Oxygen Desaturation Index Differs Significantly Between Types of Sleep Software

Yvonne Ng, MBBS(Hons), BMedSc, FRACP; Simon A. Joosten, MBBS, BMedSc, FRACP, PhD1,2; Bradley A. Edwards, PhD3,4; Anthony Turton, BSc(Hons)1; Helen Romios, BSc (MLS)1; Thilini Samarasinghe, BBiomedSc(Hons), PhD1; Shane Landry, PhD4,5; Darren R. Mansfield, MBBS, FRACP, PhD1,3,5; Garun S. Hamilton, MBBS, FRACP, PhD1,2,3

1Department of Lung and Sleep Medicine, Monash Health, Victoria, Australia; 2School of Clinical Sciences, Monash University, Victoria, Australia; 3Monash Partners, Epworth, Victoria, Australia; 4Sleep and Circadian Medicine Laboratory, Department of Physiology, Monash University, Victoria, Australia; 5School of Psychological Sciences and Monash Institute of Cognitive and Clinical Neurosciences, Monash University, Victoria, Australia; 6Hudson Institute of Medical Research, Victoria, Australia

Study Objectives: The aim of this study was to compare the oxygen desaturation index (ODI) generated by two different sleep software systems.

Methods: Participants undergoing diagnostic polysomnography for suspected obstructive sleep apnea underwent simultaneous oximetry recording using the RespMed ApneaLink Plus device (AL) and Compumedics Profusion PSG3 system (Comp). The ODI was calculated by the algorithms in the respective software of each system. To determine if differences were due to algorithm or recording devices, the Comp software was also used to generate ODI values using oximetry data from the AL.

Results: In 106 participants, there was good correlation but poor agreement in the ODI generated by the two systems. AL ODI values tended to be higher than Comp ODI values, but with significant variability. For ODI4%, bias was 4.4 events/h (95% limits of agreement −5.8 to 14.6 events/h). There was excellent correlation and agreement when the same oximetry raw data was analyzed by both systems. For ODI4%, bias was 0.03 events/h (95% limits of agreement −2.7 to 2.8 events/h). Similar results were evident when the ODI3% was used.

Conclusions: There is a clinically significant difference in ODI values generated by the two systems, likely due to device signal processing, rather than difference in ODI calculation algorithms.

Keywords: cardiovascular diseases, correlation, obstructive sleep apnea, oximetry, polysomnography


INTRODUCTION

Obstructive sleep apnea (OSA) is characterized by repetitive collapse of the upper airway during sleep. OSA is highly prevalent, with moderate-severe OSA reported to affect between 17% and 49% of middle-aged men.1,2 The gold standard for diagnosis of OSA is in-laboratory overnight polysomnography. Polysomnography is complex and labor intensive and incorporates the measurement of multiple physiological signals with manual interpretation of data. OSA is most commonly diagnosed based on the calculation of an apnea-hypopnea index (AHI). The rules for scoring hypopneas have changed over time3–5 and a key component of the definition of a hypopnea is the extent of oxygen desaturation accompanying the respiratory event.

