptoms). * = P < .01, chi-square test. ** = P < .0001, Mann-Whitney U test. AHI = apnea-hypopnea index, BMI = body mass index; ESS = Epworth Sleepiness Scale, stST = subjective total sleep time, TIB = time in bed, TST = total sleep time.
MT Bianchi and B Goparaju. Limitations of HSAT by Re-Analysis of PSG

We also examined the results using less stringent event scoring. The lack of EEG prevents scoring of non-hypoxic respiratory event-related arousals. Sleep-disordered breathing of this nature would thus be underestimated by HSAT. Among those with recalculated AHI < 5, it was observed that 66.4% had RDI > 5, and 22.9% had RDI > 15. These individuals would not be recognized by level 3 monitors, most of which lack EEG. We also re-analyzed the cohort using the AASM criteria of 3% desaturation for scoring hypopneas. Of those with a mild range AHI (≥ 5 to < 15), 38 of 264 (14.4%) would have been reclassified as AHI < 5 (no OSA). Of those with a moderate range AHI (≥ 15 to < 30), 32 of 108 (29.6%) would have been reclassified as AHI < 15. Of those with AHI > 30, 8 of 25 (32%) would have been reclassified as AHI < 30.

Finally, we evaluated the prevalence of elevated periodic limb movements of sleep (PLMS), which are not measured by current level 3 or level 4 HSAT devices. We observed elevated PLMS index values in a substantial subset of this cohort, with 28.5% having an index > 15 events/h, and 15.5% with > 30 events/h. This proportion was similar when examined in subsets

(16.9%) moved up at least one severity category. The specific category movements are shown in Figure 2C, with the most common movement being either mild to moderate or moderate to severe.

We also examined the results using less stringent event scoring. The lack of EEG prevents scoring of non-hypoxic respiratory event-related arousals. Sleep-disordered breathing of this nature would thus be underestimated by HSAT. Among those with recalculated AHI < 5, it was observed that 66.4% had RDI > 5, and 22.9% had RDI > 15. These individuals would not be recognized by level 3 monitors, most of which lack EEG. We also re-analyzed the cohort using the AASM criteria of 3% desaturation for scoring hypopneas. Of those with a mild range AHI (≥ 5 to < 15), 38 of 264 (14.4%) would have been reclassified as AHI < 5 (no OSA). Of those with a moderate range AHI (≥ 15 to < 30), 32 of 108 (29.6%) would have been reclassified as AHI < 15. Of those with AHI > 30, 8 of 25 (32%) would have been reclassified as AHI < 30.

Finally, we evaluated the prevalence of elevated periodic limb movements of sleep (PLMS), which are not measured by current level 3 or level 4 HSAT devices. We observed elevated PLMS index values in a substantial subset of this cohort, with 28.5% having an index > 15 events/h, and 15.5% with > 30 events/h. This proportion was similar when examined in subsets
above or below a recalculated AHI (4%) value of 5, and similar to our recent report in a sleep apnea cohort during diagnostic and titration PSG nights.\textsuperscript{10} For those with AHI < 5, the PLMS index was > 15 in 26.3% and > 30 in 13.8%. For those with AHI > 5, the PLMS index was > 15 in 31.0% and > 30 in 17.4%.

### DISCUSSION

The current study indicates that the common method of using TRT as the denominator in HSAT would result in substantial underestimation of OSA severity when applied to a large clinical cohort of diagnostic PSG. The extent of AHI underestimation affected patients at all OSA severity levels, and was strongly correlated with age, because older age was strongly correlated with time awake during PSG and thus extent of AHI underestimation. The results emphasize that lack of sleep-wake staging in most HSAT devices carries substantial risk of OSA underestimation. Using sTST instead of the TIB does not mitigate the OSA underestimation risk. In addition, sTST introduces a potential AHI overestimation risk, if individuals underestimate sleep duration compared to objective measures. Clinicians should use caution when interpreting HSAT results, and recognize that OSA underestimation directly affects diagnostic phenotyping and thus therapeutic decisions.

The diagnosis of OSA carries uncertainty whether HSAT or in-laboratory PSG is performed. A single night of either method is an inherently undersampled view of a pathophysiology that may vary night to night.\textsuperscript{11,12} Underestimation of OSA severity is arguably the more important risk than overestimation, though this has not been formally studied. The underestimation related to use of TRT instead of TST as the denominator is distinct from other important inferential challenges related to the sensitivity and specificity of HSAT to detect OSA. This includes, for example, risks associated with automated versus manual scoring,\textsuperscript{3} as well as inferential risks\textsuperscript{13,14} when HSAT is conducted in low-risk populations, and falsely positive results are more likely, whereas in high-risk populations falsely negative results are more likely.

