SUPPLEMENTARY DATA

Sleep spindle detection: crowdsourcing and evaluating performance of experts, non-experts, and automated methods

Simon C. Warby, Sabrina L. Wendt, Peter Welinder, Emil G.S. Munk, Oscar Carrillo, Helge B.D. Sorensen, Poul Jennum, Paul E. Peppard, Pietro Perona, Emmanuel Mignot.

Nature Methods: doi:10.1038/nmeth.2855

Supplementary Note:

In an attempt to further increase the performance of automated detection, we tried all possible combinations of detectors (all single detectors, all combinations of two detectors, all combinations of three detectors,.etc) with all possible levels of consensus (only one detector in the combination needs to find a spindle, two detectors must agree, etc). The maximum performance obtained with any combination was not greater than using the group consensus rule.

Further, to determine whether using a minimum spindle duration of 0.3 seconds rather than 0.5 seconds had impaired the performance of the automated detectors, we re-tested them with a 0.5 second spindle minimum duration criteria, both for the detector output and for spindles in the gold standard. The precision and recall of each detector changed slightly, but did not improve performance overall. Notably, performance of detector a5 decreased substantially using a minimum spindle duration of 0.5 seconds (F_1-score = 0.36) due to its tendency to underestimate spindle duration.

We also wanted to determine whether we had impaired the performance by using a $T_{overlap}$ of 0.2. This is not the case, as all detectors and annotators seemed to achieve maximal performance up to $T_{overlap}$ = 0.4. Beyond this threshold, automated detector and non-expert performance started to decay. The performance of the individual experts did not start to decay until $T_{overlap}$ =0.6, suggesting that the expert annotators were better able to estimate the location of the spindle in the EEG data. Although the overlap of spindle detections to events will have no effect on the estimation of spindle density, it will be an important consideration for the estimation of sleep spindle characteristics such as duration and frequency content. In this regard, when testing and optimizing automated spindle detectors, it would be preferable to optimize using a by-event analysis for detectors aiming to estimate spindle density, while a by-sample analysis would be preferable for detectors being used to understand specific spindle characteristics.

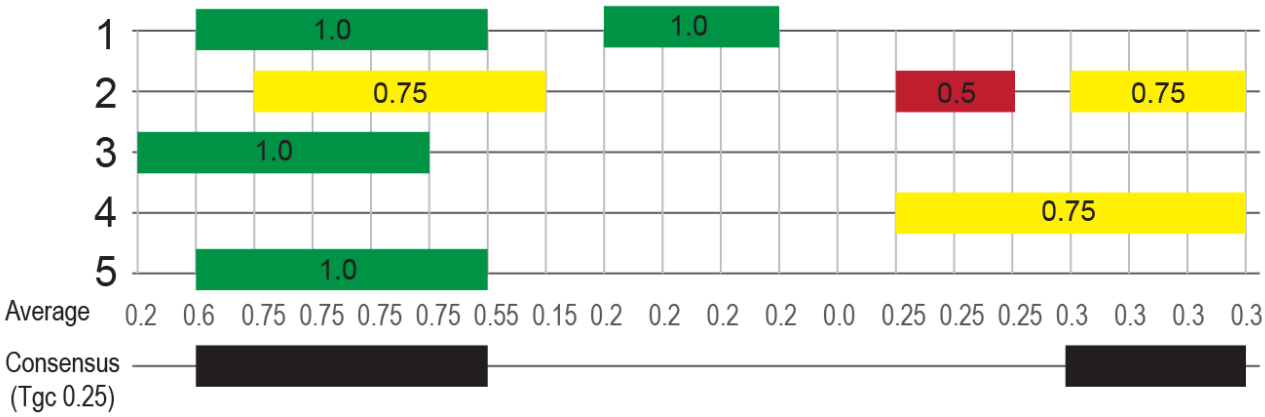
To determine whether the automated detectors and humans have systematic differences in spindle detection, we assessed the spindles detected by the 6 detectors against the gold standard on four primary spindle characteristics: absolute sigma power, spindle duration, oscillation frequency and amplitude (**Supplementary Fig. 12**). From this data, it is clear that each detector has a slightly different pattern underlying the differences.

Relative to the gold standard, each detector had a different bias towards detecting spindles with higher or lower sigma power. Many of the detectors systematically under- or over-estimate spindle duration (which is consistent with the data we present in **Figure 5c**). In general, automatically detected spindles tended to fall in a more restricted oscillation frequency range relative to the gold standard. Three detectors (a4, a5, a6) found spindles with a similar amplitude distribution to the gold standard, while other detectors had a bias for finding spindles with higher (a2) or lower amplitudes (a1, a3) than the gold standard. In summary, each detector seems to have specific differences relative to the gold standard that are difficult to generalize.

Finally, we wanted to determine whether the duration estimates of the automated spindle detectors were well-correlated with the measured duration in the gold standard. If this were the case, then the detectors would provide an adequate relative value, which could be corrected with a correction factor (i.e. multiply by 2 if it is consistently off by 50%). The maximum correlation was low (a2, R^2 = 0.213) suggesting that the duration estimates are not reliable, and not easily corrected by a standard correction factor (**Supplementary Fig. 13**).