Over recent years there has been increasing use of home-based, limited-channel recording devices to diagnose OSA. These devices are classified either as type 3 or type 4 monitors, depending on the number of signals recorded.6,7 The cornerstone signal of type 3, and particularly, type 4 devices is the oxygen saturation (SpO2), measured via pulse oximetry. This enables automatic calculation of an oxygen desaturation index (ODI), which is the average number of desaturation episodes per hour of recording. Typically, ODI is reported as the number of 3% desaturations (ODI3%) and/or the number of 4% desaturations (ODI4%). ODI has validity in the diagnosis of OSA.8 When there is a high clinical probability of OSA, an elevated ODI3% or ODI4% performs well at “ruling in” the diagnosis,9–13 particularly in obese subjects, and has higher diagnostic utility in severe compared to moderate-severe OSA.14 However, a “negative” result does not exclude OSA in those at high risk and still requires further evaluation. Furthermore, there has been increasing interest in the role of ODI4% in determining the risk for hypertension and cardiovascular disease in patients with OSA. In a report from
the European Sleep Apnea Database, although both ODI4% and AHI were associated with prevalent hypertension, only ODI4% remained independently associated with hypertension when both were entered into a multiple regression model.\textsuperscript{15} Moreover, the majority of studies showing an independent association between OSA and cardiovascular disease have embedded a 4% desaturation in the hypopnea definition.\textsuperscript{16,17} This hypopnea definition correlates very strongly to ODI4%.\textsuperscript{18} A large randomized controlled trial assessing whether CPAP lowers cardiovascular risk—the SAVE (Sleep Apnea Cardio-Vascular Endpoints) study—utilized ODI4% ≥ 12 events/h as the main inclusion criterion for OSA diagnosis and treatment.\textsuperscript{19} Importantly, this study used the ResMed ApneaLink device to measure the ODI4%.

A potential advantage of ODI over AHI is that it is based on information from a single channel (oximetry) rather than multiple respiratory channels and is calculated by an algorithm rather than manual scoring, and thus is highly reproducible if recorded and analyzed using the same device and software. The variability of AHI between and within observers is well characterized and appreciated by clinicians.\textsuperscript{20} However, there has not been any systematic comparison between ODI values measured using different oximeters and software. Such a comparison is important given that oximeter data processing and ODI algorithms differ between manufacturers,\textsuperscript{7} and it is unclear whether different systems would produce the same ODI values in a given patient, and whether this affects either OSA diagnosis or the assessment of cardiovascular risk. We hypothesize that there will be systematic differences between ODIs measured using different systems due to these algorithm differences. As a result, the current study aims to compare the ODI measured using two common home- and laboratory-based OSA testing systems, the ResMed ApneaLink Plus device (ResMed, Sydney, Australia)—hereafter referred to as AL and Compumedics Grael Profusion PSG3 system (Compumedics Limited, Abbotsford, Victoria, Australia)—hereafter referred to as Comp.

**METHODS**

Institutional ethics approval was obtained from the Monash Health Human Research Ethics Committee (approval number 14358L). Consecutive adult patients with suspected OSA who were undergoing an overnight in-laboratory diagnostic sleep study at Monash Medical Centre from Sunday to Thursday nights between October 9, 2014 and February 8, 2015 were eligible for the study. Up to three participants were enrolled each night. When the number of eligible patients exceeded the number of AL available, they were randomly selected for inclusion. Written informed consent was obtained from each participant. Enrolled participants underwent simultaneous oximetry recording using both the AL and Comp. The Comp was part of the patients’ diagnostic sleep study recording, which proceeded according to standard criteria.\textsuperscript{21} The AL was used in addition to the full polysomnography montage. Each system’s in-built oximeter was connected to individual Nonin pulse oximetry sensors (Nonin Medical, Inc., Minnesota, United States), which were attached to separate fingers on opposite hands. Consequently, the Comp ran from a finger sensor on one hand and the AL ran from a finger sensor on the opposite hand. The ODI4% and ODI3% for the simultaneous, matched total recording time (irrespective of whether the participant was awake or asleep) were automatically calculated by the software of each system using its respective oximetry data. Recording time rather than sleep time ODI was used for analysis so that the denominator (ie, time) for ODI calculation was the same for both the AL and Comp. This is because the AL does not have EEG input unlike the Comp in our current setup. This also mimics how the AL and other limited-channel devices perform in the home setting. The ResMed ApneaLink software system version 9.20 (ResMed, Sydney, Australia) and the Compumedics Profusion PSG3 software system version 3.4, build 401 (Compumedics Limited, Abbotsford, Victoria, Australia) were used for analysis. To determine if discrepancies were due to software algorithm or recording device data processing differences, the Comp software was used to generate ODI values utilizing the AL oximetry data exported in European Data Format (EDF). Data were entered into an Excel spreadsheet (2010, Version 14.5.7, Microsoft Corporation, Redmond, Washington, United States) and subsequently exported into Prism version 6.0, 2014 (GraphPad, San Diego, California, United States) for analysis. Correlation between the AL ODI and Comp ODI was assessed using the Pearson correlation coefficient for ODI3% and ODI4%. Bland-Altman plots were used to determine agreement between the ODIs calculated by the two systems.\textsuperscript{22} We aimed to enroll at least 100 participants to achieve a confidence interval (CI) for the 95% limits of agreement of ± 0.34\textit{s}, where \textit{s} is the standard deviation of the differences between measurements by the two methods.\textsuperscript{22,23} To determine the effect of differences in ODI between systems on potential clinical decision-making, categories of OSA diagnosis for the different systems were made utilizing ODI3% ≥ 5 events/h for “any” OSA and ≥ 15 events/h for “moderate-severe” OSA. These were compared with the AHI calculated through scoring of the respiratory signals recorded by the Comp using the American Academy of Sleep Medicine 2012 rules.\textsuperscript{21} The chi-square test and odds ratio were used to compare proportions of OSA diagnosis.

