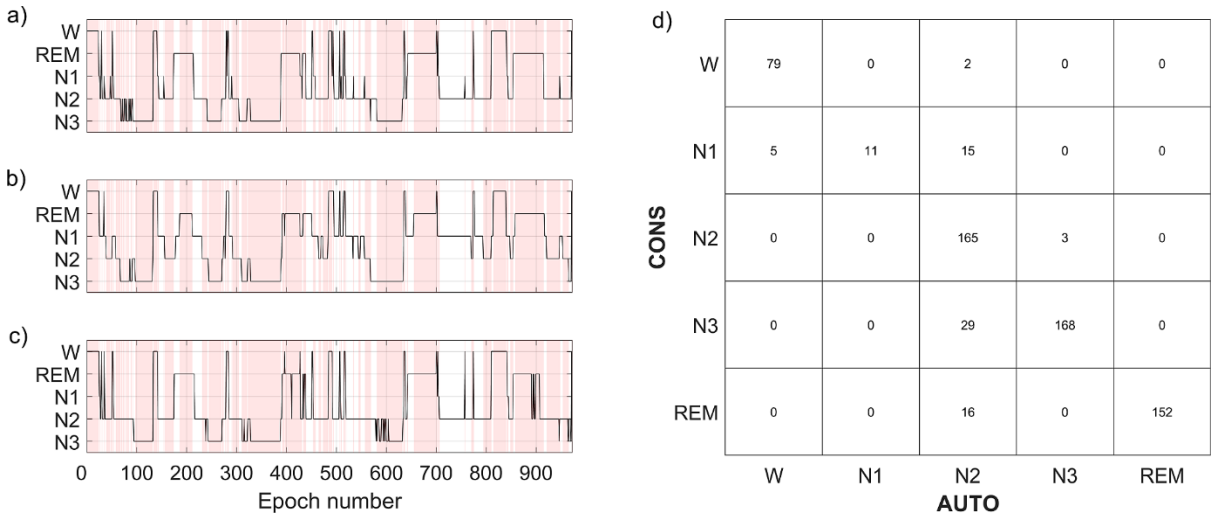


Figure S1: Consensus epochs and confusion matrix CONS-vs-AUTO. For each participant, the consensus (CONS) epochs were the epochs for which manual scorings from Innsbruck (a) and Berlin (b) were in agreement (highlighted in red). These epochs were compared to the respective epochs in the automatically (AUTO) scored hypnogram (c) to obtain the confusion matrix (d). The confusion matrix reports in the diagonal the number of epochs where there was agreement between the consensus and automatic scoring. The elements out of the diagonal show the number of epochs for which there was disagreement and the type of disagreement (e.g. the element in {row 2, column 1} indicates that five epochs were epochs scored as N1 in by human scorers in Innsbruck and Berlin, but as W by the Stanford-STAGES algorithm).

Abbreviations: W: wakefulness; REM: rapid eye movement sleep; N1: non-REM stage 1 sleep; N2: non-REM stage 2 sleep; N3: non-REM stage 3 sleep.

