Sleep Medicine 107 (2023) 18–25

Contents lists available at ScienceDirect

Sleep Medicine

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

Automatic sleep staging for the young and the old – Evaluating age bias in deep learning

Mathias Baumert [a, *], Simon Hartmann [b], Huy Phan [c, 1]

[a] Discipline of Biomedical Engineering, School of Electrical and Mechanical Engineering, The University of Adelaide, Adelaide, Australia
[b] Adelaide Medical School, The University of Adelaide, Adelaide, Australia
[c] Amazon Alexa, Cambridge, MA, 02142, United States

ARTICLE INFO

Article history:
Received 4 October 2022
Received in revised form
26 March 2023
Accepted 1 April 2023
Available online 13 April 2023

Keywords:
Sleep staging
machine learning
polysomnography

ABSTRACT

Background: Various deep-learning systems have been proposed for automated sleep staging. Still, the significance of age-specific underrepresentation in training data and the resulting errors in clinically used sleep metrics are unknown.
Methods: We adopted XSleepNet2, a deep neural network for automated sleep staging, to train and test models using polysomnograms of 1232 children (7.0 ± 1.4 years) and 3757 adults (56.9 ± 19.4 years) and 2788 older adults (mean 80.7 ± 4.2 years). We developed four separate sleep stage classifiers using exclusively pediatric (P), adult (A), older adults (O) as well as PSG from mixed cohorts: pediatric, adult, and older adult (PAO). Results were compared against an alternative sleep stager (DeepSleepNet) for validation purposes.
Results: When pediatric PSG was classified by XSleepNet2 exclusively trained on pediatric PSG, the overall accuracy was 88.9%, dropping to 78.9% when subjected to a system trained exclusively on adult PSG. Errors performed by the system staging PSG of older people were comparably lower. However, all systems produced significant errors in clinical markers when considering individual PSG. Results obtained with DeepSleepNet showed similar patterns.
Conclusion: Underrepresentation of age groups, in particular children, can significantly lower the performance of automatic deep-learning sleep stagers. In general, automated sleep stagers may behave unexpectedly, limiting clinical use. Future evaluation of automated systems must pay attention to PSG-level performance and overall accuracy.

1. Introduction

Overnight polysomnography (PSG) is the cornerstone of sleep medicine, collecting a range of physiological signals. Experts manually interpret these physiological signals based on consensus rules, which are prone to bias and error. By convention, sleep is separated into a discrete set of stages, and observing the temporal development yields the architecture of sleep upon which sleep processes are assessed. Historically, sleep staging was performed manually on 30-s time frames by a sleep technician visually evaluating EEG rhythms, EOG and EMG patterns.

Recent advances in digital signal processing and machine learning have paved the way to automate this labour-intensive and tedious process fully. In particular, deep learning approaches have achieved a sleep staging performance that is on par with human experts [1]. Under the current American Academy of Sleep Medicine (AASM) scoring framework, interscorer variability between human experts is, on average, about 83%; it varies substantially for stage N1 but also for N3 [2,3]. The current AASM position is that 'artificial intelligence' should augment but not replace the expert evaluation of PSG [4]. Technology for automated sleep staging is considered mature and has found its way to commercial sleep evaluation systems, including wearables [5].

Key to the tremendous success of deep learning is the availability of large datasets providing a broad sample of sleep stages that the machine seeks to 'learn' by generalizing the patterns underpinning sleep. Large publicly available PSG datasets, such as the

* Corresponding author. The University of Adelaide, School of Electrical and Mechanical Engineerin, Adelaide, SA5005, Australia.
E-mail address: mathias.baumert@adelaide.edu.au (M. Baumert).
[1] The work was done when H. Phan was at the School of Electronic Engineering and Computer Science, Queen Mary University of London, UK and the Alan Turing Institute, UK and prior to joining Amazon.

https://doi.org/10.1016/j.sleep.2023.04.002

Sleep Heart Health Study (SHHS) [6] and National Sleep Research Resource [7], and Montreal Archive of Sleep Studies (MASS) [8] provide sufficiently large training and testing data for deep learning systems.

Of concern, most PSG databases typically used for developing sleep stagers include only recordings from adults, creating a significant inherent sample bias and potentially large scoring errors that may result in substantial errors when applied in the pediatric or geriatric sleep setting. Aside from age bias, databases often feature particular clinical populations, introducing further sampling bias.

Compared to adults, NREM sleep in children contains more slow-wave sleep and displays higher theta, alpha, sigma and beta frequency band power in EEG [9]. The sleep architecture in children is characterized by an increase in N2 sleep at the expense of N3 [10]. In contrast, older people produce lower slow-wave amplitudes, slow delta waves and discrete spindles, lowering N3 sleep but increasing N2 sleep [11]. Further, most sleep stagers are evaluated by the overall accuracy they achieve for specific sleep stages across all epochs and total sleep time sleep efficiency. Also, the error at the level of the individual PSG and its effect on typical clinically used markers of sleep architecture often receives little attention when evaluating algorithms.

