Closing the Gender Gap in Sleep Apnea Diagnosis: AI-Enhanced Home Sleep Testing

Þ. Sigmarsdóttir[1], V. Valsdóttir[1], S. Jónsson[1], E. Erlingsson[1], K. Montazeri[1], E. Finnsson[1], D. Wilcox[1], G. Árnadóttir[1], H. Katrínardóttir[1], M. McPhee[1], R. Hugosdóttir[1], S. Helgason[1], H. Riney[2], J.S. Ágústsson[1]

1. Nox Research, Nox Medical ehf, Reykjavík, Iceland
2. Nox Research, Nox Health llc., Alpharetta, GA, United States.

Presented at SLEEP 2025, Seattle, Washington, USA

thora@noxmedical.com

Introduction

Home sleep testing (HST) is the primary method for diagnosing sleep apnea. However, traditional HST cannot measure sleep and arousals, leading to the systematic underdiagnosis of patients who predominantly experience hypopneas that terminate in arousals. This issue disproportionately affects women. In this study, we present the validation of a novel AI algorithm, Nox AI Scoring (FDA cleared software medical device DeepResp K241960), which uses respiratory inductance plethysmography (RIP) signals to score sleep and arousals from a conventional HST.

The aim of this study was to validate an AI-based algorithm (DeepRESP) to improve sleep apnea diagnosis in women using HST, thereby addressing the diagnostic gender gap and shortening the path to treatment.