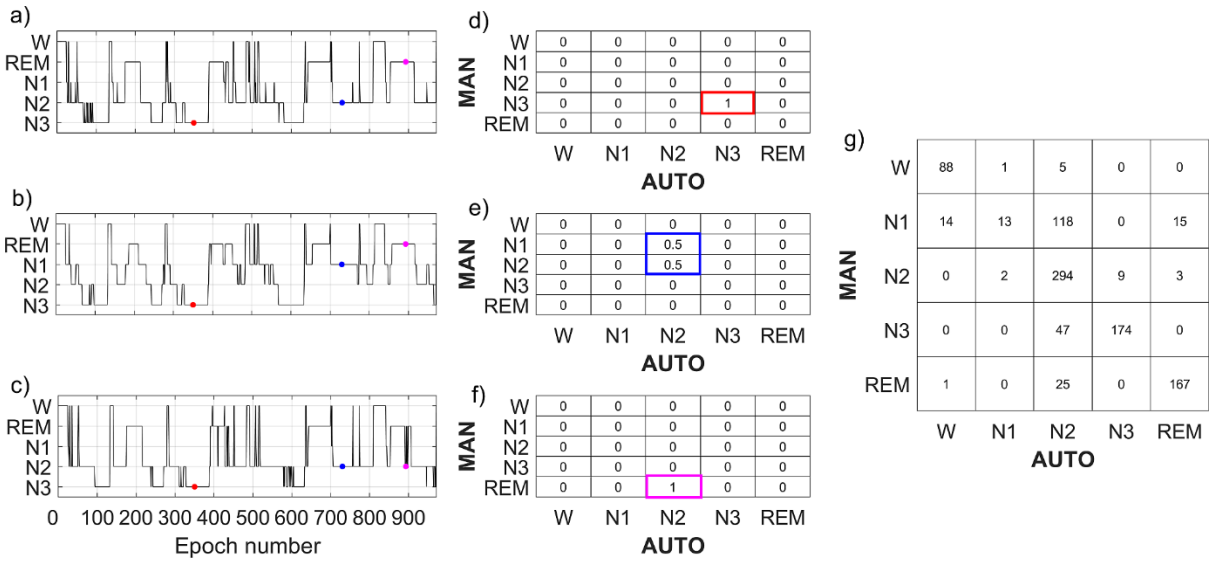
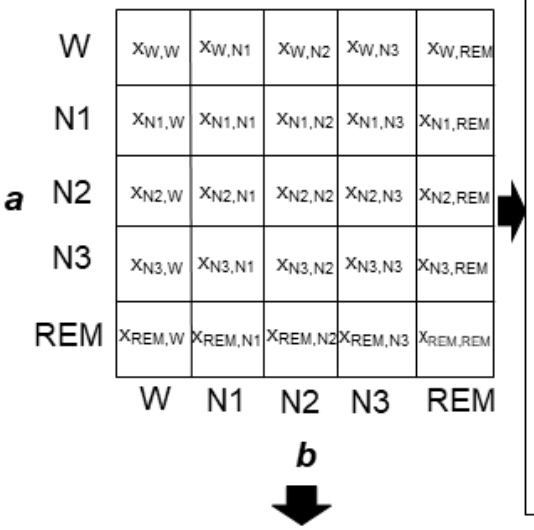


Figure S2: Manual, automatic hypnograms and MAN-vs-AUTO confusion matrix. On the left, the hypnograms scored manually by experts in Innsbruck (a), Berlin (b) and automatically scored (c) for the same PSG are shown. To understand how the confusion matrix comparing both manual (MAN) to automatic (AUTO) scoring was built, three examples are shown. For the epoch highlighted by the red dot, the two manual and the automatic scorings agreed to score it as N3 sleep, thus this epoch is reported in the confusion matrix as in (d). For the epoch highlighted with the blue dot, the scorer in Innsbruck and the automatic scoring scored it as N2 sleep, while the scorer in Berlin as N1 sleep, thus the epoch is reported in the confusion matrix as in (e). Finally, for the epoch highlighted in magenta, both human scorers scored it as REM sleep, while the algorithm as N2 sleep. This epoch is reported in the confusion matrix as in (f). The final confusion matrix including all the epochs is reported in (g).

Abbreviations: W: wakefulness; REM: rapid eye movement sleep; N1: non-REM stage 1 sleep; N2: non-REM stage 2 sleep; N3: non-REM stage 3 sleep.



IRR measures from the CM

$$A = \frac{\sum_i x_{i,i}}{N}$$

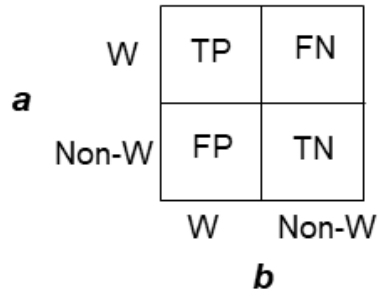
with $i \in \{W, N1, N2, N3, REM\}$ and N being the sum of all elements of the CM

$$\kappa = \frac{A - p_e}{1 - p_e}$$

with $p_e = \frac{1}{N^2}\sum_i n_{i,a} n_{i,b}$ with $i \in \{W, N1, N2, N3, REM\}$, *A* and *N* defined as above, $n_{i,a}$ the number of times **a** scores sleep stage *i*, and $n_{i,b}$ the number of times **b** scores sleep stage *i*.