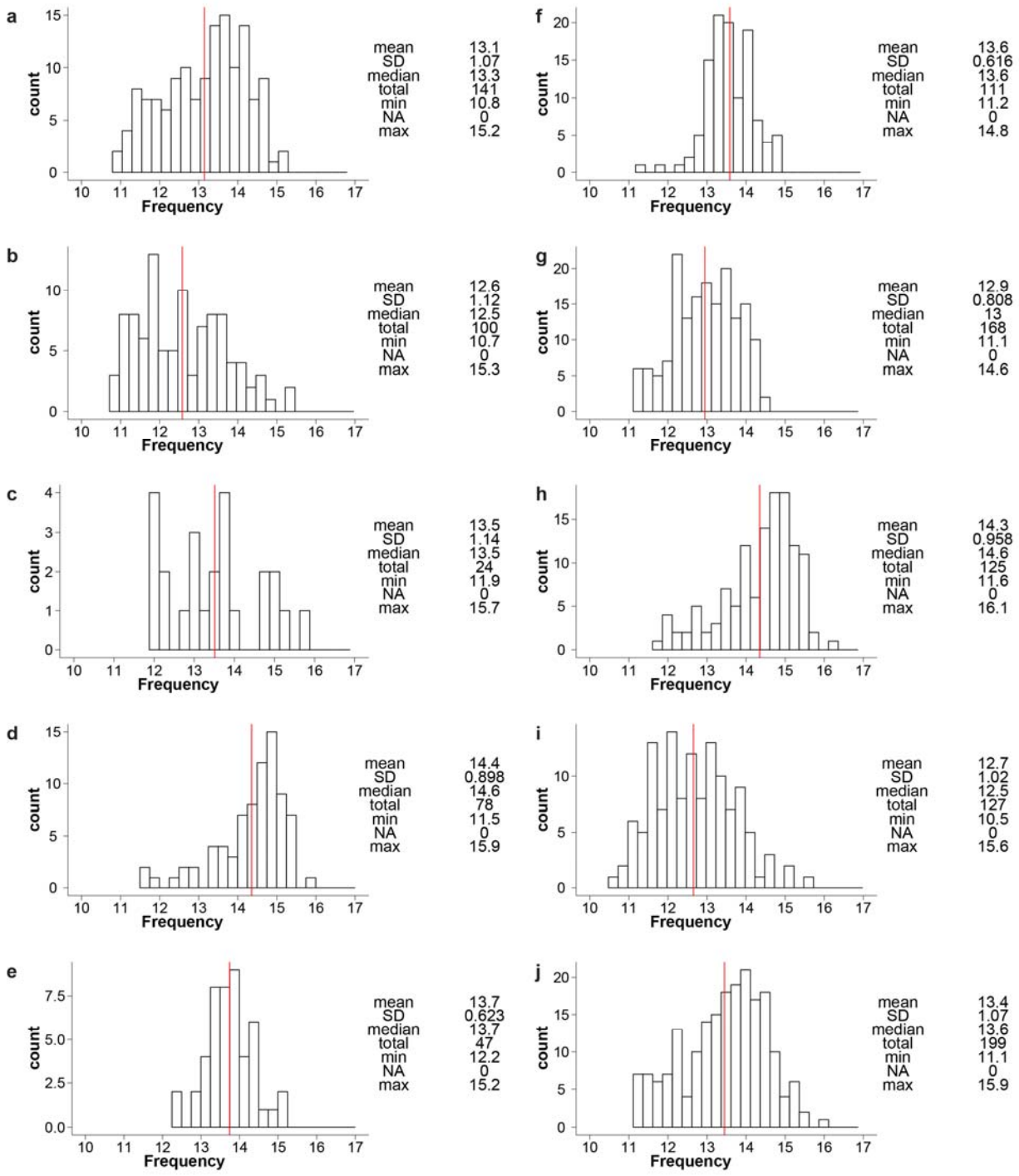
Nature Methods: doi:10.1038/nmeth.2855

Supplementary Figure 1: Group consensus rule. Example demonstrating how the annotations of 5 annotators (colored boxes) with weighted confidence scores (green = 'Definitely' =1.0, yellow = 'Probably' = 0.75, red = 'Maybe'/'Guessing' = 0.5, no spindle = 0) are averaged at each sample point and aggregated into a group consensus. Each data sample point is included in the consensus (black bars) if the average confidence score exceeds the threshold for group consensus (*Tgc*). In epochs that have been viewed by at least 5 experts, a T_{egc} of 0.25 requires that at least two experts identify the spindle with confidence equal or greater than 'Probably' in order for it to be included in the group consensus.