Here, we evaluate the significance of age-specific underrepresentation in training data for automated sleep stagers and the resulting errors in clinically used sleep metrics.

2. Methods

2.1. Automatic sleep stager architecture

We adopted XSleepNet2, the deep neural network presented in Ref. [12], as the automatic sleep staging model (Fig. 1). It achieves state-of-the-art sleep-staging performance on a wide range of sleep databases. This model adheres to the sequence-to-sequence automatic sleep staging concept, which recently achieved expert-level performance in automatic sleep staging [12]. More specifically, the model is devised to handle a sequence of consecutive epochs (20 epochs in this study) as input and maps them into a sequence of sleep stages at once. The capability of processing a series of epochs allows the model to incorporate possible long-range dependencies between sleep epochs in the input sequence, which is vital in improving sleep staging performance. It resembles how sleep experts perform manual scoring, who typically need to attend to a large context around a target epoch to determine its stage.

XSleepNet2 features a hybrid architecture and simultaneously ingests two different signal representations, the raw signal and time-frequency image, as the input; the time-frequency image was extracted as described in Ref. [12]. More specifically, the raw signal of a 30-s epoch (i.e. EEG or EOG) was transformed to a time-frequency image via short-time Fourier transform (STFT) with a window size of 2 s, 50% overlap, Hamming window, and 256-point Fast Fourier Transform (FFT), followed by logarithm scaling. Two complementary deep neural subnetworks handle the two representations separately. One subnetwork receives the raw signal and combines a convolutional neural network (CNN) and a recurrent neural network (RNN) for epoch-wise encoding and sequence-wise encoding, respectively. The other processes time-frequency representation as input and relies on two RNNs, one for epoch-wise encoding and one for sequence-wise encoding. The first subnetwork has many parameters (i.e. strong modelling capacity) to leverage a large training data size by design. In contrast, the second subnetwork has few parameters (i.e. limited modelling capacity) to avoid overfitting when the training data size is small. During model training, the learning pace of the two subnetworks is adapted according to their generalization and overfitting behaviour such that the one generalizing well is rewarded while the one overfitting is penalized. As a result, the network as a whole learns a good representation from both the raw-signal and time-frequency image inputs and gains robustness to the amount of training data.

To explore the generalisability of our findings, we tested the performance of an alternative, popular deep learning-based sleep stager, DeepSleepNet, widely used in the literature [13]. In brief, 30-s EEG epochs are subjected to convolution layers for representation learning. Temporal associations are learned via bidirectional long short-term memory (LSTM) layers.

2.2. Polysomnographic data

The experimental data comprised 7777 overnight PSG recordings with at least 5 h of valid EEG data pooled from the Childhood Adenotonsillectomy Trial (CHAT) [14,15], Cleveland Family Study (CFS) [16], Multi-Ethnic Study of Atherosclerosis (MESA) [17], Osteoporotic Fractures in Men Study (MrOS Sleep) [18], Cleveland Children's Sleep and Health Study (CCSHS) [19], and Study of Osteoporotic Fractures [20]. Epochs with all zeros (due to poor electrode contact) or annotated "UNKNOWN" (due to signal artifacts) were considered invalid and discarded from analysis. Further, EEG values exceeding six standard deviations were clipped.

We used single-channel EEG (C4-A1) and EOG (ROC-LOC) as model inputs for the automated sleep staging systems. The EEG and EOG signals, originally sampled with different sampling frequencies, were resampled to 100 Hz and bandpass filtered with a low cut-off frequency of 0.3 Hz and a high cut-off frequency of

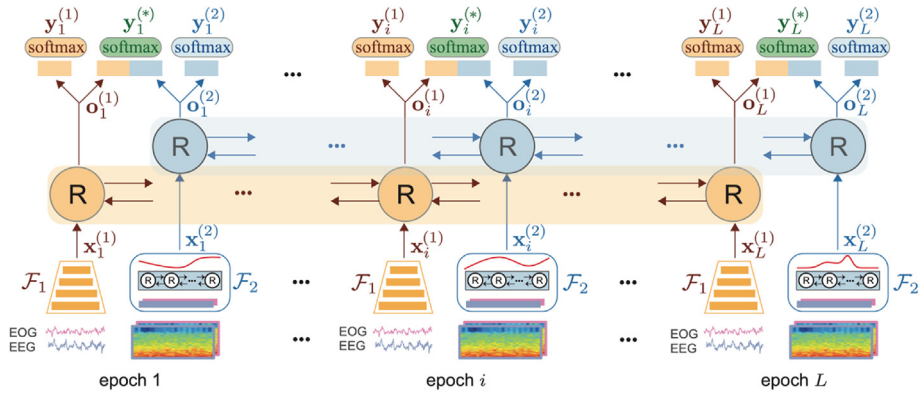

Fig. 1. Schematic representation of the XSleepNet2 deep neural network, adapted from Ref. [12].