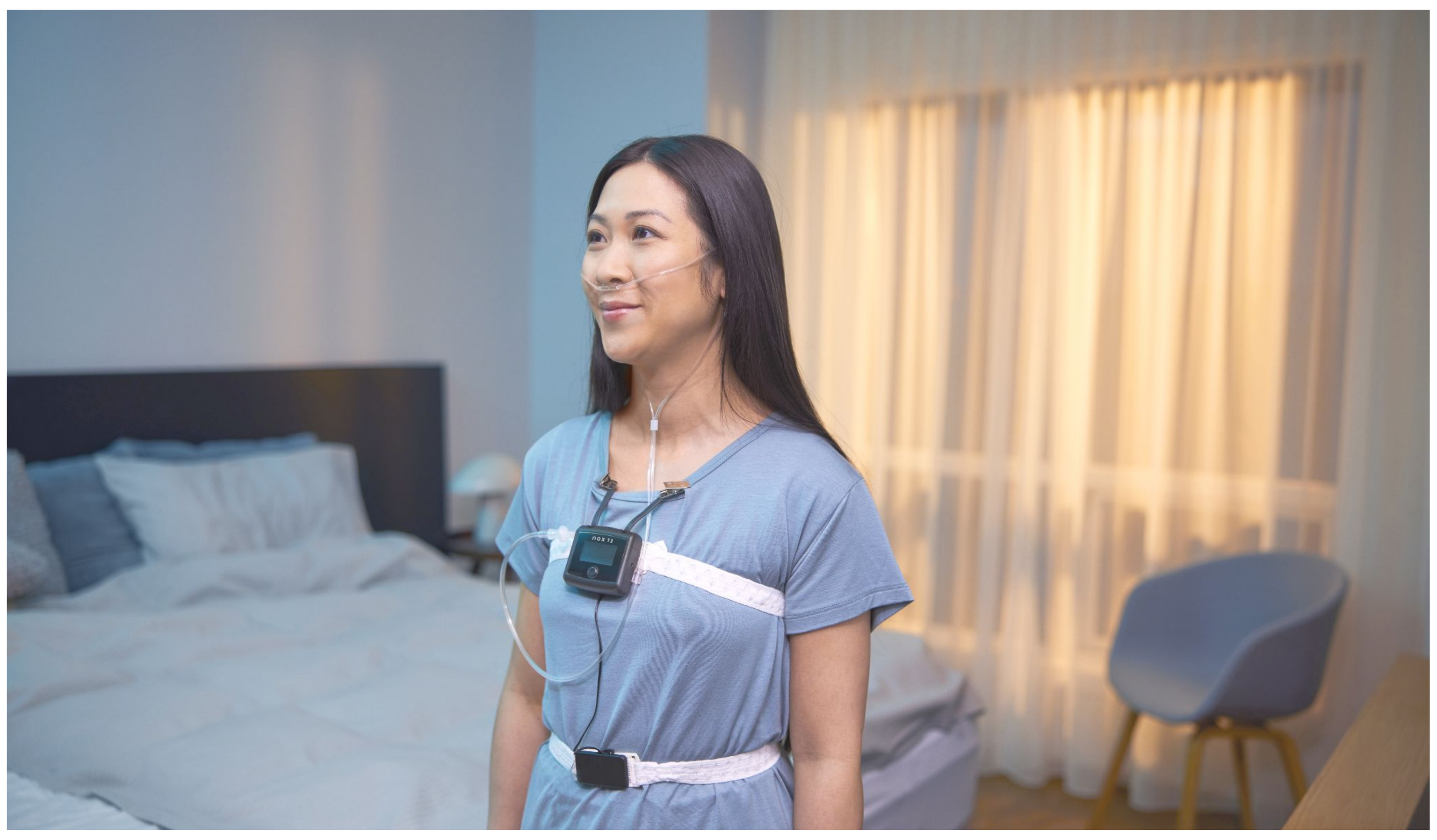

Figure 1: Image of the Nox T3s Home Sleep Testing Device. Using signals from the RIP belts, the DeepRESP algorithm can aid in improving detection of sleep apnea in women from home testing, by enabling sleep state and arousal scoring.

Methods

The algorithm was validated using 1,017 clinical PSG sleep recordings from female subjects. These recordings were manually scored using EEG, EOG, and EMG to determine sleep stages and arousals. The **same recordings** were then automatically scored as if they were HST recordings, omitting any ExG signals, using **two methods.** First, a standard HST scoring was used as a reference, where respiratory events and oxygen desaturations are scored without considering sleep and arousal scoring, resulting in an **REI value**. Then, the test method was the described algorithm, which used RIP signals to score sleep and arousals, along with respiratory events, and incorporated oximetry signal solely to detect oxygen desaturations, resulting in an **AHI value.** The results from these two methods were compared to original **manual scoring**

Figure 2 illustrates the comparison process leading to the reported results. Sensitivity, specificity, and accuracy were calculated when classifying the sleep recordings from women as having an REI/AHI value of 5 or 15 and above.