**RESULTS**

A total of 116 patients consented to the study. Data from 10 participants had to be excluded due to technical issues with the oximetry recording, as demonstrated in Figure 1. Table 1 summarizes the baseline characteristics of the remaining 106 participants.

**Device ODI Comparison**

Correlation and agreement between the ODI measured by each device, and calculated by their respective software, are summarized in Figure 2 and Figure 3. For AL ODI4% versus Comp ODI4%, \(r = 0.96 \text{ (} P < .0001; 95\% \text{ CI 0.94–0.97)} \) and \( r^2 = 0.92 \). For AL ODI3% versus Comp ODI3%, \( r = 0.94 \text{ (} P < .0001; 95\% \)
CI 0.91–0.96) and $r^2 = 0.87$. AL ODI values tended to be higher than Comp ODI values, but with significant variability. For ODI4% bias = 4.4 events/h (95% limits of agreement −5.8 to 14.6 events/h). For ODI3%, bias = 7.1 events/h (95% limits of agreement −6.4 to 20.6 events/h).

**Software ODI Comparison**

The performance of the two software systems was assessed by using a single recording source of data. The data recorded using the AL were analyzed by both the AL software and the Comp software (having been exported in EDF from the AL), thus allowing for a specific comparison of the effect of software algorithm on the ODI via correlation and agreement. For ODI4%, $r = 0.998$ ($P < .0001$, 95% CI 0.997–0.999) and bias = 0.03 events/h (95% limits of agreement −2.7 to 2.8 events/h). For ODI3%, $r = 0.997$ ($P < .0001$; 95% CI

### Table 1—Characteristics of study population (n = 106).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, y, mean (SD)</td>
<td>47 (15.5)</td>
</tr>
<tr>
<td>Sex, n (%)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>62 (58.5)</td>
</tr>
<tr>
<td>Female</td>
<td>44 (41.5)</td>
</tr>
<tr>
<td>Body mass index, kg/m², mean (SD)</td>
<td>31.9 (7.1)</td>
</tr>
<tr>
<td>Epworth Sleepiness Scale score, mean (SD)</td>
<td>7.8 (4.8)</td>
</tr>
<tr>
<td>Apnea-hypopnea index, events/h, mean (SD)</td>
<td>23.2 (25.0)</td>
</tr>
<tr>
<td>ODI4%, events/h, mean (SD)</td>
<td></td>
</tr>
<tr>
<td>ApneaLink</td>
<td>15.0 (17.8)</td>
</tr>
<tr>
<td>Compumedics</td>
<td>10.6 (16.3)</td>
</tr>
<tr>
<td>ODI3%, events/h, mean (SD)</td>
<td></td>
</tr>
<tr>
<td>ApneaLink</td>
<td>22.0 (19.4)</td>
</tr>
<tr>
<td>Compumedics</td>
<td>14.9 (18.8)</td>
</tr>
</tbody>
</table>