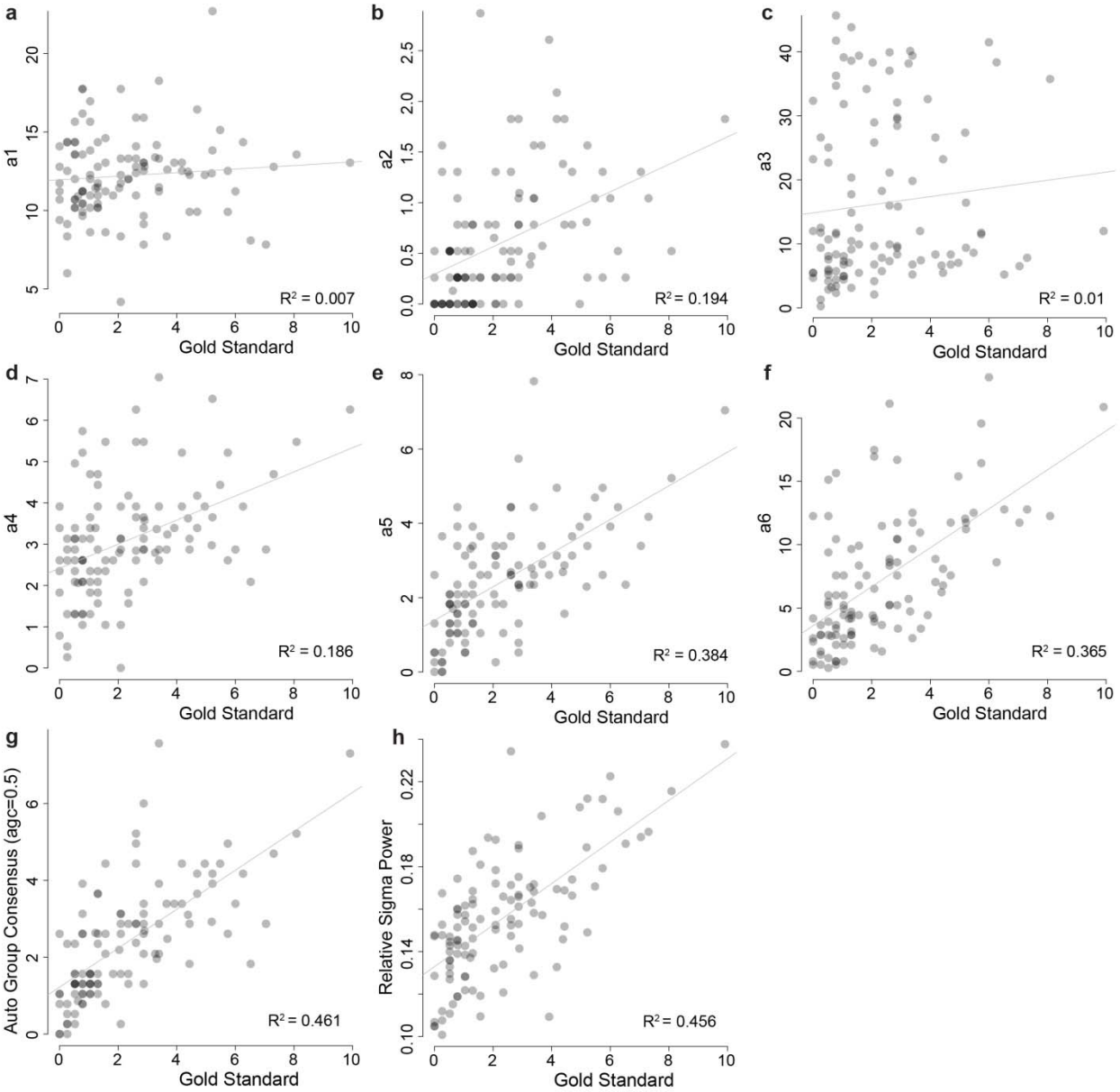
Nature Methods: doi:10.1038/nmeth.2855

Supplementary Figure 2: Spindle oscillation frequency distribution in 10 subjects (a-j). The number of spindles of each oscillation frequency bin is indicated in the histogram. The total number of spindles for each individual, as well as the mean, median, minimum and maximum oscillation frequency is reported in the table. Shapiro-Wilks normality test p-value of the distribution is also reported. These 10 subjects each had 100 epochs of N2 sleep at C3-M2 annotated for sleep spindles in the dataset. There is not clear evidence for bimodal distributions that would suggest discrete populations of fast and slow spindles in these individuals at this one scalp location. Rather, each individual has a unique distribution of spindle frequencies.



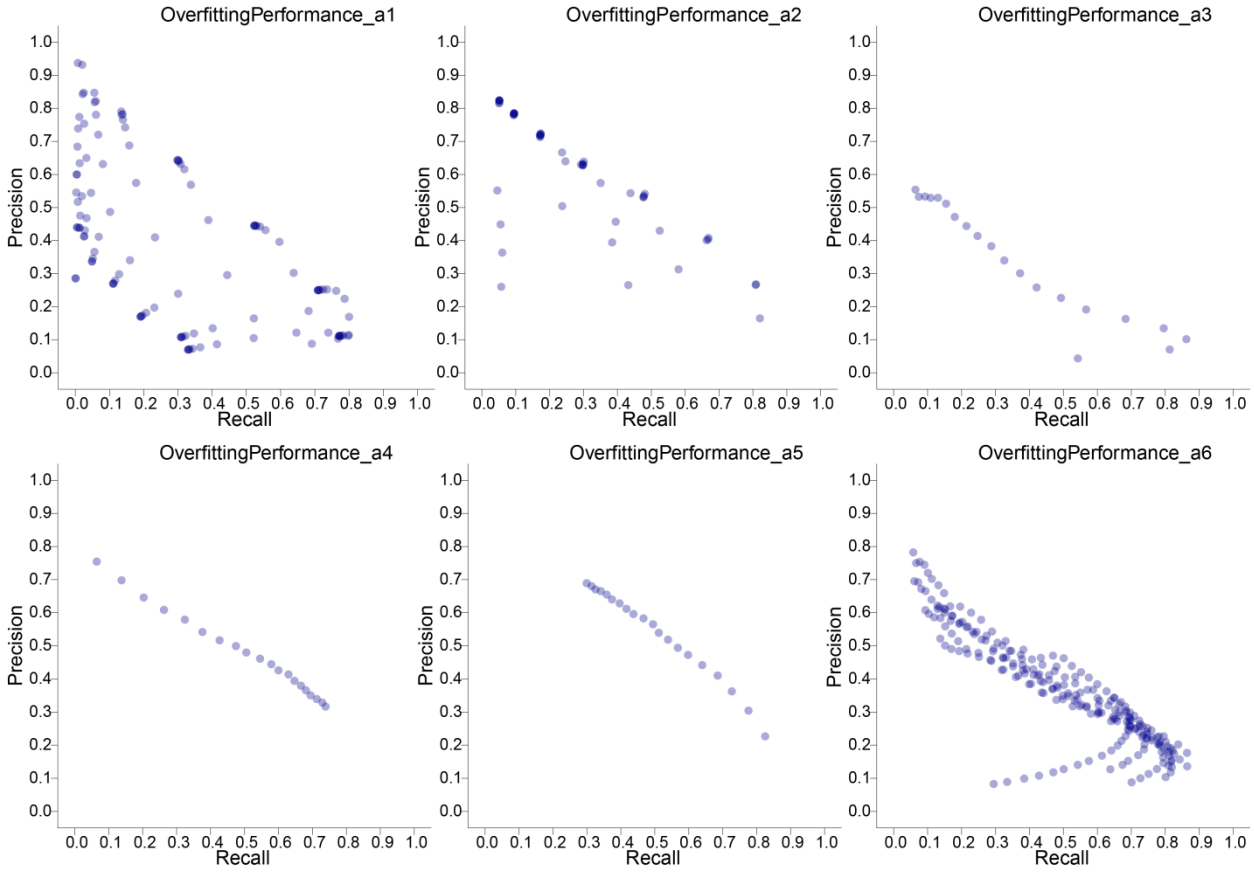
Nature Methods: doi:10.1038/nmeth.2855

Supplementary Figure 3: Correlation of spindle density in the gold standard versus automated detectors or relative sigma power. Linear regression line and coefficient of determination (R^2) is shown for each automated detector (a-f; a1-a6), the auto group consensus (g; T_{agc} = 0.5), and relative sigma power (h). Each dot in the plots is one subject. Comparison is made against the gold standard (expert group consensus with T_{egc} = 0.25). Based on this by-subject analysis, relative sigma power estimates spindle activity better than any single automated spindle detector, but not better than the group consensus of the automated detectors combined.