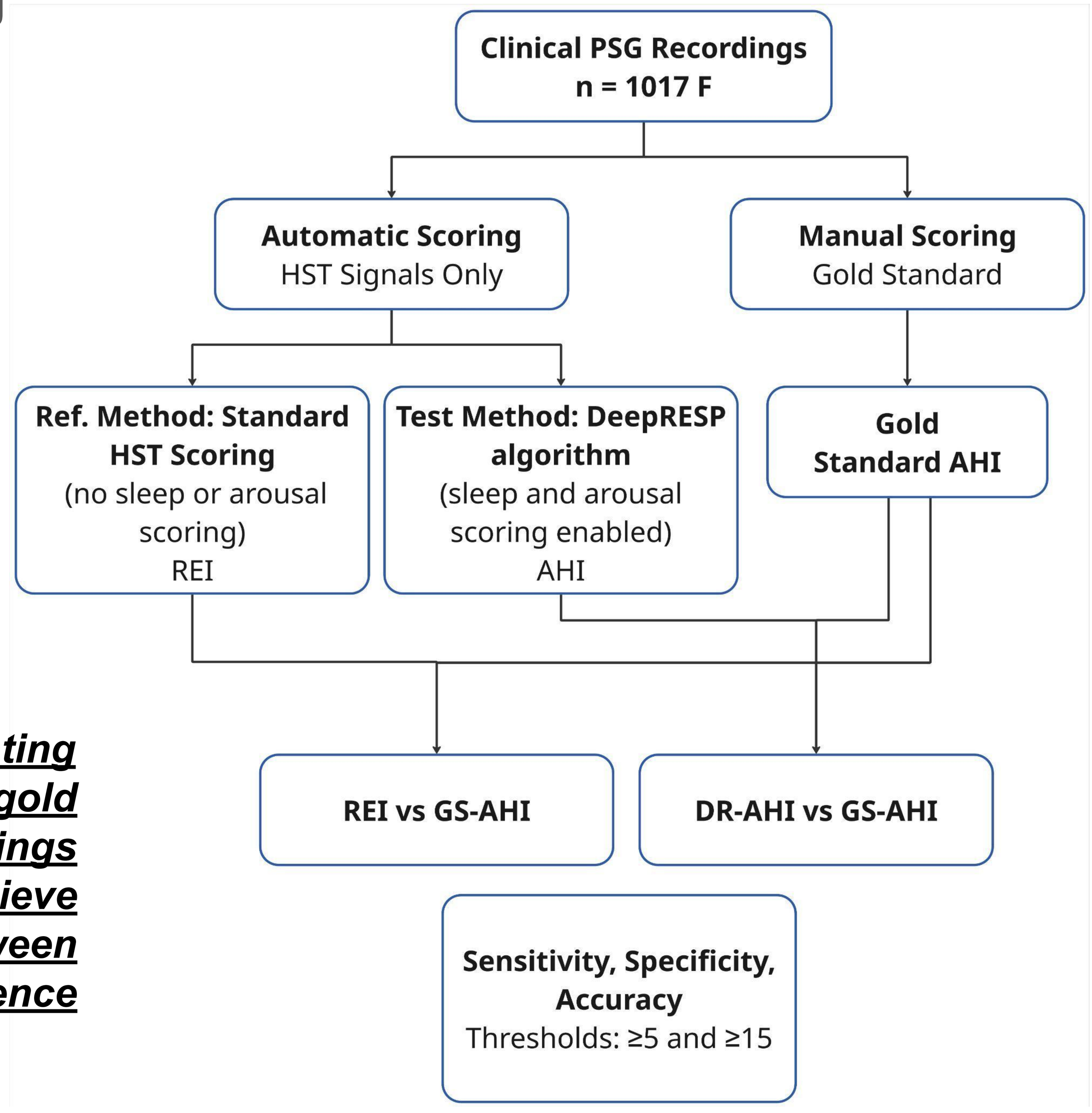

Figure 2: A flow diagram illustrating how the original gold standard PSG recordings were utilised to achieve comparable results between the test and reference methods.
GS: Gold Standard

Results

The test method demonstrated sensitivity, specificity, and accuracy of 85%, 86%, and 85%, respectively, for classifying AHI ≥ 5. In comparison, the reference method showed 73%, 76%, and 73% for the same threshold. For AHI ≥ 15, the test method had sensitivity, specificity, and accuracy of 65%, 94%, and 79%, while the reference method achieved 46%, 97%, and 69%. Most misclassifications occur close to the cut-off threshold. DeepRESP has significantly fewer misclassifications than the reference method. **Notably, the test method correctly classified 122 (12%) more sleep recordings as having AHI ≥ 5 compared to the reference method.**

Table 1: Sensitivity, specificity and accuracy of DeepRESP, vs. standard HST scoring compared to the gold standard, with regards to classifying AHI ≥ 5 and AHI ≥ 15

	Sensitivity % [95%CI]	Specificity % [95%CI]	Accuracy % [95%CI]
	AHI ≥ 5		
DeepRESP	85 [83, 87]	86 [79, 92]	85 [83, 87]
Standard HST	73 [70, 76]	76 [68, 84]	73 [71, 76]
	AHI ≥ 15		
DeepRESP	65 [62, 69]	94 [92, 96]	79 [76, 81]
Standard HST	46 [42, 50]	97 [95, 98]	69 [67, 72]