Sleep stage specific IRR measures

From each CM, five stage specific 2-by-2 CMs are obtained. Below the example for W (***similar equations apply to other sleep stages***):



$$TP = x_{W,W}$$

$$TN = \sum_{i,j \in \{N1,N2,N3,REM\}} x_{i,j}$$

$$FP = \sum_{i \in \{N1,N2,N3,REM\}} x_{i,W}$$

$$FN = \sum_{i \in \{N1,N2,N3,REM\}} x_{W,i}$$

From the 2-by-2 CM the following stage-specific measures are calculated:

$$A_W = \frac{TP+TN}{N} \text{ with } N = TP + TN + FP + FN$$

$$\kappa_W = \frac{A_w - p_W}{1 - p_W} \text{ with } p_W = \frac{1}{N^2}[(TP + FN) \cdot (TP + FP) + (FP + TN) \cdot (FN + TN)] \text{ and N defined as above}$$

$$F1_W = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

Overall F1 score

$$F1 = \frac{F1_W + F1_{N1} + F1_{N2} + F1_{N3} + F1_{REM}}{5}$$

Figure S3: Overview of the performance measures to evaluate inter-rater reliability. From the confusion matrix (CM) the overall accuracy (A) and Cohen's kappa (κ) are calculated. The inter-rater reliability (IRR) measures for single stages are calculated from 2-by-2 CMs (which are obtained from the 5-by-5 CM). The figure reports the example for accuracy, Cohen's kappa and F1 score calculated for wakefulness (W), but similar equations apply for the other sleep stages. Finally, the overall F1 score is calculated as the mean of the sleep stage-specific F1-scores.

Abbreviations: W: wakefulness; REM: rapid eye movement sleep; N1: non-REM stage 1 sleep; N2: non-REM stage 2 sleep; N3: non-REM stage 3 sleep.

Table S1: Overall and stage-specific values of accuracies (A) for the different comparisons. The values are shown as mean ± one standard deviation and as (5^{th}-95^{th}) percentile across the participants. *Abbreviations*: INN vs BER: comparison of manual hypnograms scored in Innsbruck and Berlin; INN vs AUTO: comparison of manual hypnograms scored in Innsbruck to the automatic ones; BER vs AUTO: comparison of manual hypnograms scored in Berlin to the automatic ones; CONS vs AUTO: comparison of the manual consensus hypnograms (obtained by including only epochs where the two manual scorings were in agreement) to the automatic ones; MAN vs AUTO: comparison of both manual hypnograms to the automatic one (in case of disagreement between manual scorers, an epoch was equally weighted between the two stages).

Parameter	Measure	Comparison				
		INN-vs-BER	**INN-vs-AUTO**	**BER-vs-AUTO**	**CONS-vs-AUTO**	**MAN-vs-AUTO**
A	μ±σ	0.75±0.10	0.78±0.10	0.67±0.12	0.82±0.10	0.72±0.01
	m [5^{th} – 95^{th}]	0.76 [0.56-0.88]	0.80 [0.59-0.90]	0.69 [0.45-0.84]	0.84 [0.62-0.94]	0.74 [0.54-0.93]
A_W	μ±σ	0.94±0.05	0.92±0.08	0.90±0.09	0.94±0.08	0.91±0.09
	m [5^{th} – 95^{th}]	0.96 [0.86-0.99]	0.94 [0.77-0.98]	0.93 [0.72-0.98]	0.96 [0.80-0.99]	0.94 [0.76-0.98]
A_{N1}	μ±σ	0.83±0.09	0.89±0.06	0.81±0.10	0.91±0.07	0.85±0.07
	m [5^{th} – 95^{th}]	0.85 [0.66-0.94]	0.91 [0.78-0.96]	0.83 [0.62-0.93]	0.93 [0.77-0.98]	0.87 [0.71-0.94]
A_{N2}	μ±σ	0.83±0.07	0.86±0.07	0.77±0.09	0.88±0.07	0.81±0.07
	m [5^{th} – 95^{th}]	0.84 [0.69-0.93]	0.87 [0.73-0.94]	0.78 [0.60-0.90]	0.89 [0.75-0.97]	0.82 [0.69-0.91]
A_{N3}	μ±σ	0.94±0.04	0.94±0.04	0.90±0.06	0.94±0.05	0.92±0.04
	m [5^{th} – 95^{th}]	0.95 [0.85-0.99]	0.95 [0.86-1.00]	0.91 [0.80-0.98]	0.95 [0.83-1.00]	0.93 [0.84-0.98]
A_{REM}	μ±σ	0.95±0.04	0.96±0.04	0.96±0.03	0.98±0.03	0.96±0.03
	m [5^{th} – 95^{th}]	0.96 [0.87-0.99]	0.97 [0.88-0.99]	0.96 [0.90-0.99]	0.99 [0.93-1.00]	0.96 [0.90-0.99]