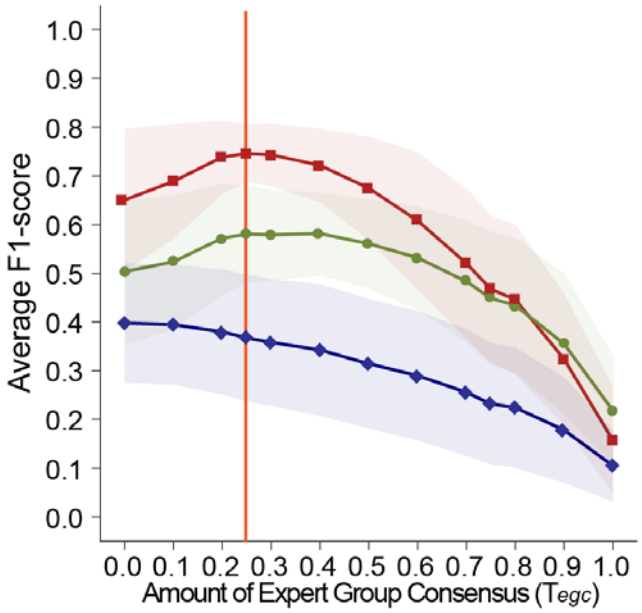
Nature Methods: doi:10.1038/nmeth.2855

Supplementary Figure 4: Over-fitting performance results of automated detectors. Precision-recall curves are plotted for each of the 6 automated detectors (a1-a6). Detector parameters were varied to attempt to optimize performance against the gold standard. The parameter(s) being varied, and the resulting maximal F_1-score are presented in **Supplementary Table 5.**