40 Hz using a 100th-order bandpass finite impulse response (FIR) filter. Each recording was normalized to the range [−1, 1].

To quantify the effect of the training sample's age bias on the automatic sleep stager output, we divided the data set into three age groups (Table 1): children aged from 4.5 to 10 years, adults aged from 16.1 to 74.8 years, and older adults aged from 75.0 to 96.0 years. The age cut-off values were chosen based on the age characteristics of the datasets. We excluded participants older than ten and younger than 16 years from our pediatric sample to reduce the impact of developmental changes on sleep EEG associated with adolescence.

2.3. Statistics

The set of pediatric PSG was randomly divided into five equal subsets for 5-fold cross-validation. Similar splitting was performed for the adult and older adult datasets. We developed four separate sleep stage classifiers using exclusively pediatric (P), adult (A), older adult (O), as well as PSG from mixed cohorts: pediatric, adult, and older adult (PAO). The deep neural network, i.e. XSleepNet2, was configured as in the original work [12]. The performance of the four sleep stager models was assessed with confusion matrices. To evaluate the clinical relevance of sleep staging performance, we calculated several typical hypnogram metrics used in the clinical setting: total sleep time (TST), sleep efficiency (SE), time of Wake, N1, N2, N3 and REM sleep, Wake after sleep onset (WASO) and REM sleep latency.

3. Results

The characteristics of the study sample are summarised in Table 2. As expected, TST shortened with increasing age, and conversely, SE was highest in children and lowest in older adults. Likewise, wake time and WASO were shortest in children and longest in older adults. Children spent marginally less per cent of sleep in N1 than adults and older adults. The percentage of N2 increased with age while N3 decreased. There was no notable age trend in the portion of REM sleep.

Confusion matrices for the XSleepNet2 sleep stager models are shown in (Fig. 2, top). When pediatric PSG was classified by the system exclusively trained on pediatric PSG (P), the overall accuracy was 88.9%. But the accuracy significantly dropped when pediatric PSG was subjected to a system trained exclusively on adult PSG (A) (78.9%). Using a system trained on mixed cohort PSG (PAO), the overall accuracy for pediatric PSG was comparable to that of the pediatric sleep stager (87.5%). Across all three models, stage N1 classification was most error-prone, dropping as low as 21.7% when using the adult system compared to 66.6% and 62.1% for P and PAO. In contrast, the Wake and REM classification was most accurate, reaching >90% for P and PAO and 71.3% and 95.0% for the adult sleep stager. Stage N2 and N3 classification exceeded 85% for all but the adult sleep stager.

Considering the sample of older people (Fig. 2, bottom), the overall sleep stager accuracy was 88.4% when the system was trained exclusively on an older population and 88.1% when solely trained on the adult sample. When trained on the entire cohort, including children (PAO), the accuracy was 88.1%. Similar to the PSG of children, N1 classification accuracy was the poorest (around 40%), while Wake and REM classification exceeded 90%, with N2 classification marginally worse. Of interest, N3 classification was notably less accurate (around 60–65%), particularly when compared to pediatric data (87%–92% accuracy).

DeepSleepNet achieved marginally lower overall classification accuracies than XSleepNet2 but displayed a similar pattern of performance drop when pediatric PSG was scored on a system trained exclusively on adult data (Table S1). Sleep stage classification accuracy showed a similar pattern (Fig. S1),

Fig. 3 shows the errors common to all XSleepNet2 models, stager-specific errors distributed over the five sleep stages, and the distances of misclassified epochs to their nearest sleep stage transition. In the case of pediatric sleep staging (top left), light sleep contributed the most to the set of common errors, and N2 is the most misclassified stage. The stager-specific errors exhibit clear discrepancies between the stager trained with adult data only (A) and those trained with pediatric PSG (P and PAO) data. Adult PSG classification resulted in a noticeably higher amount of misclassified Wake and N1, while a significantly lower amount of misclassified N3 and REM. Such a difference was not observed in the case of the sleep stager developed on older adult data (top-right), where stager-specific errors are distributed across the sleep stages similarly, suggesting fewer differences between adult and older adult PSG compared to pediatric data.

Regarding the relative position of misclassified epochs to their closest sleep stage transitions, 55% of the common pediatric stager errors occurred within four epochs from their closest transitions. Out of 55%, less than 20% are rapid-transition epochs. In the case of the older adult sleep stager, the corresponding percentage is around 65%, of which 25% are rapid-transition epochs. The stager-specific errors generally saw fewer epochs distributed in the vicinity of 4 epochs than the common ones. This is observed in both cases of pediatric and older adult staging. However, no striking differences are seen among the distributions of the stager-specific errors. The sleep stage classification errors produced by DeepSleepNet were very similar (Fig. S2).