Table S2: Overall and stage-specific values of F1-score (F1) for the different comparisons. The values are shown as mean ± one standard deviation and as (5^{th}-95^{th}) percentile across the participants. *Abbreviations*: INN vs BER: comparison of manual hypnograms scored in Innsbruck and Berlin; INN vs AUTO: comparison of manual hypnograms scored in Innsbruck to the automatic ones; BER vs AUTO: comparison of manual hypnograms scored in Berlin to the automatic ones; CONS vs AUTO: comparison of the manual consensus hypnograms (obtained by including only epochs where the two manual scorings were in agreement) to the automatic ones; MAN vs AUTO: comparison of both manual hypnograms to the automatic one (in case of disagreement between manual scorers, an epoch was equally weighted between the two stages).

Parameter	Measure	Comparison				
		INN-vs-BER	**INN-vs-AUTO**	**BER-vs-AUTO**	**CONS-vs-AUTO**	**MAN-vs-AUTO**
F1	μ±σ	0.70±0.10	0.65±0.13	0.59±0.12	0.70±0.13	0.62±0.12
	m [5^{th} – 95^{th}]	0.71 [0.50-0.85]	0.66 [0.42-0.82]	0.59 [0.37-0.77]	0.71 [0.46-0.87]	0.63 [0.40-0.78]
$F1_W$	μ±σ	0.82±0.13	0.79±0.14	0.76±0.16	0.85±0.14	0.77±0.14
	m [5^{th} – 95^{th}]	0.85 [0.56-0.96]	0.83 [0.50-0.95]	0.80 [0.43-0.94]	0.89 [0.55-0.98]	0.81 [0.50-0.94]
$F1_{N1}$	μ±σ	0.49±0.15	0.34±0.16	0.28±0.15	0.42±0.19	0.30±0.14
	m [5^{th} – 95^{th}]	0.49 [0.24-0.72]	0.35 [0.07-0.61]	0.27 [0.05-0.53]	0.43 [0.08-0.73]	0.30 [0.07-0.54]
$F1_{N2}$	μ±σ	0.76±0.13	0.84±0.10	0.71±0.14	0.84±0.11	0.78±0.10
	m [5^{th} – 95^{th}]	0.78 [0.50-0.91]	0.86 [0.68-0.93]	0.74 [0.45-0.89]	0.87 [0.64-0.96]	0.79 [0.61-0.90]
$F1_{N3}$	μ±σ	0.68±0.29	0.48±0.35	0.43±0.33	0.50±0.37	0.45±0.33
	m [5^{th} – 95^{th}]	0.80 [0.00-0.94]	0.56 [0.00-0.92]	0.47 [0.00-0.88]	0.60 [0.00-0.95]	0.50 [0.00-0.89]
$F1_{REM}$	μ±σ	0.77±0.22	0.80±0.21	0.76±0.23	0.86±0.23	0.78±0.20
	m [5^{th} – 95^{th}]	0.84 [0.25-0.96]	0.88 [0.31-0.97]	0.84 [0.12-0.96]	0.94 [0.18-0.99]	0.85 [0.32-0.95]

Table S3: Results of the multiple regression linear analyses for overall accuracies. For each analysis, the overall accuracy was the outcome variable and age, sex (categorical), PLMS index, AHI and BMI the predictors. Z-score transformations were applied to both outcome variable and predictors (except sex). For each model, the overall p-value is reported, as well as the slope estimate (b) and the p-value of each predictor. Cubic transformation was applied to accuracies in the highlighted comparisons (*) to meet the normality assumption of the model residuals. *Abbreviations*: INN-vs-BER: comparison of manual hypnograms scored in Innsbruck and Berlin; INN-vs-AUTO: comparison of manual hypnograms scored in Innsbruck to the automatic ones; BER-vs-AUTO: comparison of manual hypnograms scored in Berlin to the automatic ones; CONS-vs-AUTO: comparison of the epochs where manual scorers from Innsbruck and Berlin were in consensus to the respective epochs automatically scored; MAN-vs-AUTO: comparison of both manual hypnograms to the automatic one (in case of disagreement between manual scorers, an epoch was equally weighted between the two manually scored stages). PLMS: periodic limb movement during sleep; AHI: apnea-hypopnea index; BMI: body-mass index.