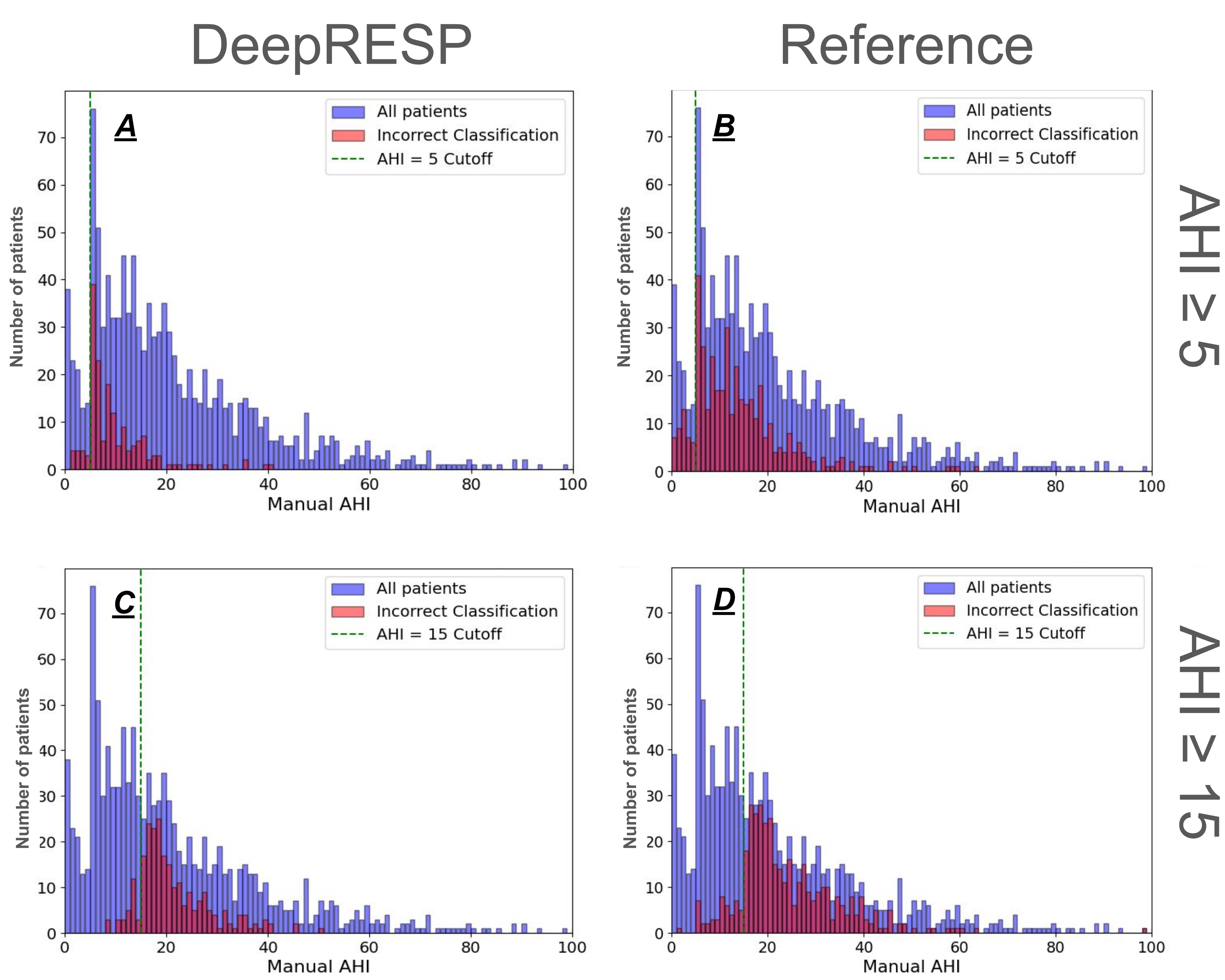

Figure 3: Histograms of patients' AHI values (purple), mis-classified sleep recordings (red) by DeepRESP (A) and (C) and standard HST scoring (B) and (D). The green dashed line shows the decision boundary. DeepRESP has fewer misclassifications than the standard HST scoring when compared to gold standard AHI. Most misclassifications for DeepRESP occur close to the decision boundary.

Conclusions

Females are more likely than males to have sleep apnea characterized by hypopneas that terminate in arousals. Consequently, they often receive inconclusive results from HST and require follow-up tests. This increases their risk of discontinuing medical assistance for sleep apnea. By incorporating respiratory-based sleep and arousal scoring, females are more likely to receive conclusive HST results, thereby shortening their path to treatment.