Table 1
Study samples.

Database	Children N = 1232	Adults N = 3757	Older Adults N = 2788
Age (mean ± sd)	7.0 ± 1.4 y	56.9 ± 19.4 y	80.7 ± 4.2 y
[Range]	[4.5–10 y]	[16.1–74.8 y]	[75.0–96.0 y]
Sex (male, female, unknown)	578 m, 633f,21u	2415 m, 1342f	1968 m, 820f
Race			
white	482	2166	2205
black	577	964	263
Other/unknown	173	627	320
Source databases (N)	CFS (N = 18), CHAT (N = 1214)	CFS (N = 606), MESA (N = 1401), MrOS (N = 1235), CCSHS(N = 515)	CFS (N = 33), MESA (N = 652), MrOS(N = 1651), SOF(N = 452)

CFS = Cleveland Family Study, CHAT = Childhood Adenotonsillectomy Trial, MESA = Multi-Ethnic Study of Arteriosclerosis, MrOS = Osteoporotic Fractures in Men, SOF = Study of Factures.

Table 2
Hypnogram characteristics of the pediatric, adult and older adults study sample obtained by manual expert scoring.

	Children (n = 1232)	Adults (n = 3757)	Older Adults (n = 2788)	p-value
Age [years]	7.0 ± 1.4	56.9 ± 19.4	80.7 ± 4.2	<0.0001
TST [min]	451.4 ± 55.2	375.2 ± 83.4	344.8 ± 76.5	<0.0001
Wake [min]	109.7 ± 53.9	253.4 ± 105.1	286.9 ± 115.4	<0.0001
N1 [%]	8.2 ± 4.1	8.6 ± 7.3	9.1 ± 7.4	0.0003
N2 [%]	41.7 ± 8.6	58.9 ± 10.9	62.2 ± 13.9	<0.0001
N3 [%]	32.0 ± 8.9	14.0 ± 10.4	12.8 ± 12.1	<0.0001
R [%]	18.3 ± 4.3	19.6 ± 6.7	18.5 ± 7.4	<0.0001
WASO [min]	43.3 ± 38.2	84.1 ± 60.5	117.5 ± 70.9	<0.0001
SE [%]	81.5 ± 9.0	66.1 ± 12.6	61.2 ± 13.5	<0.0001
REM onset latency [min]	223.8 ± 68.3	219.3 ± 97.4	212.8 ± 103.0	<0.0001

TST – total sleep time; WASO – Wake after sleep onset; SE – sleep efficiency.

Fig. 2. Confusion matrices of XSleepNet2 output subjected to pediatric PSG (top) and PSG of older people (bottom) compared to manual scoring. A-trained – Sleep stager trained exclusively on adult PSG; P-trained – Sleep stager trained exclusively on pediatric polysomnograms; O-trained – Sleep stager solely trained on older people; PAO-trained – Sleep stager trained on a mixed cohort of children, adults and older adults.

Table 3 summarises the relative errors in the clinical hypnogram-based metrics for the pediatric study sample introduced by XSleepNet2 compared to manual scoring. Considering the system trained exclusively on pediatric data, the interquartile range of errors exceeded 10% for N1, WASO and REM latency. Similar results were obtained using the system trained on the mixed cohort. However, when using the system exclusively trained on adults, errors were notably more significant, with the IQR exceeding 10% for all metrics but total sleep time and efficiency. Considering the sample of older adults, the IQR of the relative error exceeded 10% for all sleep stages except Wake. The IQR of errors was below 10% for total sleep time, REM latency and sleep efficiency across all XSleepNet2 models. The errors in hypnogram-based metrics obtained with DeepSleepNet were comparable in magnitude (Table S2).

Fig. 4 visually represents the complete range of relative errors by the XSleepNet2 subjected to pediatric data as massive individual errors are introduced by the automated sleep stager, particularly for the relative percentage of sleep stages when staging pediatric PSG with the adult sleep stager. Also, WASO and REM latency may be estimated with more than 100% error. Fig. 5 visualizes the range of errors produced by XSleepNet2 when staging the PSG of older people. Similar to children, sleep stage distribution, in particular, can lead to significant errors. But also, other metrics were estimated with errors larger than 100% in individual cases. Individual errors produced by DeepSleepNet can become similarly significant (Figs. S3 and S4).