Although the risk of falsely negative results carries clear implications, clinical risk is potentially incurred even in those who meet criteria for OSA with AHI > 5. OSA severity misclassification between mild, moderate, and severe may compromise care planning. First, OSA severity category predicts medical risk.\textsuperscript{15–17} Medical risk perception and baseline AHI severity may motivate patients to pursue treatment in the setting of a new diagnosis.\textsuperscript{18,19} As much uncertainty surrounds the effect of and need to treat mild OSA, assigning this severity category by HSAT could be falsely reassuring if the HSAT underestimated a true moderate severity case. If HSAT is used to assess severity after an intervention such as weight loss, dental appliance, or surgery, the results may be falsely optimistic compared to baseline assessments of severity made by PSG. OSA severity may inform the choice of non-positive airway pressure (PAP) options for those who cannot or will not comply with PAP. For example, dental appliance therapy is mainly recommended for mild to moderate cases (although recent evidence suggests some consideration for severe cases\textsuperscript{20}). Furthermore, surgical approaches that might be more favorable for severe cases may not be considered for those falsely categorized as moderate. Although weight loss should be recommended to any patient with OSA and elevated BMI, it may be that those with mild OSA could reasonably choose a trial of weight loss as their sole initial therapy.

Several potential solutions may mitigate this underestimation risk attributable to using TRT rather than oTST as the denominator for calculating severity. Advances in quantifying sleep through non-EEG signals, such as autonomic physiology\textsuperscript{21–23} may be more easily implemented than, for example, adding EEG channels to existing systems. However, there is little incentive for HSAT device manufacturers to pursue such improvements. Until and unless such methods enjoy wide clinical implementation, practitioners should be cautious when interpreting HSAT results. Even when the TRT and the sTST seem to be aligned, because of unreliability of sTST and the lack of an independent gold standard in HSAT settings, uncertainty remains regarding underestimation of OSA.

In addition to this OSA underestimation issue, another common sleep disturbance will also be missed by level 3 or level 4 kits. PLMS occurred often in patients from our laboratory, whether or not the recalculated AHI values would support a diagnosis of OSA. PLMS are of clinical relevance for several potential reasons. Epidemiology studies have linked PLMS to cardiovascular and cerebrovascular risk.\textsuperscript{24} Occult PLMS may contribute to sleep fragmentation and symptoms of insomnia or nonrefreshing sleep. This treatable cause of sleep disturbance is not assessed by currently available HSAT devices, which do not include limb sensors. Evaluation of PLMS with PSG should be considered in patients with ongoing symptoms despite treatment of OSA diagnosed by HSAT.

This retrospective study has limitations that could in principle be addressed in future work. The tertiary referral nature and heterogeneity of sleep and systemic pathology may not generalize to other clinical settings. For example, a recent study suggested minimal misclassification in patients with OSA when rescoring the AHI based on self-reported TST.\textsuperscript{25} Other work has suggested that actigraphy concurrent with home testing did not substantially alter the scoring results,\textsuperscript{26} although this may be in part due to actigraphy overestimating oTST, which is essentially the risk with using TIB as detailed here. Another point of uncertainty involves the relationship between in-laboratory oTST relative to at-home oTST within individuals. Some patients may have more consolidated sleep (higher efficiency) at home than in the laboratory because of first-night effect.\textsuperscript{27} Other patients, by contrast, will have the opposite pattern, possibly because of the so-called reverse first-night effect such as may occur in psychophysiological insomnia.\textsuperscript{28} The clinical challenge remains that when presented with an individual patient and his or her HSAT report, there is no independent predictor or gold standard to decide whether using TIB or sTST would be a reasonable estimate for calculating the OSA severity index.

### ABBREVIATIONS

AASM, American Academy of Sleep Medicine  
AHI, apnea-hypopnea index
REFERENCES


SUBMISSION & CORRESPONDENCE INFORMATION

Submitted for publication June 17, 2016
Submitted in final revised form November 9, 2016
Accepted for publication December 9, 2016
Address correspondence to: Matt T. Bianchi, MD, PhD, Wang 7 Neurology, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114; Tel: (617) 724-7426; Fax: (617) 724-6513; Email: mbianchi@partners.org

DISCLOSURE STATEMENT

This was not an industry supported study. Dr. Bianchi has received funding from Massachusetts General Hospital, the Center for Integration of Medicine and Innovative Technology, the Milton Family Foundation, and currently receives funding from the Department of Neurology, the MGH-MIT Grand Challenge, and the American Sleep Medicine Foundation. Dr. Bianchi has a patent pending on a home sleep monitoring device. Dr. Bianchi received travel funding from Servier; has consulting and research agreements with MC10, Insomnisolv, and McKesson; is a medical monitor for Pfizer; and has provided expert testimony in sleep medicine. Dr. Goparaju has indicated no financial conflicts of interest.