ODI4% = oxygen desaturation index for ≥ 4% desaturation, ODI3% = oxygen desaturation index for ≥ 3% desaturation, SD = standard deviation.
0.996–0.998) and bias = 0.08 events/h (95% limits of agreement −3.1 to 3.3 events/h).

**Device Bias on Diagnostic Categories**

When considering OSA as ODI3% ≥ 5 events/h, Comp ODI3% resulted in a diagnosis of OSA in 66 subjects, compared to 90 subjects for AL ODI3%, which is 36% more subjects with OSA when the ApneaLink was used; \( P = .0002 \). Odds ratio for OSA diagnosis with AL ODI3% compared to Comp ODI3% was 3.4 (95% CI 1.8–6.6). For moderate-severe OSA (ODI3% ≥ 15 events/h), AL ODI3% resulted in a diagnosis in 59 subjects (56%), compared to 32 subjects (30%) for Comp ODI3%, \( P = .0002 \). Odds ratio for diagnosing moderate-severe OSA with AL ODI3% compared to Comp ODI3% was 2.9 (95% CI 1.7–5.1). When looking at the diagnosis of moderate-severe OSA (based on AHI ≥ 15 events/h), the AL ODI3% and Comp ODI3% performed differently. AL ODI3% ≥ 15 events/h diagnosed moderate-severe OSA with a sensitivity of 96.4% (95% CI 87.5% to 99.6%) and a specificity of 88.2% (95% CI 76.1% to 95.6%). Comp ODI3% ≥ 15 events/h diagnosed moderate-severe OSA with a sensitivity of 58.2% (95% CI 44.1% to 71.3%) and a specificity of 100% (95% CI 93% to 100%).

**DISCUSSION**

Our study demonstrates a clinically significant difference between the ODI3% and ODI4% values generated by the AL software and Comp. Despite good correlation between devices, our results show that there is a bias for AL ODI values to be higher than Comp ODI. The bias is larger for ODI3% (7.1 events/h) than for ODI4% (4.4 events/h). More importantly, the agreement between devices was poor, with wide 95% limits of agreement, particularly for ODI3%. These differences have clinical relevance, with significantly more patients crossing potential diagnostic thresholds for OSA and moderate-severe OSA with the AL compared to Comp.

Oximetry is a key signal measured during polysomnography and is usually considered the most robust and reproducible signal. As a result, oximetry is the cornerstone of type 4, and many type 3, limited-channel home sleep test devices. The ODI is commonly used in some countries as a metric for diagnosing OSA8–13 and/or as a marker for cardiovascular risk.24–28 Implicit in the clinical utility of ODI is that the result is both accurate and reproducible. However, previous literature assessing the performance characteristics of individual oximeters has demonstrated that due to differences in internal signal processing, there is significant variability in the \( \text{SpO}_2 \) values obtained during simulated OSA.29 Although an ODI from individual oximeters has been validated as a “rule in” test for the diagnosis of moderate-severe OSA,10,13 there have been negligible data comparing the ODIs between oximeters. To date, the only previous study that directly compared ODI between devices is that by Zou et al.30 In this study, the investigators compared the Watch_PAT 100 (Itamar Medical Ltd., Caesarea, Israel) to unattended in-home polysomnography using the Embla A10 system (Medcare, Reykjavik, Iceland). As part of their broader analysis, the ODI4% automatically calculated by these two devices was compared. Similar to the current study, they demonstrated that despite good correlation there was poor agreement.
between devices. The Watch_PAT 100 tended to score higher ODI4% (particularly in mild-moderate OSA), with a bias of 4.4 (± 6.5) events/h, a bias which is identical to that observed in the current study. Importantly, the wide limits of agreement (−5.8 to 14.6 events/h) in our study highlights that clinicians cannot be confident that an ODI4% recorded in the AL is the same as that recorded in the Comp. The variability was even greater for ODI3%. In this case the bias was for the AL to record a higher ODI3% than the Comp by an average of 7.1 events/h. The limits of agreement were even wider (−6.4 to 20.6 events/h).