4. Discussion

This study explored the errors produced by two state-of-the-art automated sleep staging systems created by sampling bias towards middle-aged women and men. We observed a critical increase in classification errors if pediatric data were assessed by sleep staging models trained only on middle-aged adults. The problem can be effectively rectified when including a representative training sample of pediatric PSG in developing the sleep stager. Looking at older people, the error caused by sleep stagers trained on middle-aged adult data is substantially less than that experienced with pediatric data. Indeed, adding a pediatric PSG appears to reduce the median errors marginally. Thus, a single comprehensive automatic sleep stager for a wide age range of patients starting from preschooler with clinical-grade performance seems feasible if the training data encompasses a broad representation. However, when considering individual PSG, errors in clinical markers produced by any sleep stager models can become unacceptably high, raising

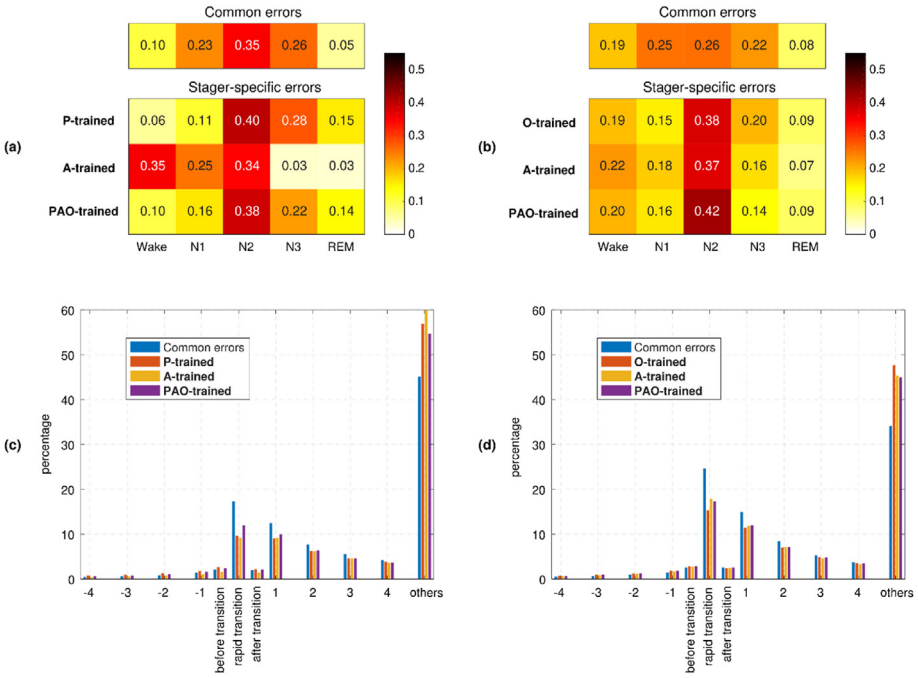

Fig. 3. The top panels show common errors across all XSleepNet2 stagers and stager-specific errors distributed over the five sleep stages. The distances of misclassified epochs to their nearest sleep stage transition are shown in the bottom panels. A-trained – Sleep stager trained exclusively on adult PSG; P-trained – Sleep stager trained on pediatric polysomnograms; O-trained – Sleep stager trained exclusively on older people; and PAO-trained – Sleep stager trained on a mixed cohort of children, adults and older adults.

Table 3
Relative errors in typical clinically used measures of sleep architecture produced by XSleepNet2 trained on children (P-trained), adults (A-trained), older participants (O-trained) or the mixed cohort (PAO-trained) when subjected to pediatric data or data from older participants. Errors are expressed in per cent as medians and interquartile ranges. Errors exceeding 10% are printed in bold letters for convenience.

	children			older adults		
Sleep metric	P-trained	A-trained	PAO-trained	P-trained	A-trained	PAO-trained
TST [min]	0.083 (-0.625 - 0.853)	1.902 (0.564–4.65)	0.098 (-0.660 - 0.999)	0.298 (-2.758 - 3.524)	0.524 (-2.521 - 3.875)	0.567 (-2.427 - 3.906)
Wake [min]	−0.354 (-4.161 - 2.962)	**−19.62** (**-40.11** - −7.792)	−0.7285 (-5.26 - 2.937)	0.4843 (-2.853 -3.968)	0.7569 (-2.375 - 4.479)	0.334 (-1.019 - 1.819)
N1 [%]	6.905 (**-10.68**–**28.06**)	**−66.45** (**-79.28 - -48.11**)	−2.817 (**-19.69**–**18.97**)	**−17.88** (**-41.17**–9.661)	**−24.61** (**-47.36**–3.374)	**−18.87** (**-41.71**–8.947)
N2 [%]	−1.481 (-8.171 - 9.215)	−4.351 (**-15.83**–**10.94**)	−0.113 (-8.100 - **10.37**)	5.93 (-2.225 - **16.36**)	5.948 (-1.696 - **16.59**)	5.016 (-2.439 - **14.74**)
N3 [%]	−2.873 (-13.77–9.193)	7.685 (-6.686 - 27.79)	−0.801 (-11.53–11.41)	**−24.7** (**-62.5**–**12.48**)	**−21.21** (**-55.94**–**20.63**)	**−17.29** (**-52.35**–**23.24**)
REM [%]	3.288 (−4.834 - **12.57**)	**35.05** (**15.1** - **66.91**)	3.398 (−4.877 - 12.66)	0.4573 (-8.182 - **10.57**)	1.538 (-7.225 - **12.3**)	0.041 (-8.393 - **10.23**)
WASO [min]	0 (**-11.01 - 11.21**)	**−33.33** (**-55.4** - **-13.48**)	−0.855 (**-12.5**–**10.49**)	0 (-9.859 - **10.69**)	−1.587 (**-12.38**–9.483)	−1.639 (**-11.99**–8.587)
REM latency [min]	−0.238 (**-17.69**–0.3017)	**−71.58** (**-92.08** - −4.32)	−0.255 (**-21.27**–0.279)	0 (−0.314 - 0.133)	0 (-0.329 - 0.139)	0 (-0.257 - 0.199)
SE [1]	0.083 (-0.625 - 0.853)	1.902 (0.564–4.65)	0.09 (-0.659 - 0.999)	0.027 (-3.295 - 3.219)	0.3669 (-3.18 - 3.491)	0.423 (-2.951 - 3.649)