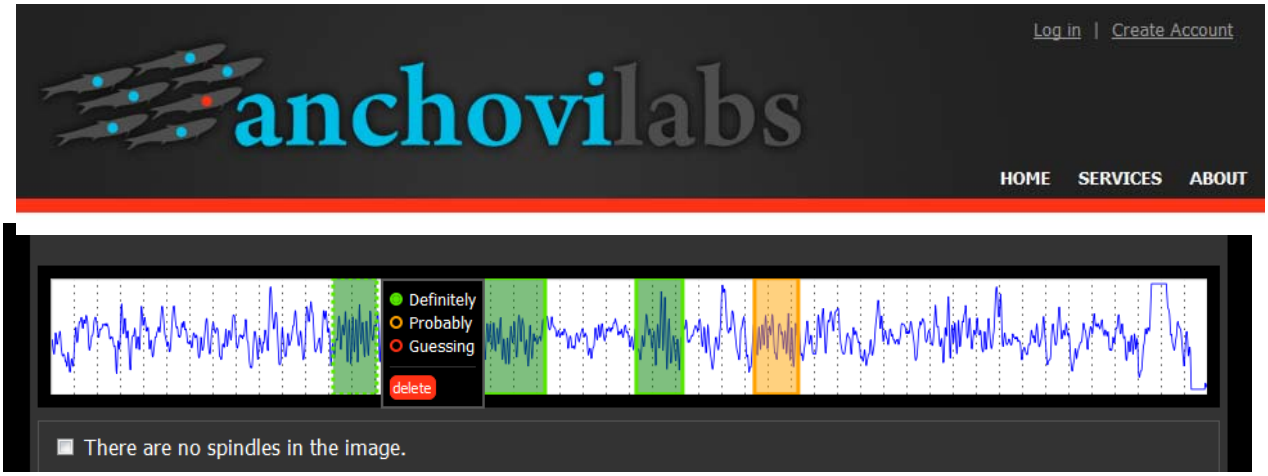
Nature Methods: doi:10.1038/nmeth.2855

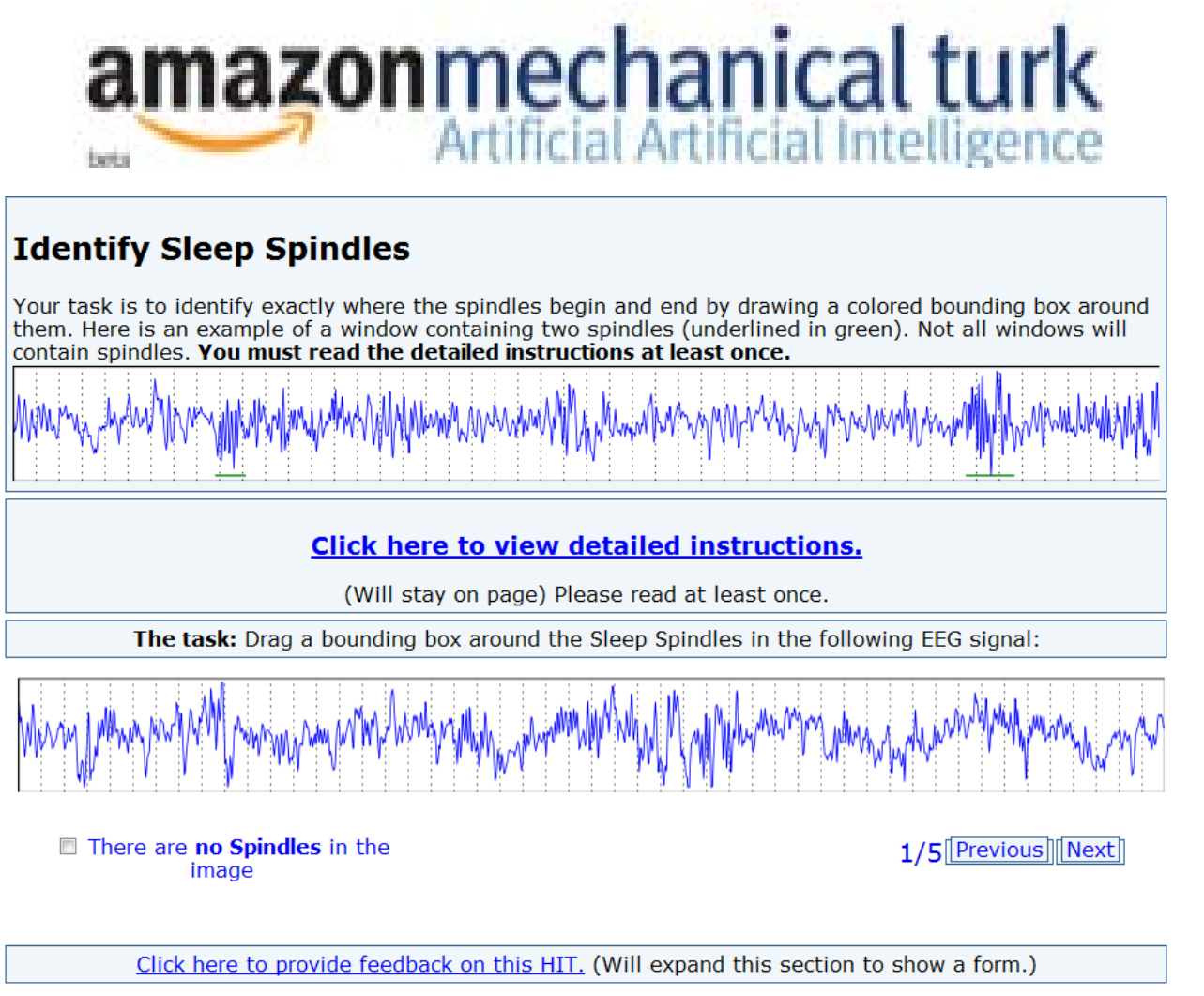
Supplementary Figure 5: Effect of varying the amount of expert group consensus (T_{egc}) required for the gold standard on the by-event performance. F_1-score for the average individual experts (squares), average non-expert group (ngc between 0.2 and 0.6; circles), and average individual automated spindle detectors (diamonds). Standard deviation of each group is indicated with shading. For this study, we used an expert group consensus of T_{egc} = 0.25 (orange line). Performance of the non-expert group or the automated detectors would not have increased overall if a higher level of expert group consensus (T_{egc}) was chosen for the gold standard.



Nature Methods: doi:10.1038/nmeth.2855

Supplementary Figure 6: Screenshots of the web interfaces for Anchovi Labs and Amazon Mechanical Turk used to collect spindle identification data from experts and non-experts.



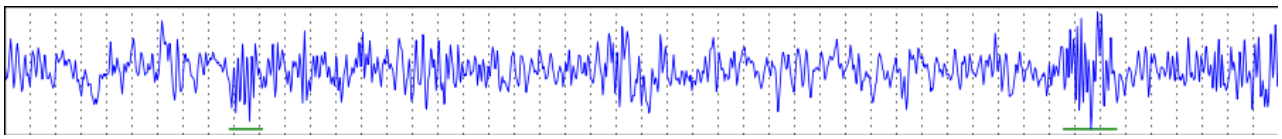


Nature Methods: doi:10.1038/nmeth.2855

Supplementary Figure 7: Written instructions and training protocol for experts and non-experts.

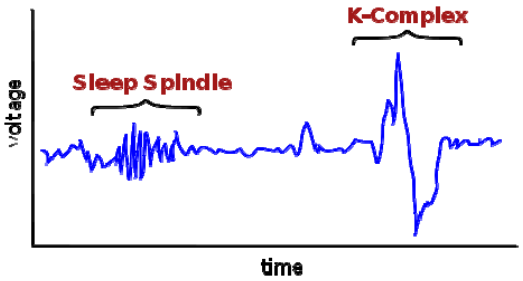
Identify Sleep Spindles

Your task is to identify exactly where the spindles begin and end by drawing a colored bounding box around them. Here is an example of a window containing two spindles (underlined in green). Not all windows will contain spindles. **You must read the detailed instructions at least once.**



Detailed Instructions

You will be presented with EEG data that measures the brain activity of a person that is in stage 2 sleep. The goal is to identify patterns in the data that are known as **Sleep Spindles.** Your task is to identify exactly where the spindles begin and end by drawing a colored bounding box around them. A description and examples of sleep spindles are presented below.



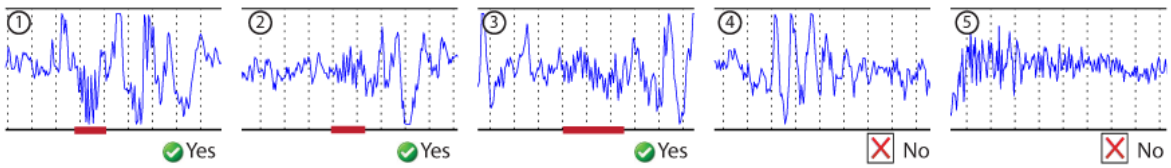
source: wikipedia

Figure 1: An example of a sleep spindle and a K-complex. These features are seen in the EEG during stage 2 sleep. Note that over time (moving to the right in the horizontal axis) a change in voltage of the EEG signal has caused the line to go up and down (vertical axis).