	INN-vs-BER		**INN-vs-AUTO(*)**		**BER-vs-AUTO**		**CONS-vs-AUTO(*)**		**MAN-vs-AUTO(*)**	
Predictors	b	p-value	b	p-value	b	p-value	b	p-value	b	p-value
Intercept	-0.122	**0.003**	-0.107	**0.008**	-0.180	**<0.001**	-0.169	**<0.001**	-0.178	**<0.001**
Age	-0.118	**<0.001**	-0.034	0.282	-0.143	**<0.001**	-0.088	**0.005**	-0.115	**<0.001**
Sex (F)	0.259	**<0.001**	0.226	**<0.001**	0.382	**<0.001**	0.361	**<0.001**	0.378	**<0.001**
PLMS index	-0.014	0.654	-0.035	0.250	-0.021	0.494	-0.046	0.125	-0.029	0.329
AHI	-0.186	**<0.001**	-0.246	**<0.001**	-0.155	**<0.001**	-0.186	**<0.001**	-0.201	**<0.001**
BMI	-0.043	0.173	-0.076	**0.015**	-0.066	**0.035**	-0.074	**0.018**	-0.077	**0.012**
Overall p-value	**<0.001**		**<0.001**		**<0.001**		**<0.001**		**<0.001**	

Table S4: Results of the multiple regression linear analyses for overall F1-scores For each analysis, the overall F1-score was the outcome variable and age, sex (categorical), PLMS index, AHI and BMI the predictors. Z-score transformations were applied to both outcome variable and predictors (except sex). For each model, the overall p-value is reported, as well as the slope estimate (b) and the p-value of each predictor. Cubic transformation was applied to F1-scores in the highlighted comparisons (*) to meet the normality assumption of the model residuals. *Abbreviations*: INN-vs-BER: comparison of manual hypnograms scored in Innsbruck and Berlin; INN-vs-AUTO: comparison of manual hypnograms scored in Innsbruck to the automatic ones; BER-vs-AUTO: comparison of manual hypnograms scored in Berlin to the automatic ones; CONS-vs-AUTO: comparison of the epochs where manual scorers from Innsbruck and Berlin were in consensus to the respective epochs automatically scored; MAN-vs-AUTO: comparison of both manual hypnograms to the automatic one (in case of disagreement between manual scorers, an epoch was equally weighted between the two manually scored stages). PLMS: periodic limb movement during sleep; AHI: apnea-hypopnea index; BMI: body-mass index.

	INN-vs-BER		**INN-vs-AUTO(*)**		**BER-vs-AUTO**		**CONS-vs-AUTO(*)**		**MAN-vs-AUTO(*)**	
Predictors	b	p-value	b	p-value	b	p-value	b	p-value	b	p-value
Intercept	-0.108	**0.006**	-0.240	**<0.001**	-0.220	**<0.001**	-0.274	**<0.001**	-0.259	**<0.001**
Age	-0.170	**<0.001**	-0.202	**<0.001**	-0.208	**<0.001**	-0.230	**<0.001**	-0.224	**<0.001**
Sex (F)	0.227	**<0.001**	0.509	**<0.001**	0.466	**<0.001**	0.583	**<0.001**	0.550	**<0.001**
PLMS index	-0.027	0.359	-0.062	**0.030**	-0.051	0.076	-0.077	**0.006**	-0.060	**0.032**
AHI	-0.237	**<0.001**	-0.191	**<0.001**	-0.177	**<0.001**	-0.165	**<0.001**	-0.178	**<0.001**
BMI	-0.011	0.723	-0.059	**0.044**	-0.039	0.192	-0.050	0.081	-0.054	0.064
Overall p-value	**<0.001**		**<0.001**		**<0.001**		**<0.001**		**<0.001**	