TST – total sleep time; WASO – Wake after sleep onset; SE – sleep efficiency.

serious questions as to whether “unsupervised” fully automated sleep staging with current state-of-the-art systems is ready for implementation in the clinic.

The explosion of deep learning research has produced a plethora of automated sleep staging systems. Aside from apparent time saving and related cost reduction, simplified EEG montages are an attractive proposition. Typically, these models are benchmarked against readily available adult PSG databases that may suffer from limited data EEG quality, such as the SHHS, due to the data's age and the data acquisition systems available at the time. While the performance of these systems has been deemed sufficient for clinical use, implementation in commercial systems and acceptance by the clinical community is still limited. One key issue is that performance is commonly reported across the entire cohort for testing data rather than individual recordings. While accuracy may be exceptionally high in many PSG recordings, it is crucial to evaluate performance at the PSG level and establish the range of scoring errors that can occur for an individual patient. Unlike human experts, deep learning models have been shown to behave in unexpected ways and occasionally produce unacceptably large errors.

Interestingly, both scoring systems used in this study produced large individual errors, and the pattern of errors appears to be similar, suggesting that automated systems may generally struggle with particular recordings, and this problem is not specific to any

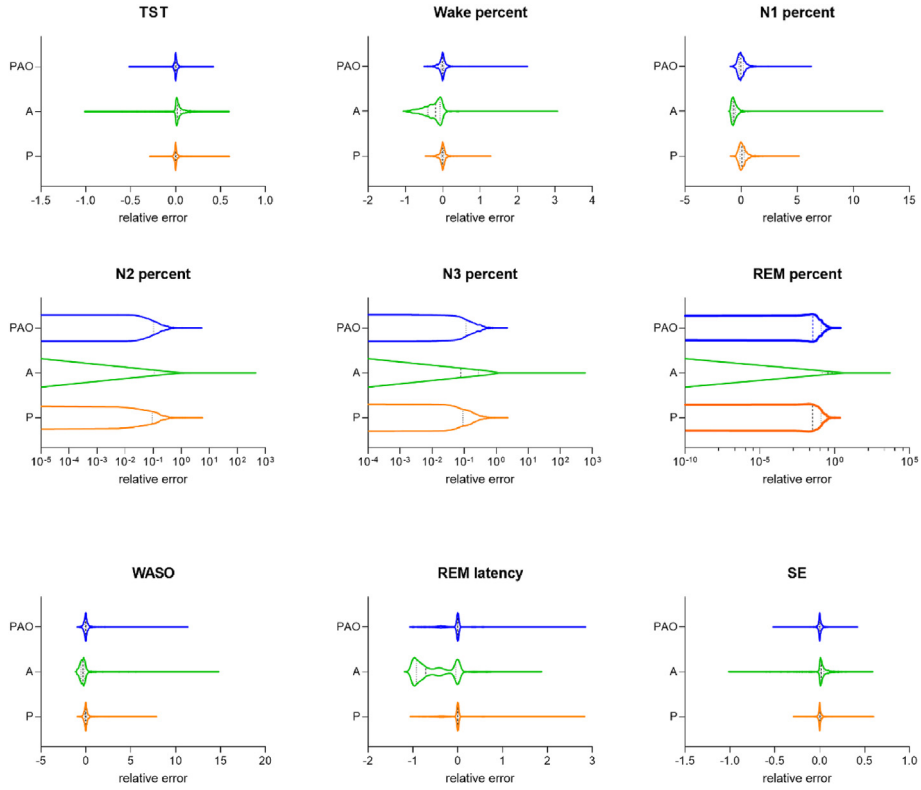

Fig. 4. The relative error in sleep staging of children created by XSleepNet2 trained exclusively on pediatric polysomnograms (P), adults' polysomnograms (A) and mixed cohorts including older adults (PAO) compared to manual scoring. Data are presented as violin plots, including median and interquartile ranges. (TST – total sleep time; WASO – Wake after sleep onset; SE – sleep efficiency).