Definition of a Sleep Spindle:

For the purpose of this study, we are defining a sleep spindle based on its shape, speed, duration, and height. **It is most important that the spindle stands out as being different from the surrounding EEG signal.**

1. Shape of spindle:
The spindle is usually shaped like a diamond or football (this is sometimes referred to as a 'waxing/waning' shape). Note that sleep spindles are often found near K-complexes (see Figure 1). Sometimes the K-complex wave might be so close to the spindle that it changes the shape of the sleep spindle. A certain amount of deformation in the shape of sleep spindle (ie the axis of spindle is not completely flat) is ok (Figure 2).



Nature Methods: doi:10.1038/nmeth.2855

Figure 2: Shape of the spindle (underlined in red) is acceptable in the first three examples, but not the fourth or fifth. Note that the third spindle shape is changed slightly because of other waves.

2. Speed of waves:
A sleep spindle is a group of waves that oscillate (go up and down) at approximately 12-15 cycles per second (this can be said as having a frequency of 12-15Hz). It can be difficult to estimate the speed. However, because the vertical dashed lines in the display mark 0.5 second intervals, one way to determine the speed is to count the number of wave peaks between the dashed lines: between 6 and 7.5 wave peaks in 0.5 seconds would be equal to 12-15 cycles per second (Figure 3). It is important that the spindle appears as a 'burst' of waves that are slightly faster (closer together) than the waves around it (Figure 4).

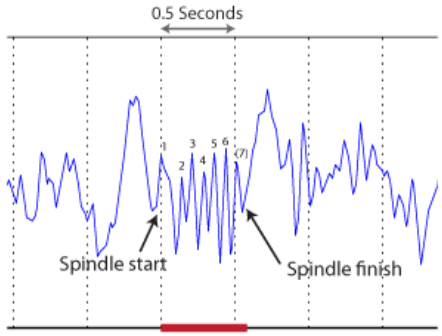


Figure 3: Estimating the speed of the spindle by counting the number of waves. In this enlarged picture, it is easy to see there are between 6 and 7 waves in 0.5 seconds, which is equal to 13 cycles per seconds. This is within the 12-15 cycle per second range of sleep spindles.

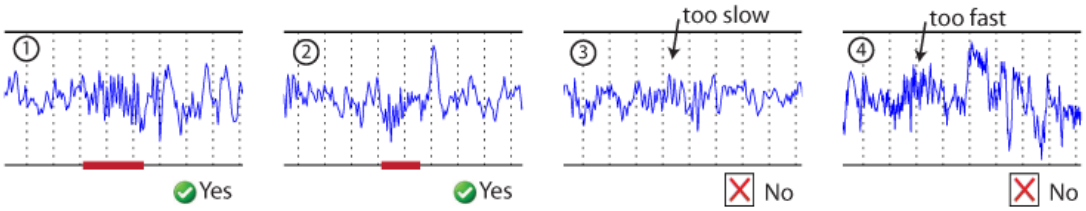


Figure 4: The first two examples are appropriate speeds for a sleep spindle. The third example is too slow, and the fourth example too fast (too many cycles per second) to be a sleep spindle. Notice that you can clearly see gaps between the waves in the third example, and you can see no gaps at all between the waves in the fourth example.

3. Duration of spindle:
Most commonly, spindles are around 0.5 to 1.0 seconds in length (duration), but can be as short as 0.4 seconds and as long as 5 seconds (Figure 5).

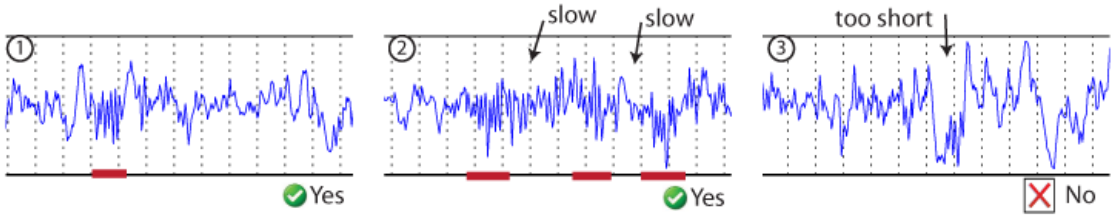


Figure 5: Example spindles of different durations. The second example could be considered one single long spindle, but there are slow segments in the middle, and in this case, it has been considered three

Nature Methods: doi:10.1038/nmeth.2855

separate spindles of shorter duration. The duration of the third example is too short to be considered a spindle (< 0.4 seconds).

3. Height of waves in spindle:
The height (amplitude) of the spindle is less important than the other criteria. The height of the waves in the spindle is usually a little larger than the waves around it. The spindle should be distinct from the other waves around it.

How to annotate Spindles:

To create a bounding box around the spindles, you need to **left click and drag** with the mouse around the spindle. In this case, a menu appears where you can select how sure you are of it:

- **Definitely**: "I am sure that this is a Spindle. It meets all of the criteria of shape, speed, duration and height and is very distinct from the surrounding waves."
- **Probably**:"I would bet that this is a Spindle, although I am not completely sure because one of the criteria is not quite right. There are some imperfections in the spindle, but I still think it is a sleep spindle."
- **Guessing**: "I think this could be a spindle, but I am not positive. Two or more of the criteria are not perfect. It would be best to have someone have a second look at this."

If you just click on the "Spindle" button, it will be assigned "Definitely".

Accuracy is important, so be sure to size the bounding box so that it only includes the spindles, not surrounding EEG waves. You can resize and move the bounding box (Figure 6) by clicking in the middle or on the edges and dragging. You can change the spindle certainty, or delete the bounding box by right clicking on it. There may be multiple spindles, or none within a window (Figure 7). If there are no spindles in the window, indicate this by clicking the box marked "There are no Spindles in the image" found at the bottom left of your screen before moving on to the next window.

If the spindle runs into the end or beginning of the window, just draw the bounding box right up to the edge of the window.