These findings are of importance to clinicians. Mean difference and limits of agreement between ODI were shown to be across diagnostic and severity thresholds for “any OSA”, and particularly for moderate-severe OSA. This is highlighted when one considers the proportion of patients who receive a “diagnosis” of either OSA (based on ODI3% ≥ 5 events/h) or moderate-severe OSA (based on ODI3% ≥ 15 events/h) with the ODI from either device. The current work demonstrates that significantly more patients would receive a diagnosis of OSA, or more particularly, moderate-severe OSA with the AL ODI compared to the Comp ODI. Furthermore, when compared to the gold standard of AHI for diagnosing moderate-severe OSA, at a threshold of ODI3% ≥ 15 events/h the AL ODI3% had a significantly higher sensitivity (96.4%) compared to Comp ODI3% (58.2%). This finding appears to partly contradict the findings of a previous study by Ward and colleagues. Contrary to our results, they found that the AL was less sensitive than the Comp at diagnosing OSA. However, there are several important differences between their study and ours. Most importantly, instead of ODI, they compared a different metric, the AHI, generated by both the AL and Comp, and found that the AL tended to underestimate the Comp AHI (mean difference 13.5 events/h during simultaneous recording). This is likely due to the different denominator used to calculate the AHI (monitoring time for the AL and sleep time for the Comp), leading to a systematically lower AHI for the AL, particularly for patients with poor sleep efficiency. In contrast, we used recording time instead of sleep time ODI in our analysis to ensure the denominator for ODI calculation was the same for both the AL and Comp. In our study, based on the calculated odds ratio, a patient has 2.9 times greater odds of a diagnosis of moderate-severe OSA if an AL oximeter is used compared to a Comp oximeter. Differences as large as this are likely to affect clinical decision-making about whether to pursue further evaluation (eg, in the patient with a high pretest probability of OSA but “normal” ODI) and when to implement treatment.

There are three potential reasons why oximeters could give different ODI values: (1) oximeter acquisition and internal processing, (2) patient factors related to the use of different fingers, or (3) algorithm differences in the way desaturation events are defined. Pulse oximeters rely on light-emitting diodes, light sensors, and the differing light absorption characteristics of oxygenated and deoxygenated hemoglobin. Multiple stages of signal processing are required to produce a SpO2. These include: light and wavelength measurement characteristics, artefact rejection, sampling frequency, signal averaging times, accuracy and reproducibility. Manufacturers typically validate oximeters to an accuracy of ± 2%. All of these may influence both the baseline SpO2 and the temporal characteristics of any change in SpO2. Signal averaging time is likely to have the greatest potential effect in the setting of OSA, as apneas and hypopneas lead to short-lived oxygen desaturations, and longer signal averaging may therefore underestimate the extent of oxygen desaturation during a respiratory event. AHI is known to vary significantly with different signal averaging times. With respect to patient factors, the variable most likely to affect SpO2 is poor peripheral tissue perfusion, which could potentially differ between fingers or hands. Poor perfusion reduces the pulsatile flow in the finger and hence the accuracy of measured SpO2. Finally, the ODI calculation algorithm is important. Just as AHI will vary according to the scoring criteria used, the “rules” used to score an oxygen desaturation event will also influence the ODI. Key factors include the SpO2 level used as baseline prior to the “dip” and rise in SpO2, but also the duration of SpO2 required for a desaturation to be recognized by the algorithm. Other temporal characteristics such as the rate of SpO2 decline could affect the ODI. Each oximeter and/or sleep system manufacturer uses its own algorithm to calculate ODI and there is no standardized set of “rules” in place to guide manufacturers or clinicians, such as there is with AHI. Moreover, the algorithm used by each device is not readily available for clinicians to refer to when interpreting ODI data. All of these factors may lead to important differences in ODI between oximeters and future research is needed to address how some of these factors differ between devices.