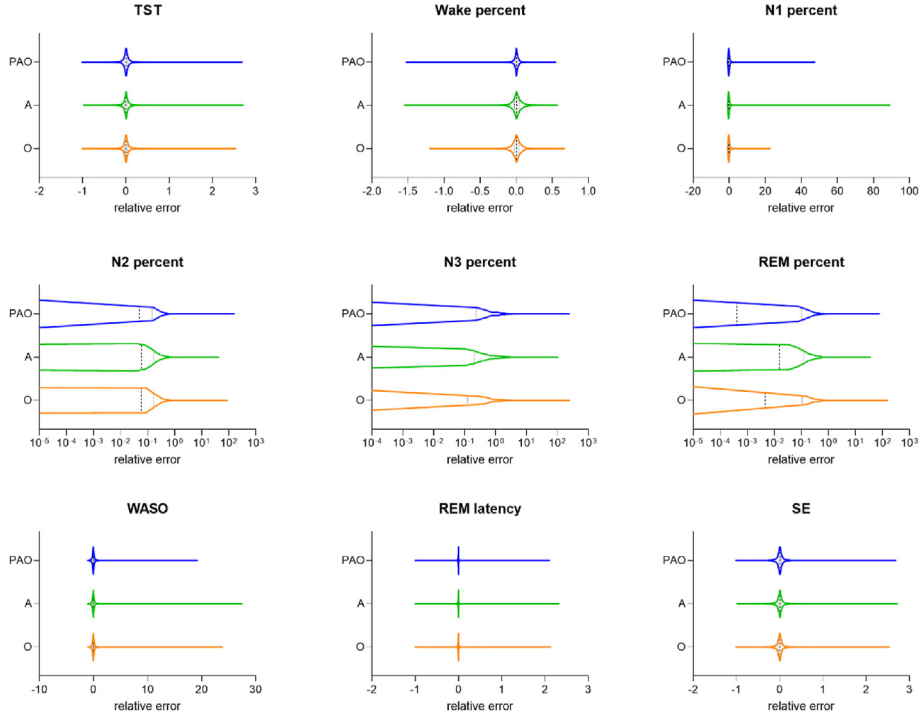

Fig. 5. The relative error in sleep staging of older people created by XSleepNet2 trained exclusively on older people's polysomnograms (O), adults' polysomnograms (A) and mixed cohorts including children (PAO) compared to manual scoring. Data are presented as violin plots, including median and interquartile ranges. (TST – total sleep time; WASO – Wake after sleep onset; SE – sleep efficiency).

particular deep-learning sleep stager. To address these issues, it would appear necessary to output some measure of classification uncertainty that will inform the sleep technician about the uncertainty or reliability of individual PSG scoring results. The system could recommend manual editing or rescoring if the uncertainty is comparably high. From a modelling perspective, this raises the need for an active learning mechanism for a model to be gradually adjusted using the sleep technician's editing/rescoring annotation in a closed-loop manner. Detailed performance assessment in various clinical scenarios will be critical to promoting broad acceptance. Inspecting how a model behaves on the anomalous PSG and understanding how sleep staging errors give rise to errors in clinical markers would pave the way to modelling development, generalization and robustness, narrowing the clinical implementation gap, where models achieve high performance in preclinical testing perform poorly in the clinical setting. Auditing cases where the systems fail may help identify and address the systemic bias of the sleep stager [21]. The lack of sufficiently large data representation of rare medical conditions may present another challenge for deep learning systems. Transfer learning may provide an effective strategy for dealing with rare diseases. Further, implementing deep learning systems into the clinical pathway and identifying practical ways of interaction between the system and the user is necessary to close the implementation gap.

4.1. Limitations

We used a large convenience sample of sleep studies pooled from multiple studies using various EEG recording equipment, constituting a rich data representation. The prevalence of sleep pathologies such as sleep-disordered breathing or periodic limb movement disorder may not reflect the general population. Aside from fragmenting the sleep microstructure, sleep architecture may be affected.

Also, a potential sex bias may affect the performance of deep learning systems. Women show larger slow wave amplitudes than men, which may affect the performance of the sleep stager model [22]. Sleep microstructure was shown to vary in older men and women [23]. Given how well our mixed model generalized age effects, we would anticipate similar effectiveness regarding sex differences.