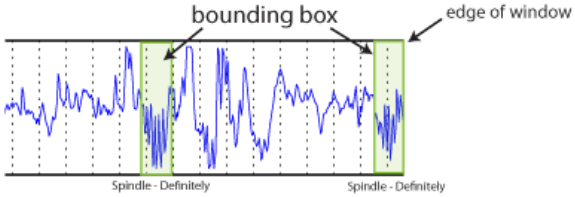
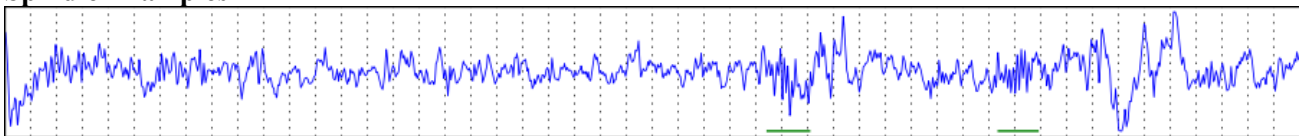
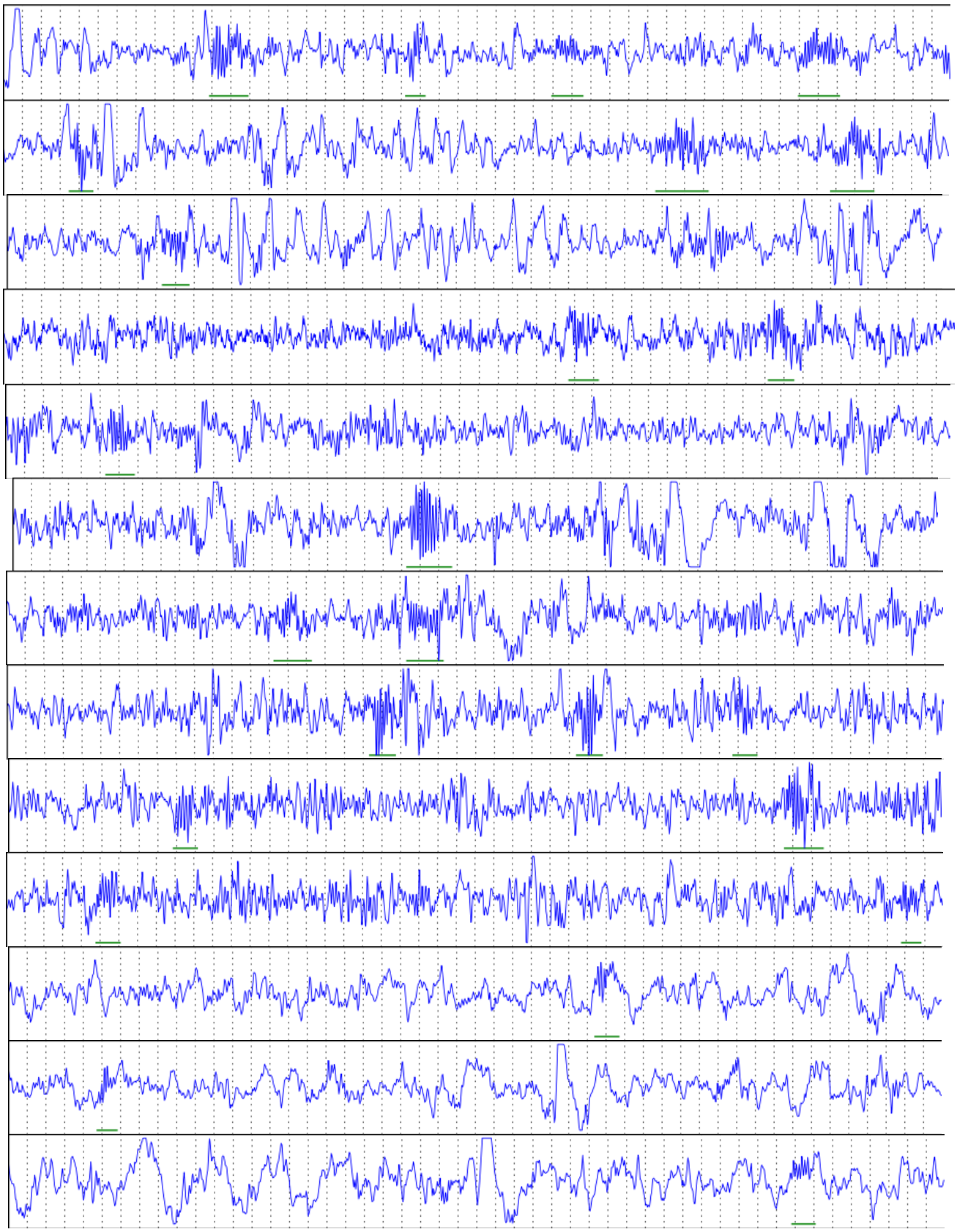


Figure 6: Use your mouse to draw a bounding box around the spindle. The size of the bounding box can be changed by clicking on the middle or on the edges of the box and dragging. In this case, the certainty of the spindle has been judged as "Definitely". *If the spindle runs into the end or beginning of the window, just draw the bounding box right up to the edge of the window.