A key strength of our study is that our results demonstrate that the differences in ODI between the AL and the Comp are likely due to oximeter acquisition and processing factors, patient-related factors or ODI calculation algorithms. We exported the oximeter data from the AL into the Comp software. ODI3% and ODI4% were then generated using the Comp software and compared to the ODIs measured and calculated by the AL software. In this situation, there was not only an extremely strong correlation between ODIs (r2 = 0.99, see Figure 2C and Figure 2D), but the Bland-Altman analysis demonstrated excellent agreement. There was also no clinically significant bias between software systems, and 95% limits of agreement were extremely narrow (see Figure 3C and Figure 3D). These results mean that there was no significant effect of the ODI calculation algorithm to explain the variability seen in our primary analysis, rather that the data including any prealgorithm processing was the difference.

Although it is possible for the differences in ODI to be due to the use of separate fingers and arms, we think this is unlikely given the consistent bias seen when AL ODI was compared with Comp ODI in Bland-Altman analysis. If random patient effects such as finger tissue perfusion were to be the cause, one would expect minimal bias as the variability would extend in either direction, but wide limits of agreement (although note that we did not use a formal randomization process in selecting fingers for pulse oximetry monitoring). Moreover, oximeter processing factors are the likely explanation for our results as the processing parameters between the AL and the Comp are known to be different. The AL samples and records SpO2 data at 1 Hz and uses a 3 second signal averaging time to create the
final output value. In contrast, the Comp samples SpO2 every heartbeat. The most recent seven samples are used to calculate the SpO2 output value; the highest and lowest values in that block of data are excluded and the remaining five data points are averaged to provide the final SpO2 data point. In addition, how each device determines and rejects artefact (as compared to a valid signal) is not known, but each is unlikely to deal with artefact in the same way.

Our study has a few limitations. Although we demonstrated that the ODI generated by the AL and the Comp are different, we cannot say which system produces a more accurate ODI. There is no universally accepted gold standard for either oximeter processing or ODI algorithm calculation, and there is not even an accepted standard for manual scoring of ODI. The most important limitation, however, is that our results relate only to these two devices and their respective software, and results cannot be generalized to other oximeters and sleep diagnostic systems. Nevertheless, the important take-home message from our study is that ODI values cannot be directly compared between patients, unless one is comparing data acquired with the same oximeter and software. Clinicians need to know how an oximeter processes SpO2 data and how an ODI is calculated, and whether there is a bias for one oximeter to overestimate or underestimate an ODI compared to the other. It is also important that other comparative studies such as ours are performed with other manufacturers’ equipment and software, and hopefully this study will stimulate further research into the area.

In summary, our study has demonstrated that there is a clinically significant difference in ODI measured by two common sleep diagnostic systems—the ResMed ApneaLink Plus device (AL) and the Compumedics Grael Profusion PSG3 system (Comp). There is a bias for the AL to report higher ODI values, both for ODI3% and ODI4%, but with wide limits of agreement. The differences are large enough to significantly affect diagnostic thresholds for OSA and, in particular, moderate-severe OSA. The differences are likely the result of signal processing rather than patient factors or manufacturer algorithms for scoring desaturations. This has implications for how clinicians use ODI as a diagnostic tool, or as a marker of cardiovascular risk in OSA. Caution is advised when comparing ODI between patients or when performing posttreatment reassessment in the same patient, unless the same oximeter and software algorithm have been used.

**REFERENCES**