We did not include PSG from toddlers. Hence our findings cannot be extrapolated to very young children. Given the substantial change in sleep formation and EEG patterns across the first years of life, particularly in infants, automated sleep stagers using deep learning may need to be adapted for those specific age groups. It remains to be seen whether an age-comprehensive single model can serve this age group or whether dedicated systems are necessary for very young children. To address the potential age bias in the deep learning model, age labels minimizing the risk of misdiagnoses may become necessary.

Further, our findings are based on only two sleep stagers. While we cannot generalize our observation to any possible deep learning systems, many current state-of-the-art systems tend to perform and behave similarly [24].

4.2. Conclusion

Underrepresentation of age groups, in particular children, can significantly lower the performance of automatic deep-learning sleep stagers. A single model trained on a broad representative sample of data generalizes well and is preferable to age-specific sleep stager models. In general, automated sleep stagers may behave unexpectedly, limiting clinical use. Future evaluation of automated systems must pay attention to PSG-level performance and overall accuracy.

Authors' contributions

All authors contributed to writing the initial draft.

Role of funding source

None.

Ethics committee approval

Not applicable.

Declaration of competing interest

None of the authors declares a conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.sleep.2023.04.002.

References

[1] Phan H, Mikkelsen K. Automatic sleep staging of EEG signals: recent development, challenges, and future directions. Physiological Measurement; 2022.
[2] Rosenberg RS, Van Hout S. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. J Clin Sleep Med 2013;9(1):81–7.
[3] Penzel T, Zhang X, Fietze I. Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules. J Clin Sleep Med 2013;9(1):89–91.
[4] Goldstein CA, et al. Artificial intelligence in sleep medicine: an American Academy of Sleep Medicine position statement. J Clin Sleep Med 2020;16(4):605–7.
[5] Baumert M, et al. Sleep characterization with smart wearable devices: a call for standardization and consensus recommendations. Sleep 2022;45(12).
[6] Quan SF, et al. The sleep Heart Health study: design, rationale, and methods. Sleep 1997;20(12):1077–85.
[7] Zhang GQ, et al. The national sleep research resource: towards a sleep data commons. J Am Med Inf Assoc 2018;25(10):1351–8.
[8] O'reilly C, et al. Montreal Archive of Sleep Studies: an open-access resource for instrument benchmarking and exploratory research. J Sleep Res 2014;23(6):628–35.
[9] Gaudreau H, Carrier J, Montplaisir J. Age-related modifications of NREM sleep EEG: from childhood to middle age. J Sleep Res 2001;10(3):165–72.
[10] Kahn A, et al. Normal sleep architecture in infants and children. J Clin Neurophysiol 1996;13(3):184–97.
[11] Muehlroth BE, Werkle-Bergner M. Understanding the interplay of sleep and aging: methodological challenges. Psychophysiology 2020;57(3):e13523.
[12] Phan H, et al. XSleepNet: multi-view sequential model for automatic sleep staging. IEEE Trans Pattern Anal Mach Intell; 2021.
[13] Supratak A, et al. DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG. IEEE Trans Neural Syst Rehabil Eng 2017;25(11):1998–2008.
[14] Redline S, et al. The Childhood Adenotonsillectomy Trial (CHAT): rationale, design, and challenges of a randomized controlled trial evaluating a standard surgical procedure in a pediatric population. Sleep 2011;34(11):1509–17.
[15] Marcus CL, et al. A randomized trial of adenotonsillectomy for childhood sleep apnea. N Engl J Med 2013;368(25):2366–76.
[16] Redline S, et al. The familial aggregation of obstructive sleep apnea. Am J Respir Crit Care Med 1995;151(3 Pt 1):682–7.
[17] Chen X, et al. Racial/ethnic differences in sleep disturbances: the multi-ethnic study of Atherosclerosis (MESA). Sleep 2015;38(6):877–88.
[18] Blackwell T, et al. Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: the osteoporotic Fractures in men sleep study. J Am Geriatr Soc 2011;59(12):2217–25.
[19] Rosen CL, et al. Prevalence and risk factors for sleep-disordered breathing in 8- to 11-year-old children: association with race and prematurity. J Pediatr 2003;142(4):383–9.
[20] Spira AP, et al. Sleep-disordered breathing and cognition in older women.

J Am Geriatr Soc 2008;56(1):45–50.
[21] Liu X, et al. The medical algorithmic audit. The Lancet Digital Health; 2022.
[22] Mourtazaev M, et al. Age and gender affect different characteristics of slow waves in the sleep EEG. Sleep 1995;18(7):557–64.
[23] Hartmann S, et al. Characterization of cyclic alternating pattern during sleep in older men and women using large population studies. Sleep 2020;43(7): zsaa016.
[24] Phan H, Mertins A, Baumert M. Pediatric Automatic Sleep Staging: a comparative study of state-of-the-art deep learning methods. IEEE Transactions on Biomedical Engineering; 2022.