Spindle Examples



Nature Methods: doi:10.1038/nmeth.2855



Nature Methods: doi:10.1038/nmeth.2855

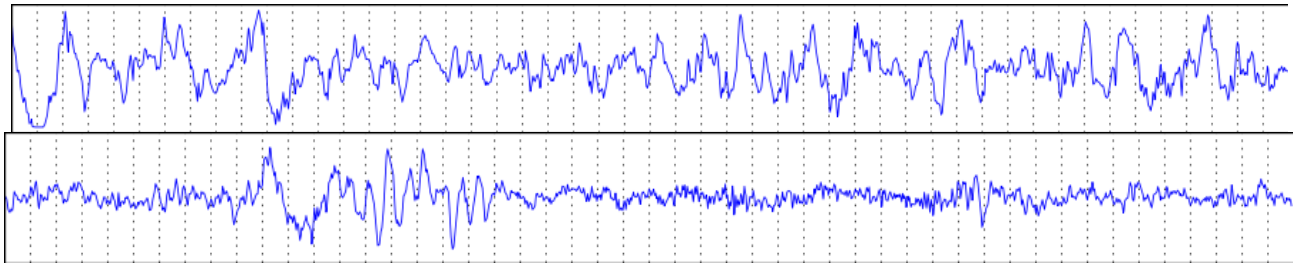


Figure 7: Here are some examples of sleep spindles, indicated with a green bar below them. These are the events you will want to identify by drawing a bounding box around them with your mouse. There are also other events that are not identified as spindles, because they are too short, too small, or don't have the correct shape. As a reference, the gridlines in the display are spaced at 0.5 seconds so that you can approximate the number of cycles by counting the number of waves. Some windows do not have any spindles.

Remember, you need to select the spindles precisely. Do not include any noise around them. **You will only get paid if you do a careful job in selecting the spindles.** We will double check some of your jobs to make sure you select events that fulfill the criteria.

The goal is to very accurately identify the spindles. Try as best you can to identify where the spindle begins and ends. Quality is more important that quantity.

Nature Methods: doi:10.1038/nmeth.2855

Supplementary Figure 8: Pseudo-code for the spindle detection algorithms.

a1. Bodizs et al, 2009 [41]

[# Derive spindle frequency boundaries and spindle detection amplitude criteria for slow and fast spindles using the all night average spectrum during S2+S3+S4 sleep. Band-pass filter EEG, calculate Hanning-corrected moving average and detect spindles when the constant threshold is exceeded. Do the spindle detection separately for slow and fast spindles and combine the results afterwards.]

```
for C3-M2 and O1-M2 do
       Calculate the average amplitude spectra of S2+S3+S4 sleep with 0.06 Hz resolution
       using 4 s non-overlapping Hanning-corrected windows zero-padded to 16.45 s and
       normalized as
       spectrum_normalized ← 2×spectrum/(16.45×fs)
       spectrum_lowres ← construct a low resolution spectrum with 0.24 Hz resolution by
       extracting every 4th sample of spectrum_normalized in the 9-16 Hz range
       for n = 2,…,N-1 do
              Fit ax^2+bx+c to spectrum_lowres(n) at points n-1:n+1
              spectrum_lowres"(n) ← 2a [# 2nd order derivative of spectrum_lowres]
       end for
end for

spectra_average"= mean(spectra_lowresC3", spectra_lowresO1")
Find the exact zero-crossing points surrounding the two largest negative peaks of
spectra_average" by linear interpolation between the sample points on either side
Round the zero-crossing points to nearest frequency bin in the high resolution spectra [#
lower and upper frequency boundaries of slow and fast spindles]

for slow and fast spindle detection do
       thresholds ← number of frequency bins between
       boundaries×mean(spectrum_normalized(bound_lower,bound_upper))

       x_m ← center frequency of spindle frequency band
       w ← width of spindle frequency band
       Gaussian filter ← e^(-|x-x_m| / w/2) , x = 0,…,fs/2

       for every 4 s non-overlapping window do
              Filter C3-M2 with the Gaussian filter and calculate the absolute value of
              the filtered signal
       end for

       Calculate the 22-points Hanning-weighted moving average multiplied by π/2

       if moving average > threshold and 0.3 s ≤ duration above threshold ≤ 3 s then
              Detect spindle
       end if
       Return detection_slow(n)or detection_fast(n)
end for
for n = 1:N do
       if detection_slow(n) + detection_fast(n) ≥ 1 then
              detection(n) ← 1
       else
              detection(n) ← 0
       end if
end for
for j = 1:no. of spindles do
       if duration_j > 3 s then
              Discard j^th spindle
       end if
end for
```


Nature Methods: doi:10.1038/nmeth.2855

a2. Ferrarelli et al, 2007 [17]

[# Band-pass filter EEG, calculate the envelope, determine upper and lower thresholds for spindle detection. When the signal exceeds the upper threshold determine the beginning and end of the spindle based on the nearest troughs below the lower threshold. Detect a spindle if it matches the duration criteria.]

```
C3_11-15Hz ← Bandpass filter signal from C3-M2 in the 11-15 Hz band
Determine the envelope of the rectified bandpass filtered signal by using the local
maxima of the rectified signal
Find the peaks and troughs of the envelope
Construct a histogram in 120 bins of the envelope peak amplitude [# only S2+S3+S4]
threshold_lower ← 2×most common envelope peak amplitude
threshold_upper ← 8×mean(|C3_11-15Hz|) [# only S2+S3+S4]
Define possible spindle boundaries as troughs of the envelope < threshold_lower
Define possible spindle peaks as envelope peaks > threshold_upper [# only S2+S3+S4]
for every spindle peak do
       Find the boundaries preceding and following the peak
       if 0.3 s ≤ duration ≤ 3 s then
              Detect spindle within these boundaries
       end if
end for
```

a3. Mölle et al, 2002 [31]

[# Band-pass filter EEG, calculate RMS in sliding windows and apply a constant threshold. Detect a spindle if the RMS signal exceeds the threshold for 0.3-3 s.]

```
Bandpass filter signal from C3-M2 in the 12-15 Hz band
Calculate the RMS of the bandpass filtered signal with a time resolution of 50 ms using a
time window of 100 ms [# 50% overlap]
threshold ← 1.5 × standard deviation of bandpass filtered signal [# only S2]
if RMS > threshold and 0.3 s ≤ duration above threshold ≤ 3 s then
       Detect spindle
end if
```

a4. Martin et al, 2012 [10]

[# Band-pass filter EEG, calculate RMS in sliding windows and apply a constant threshold. Detect a spindle if the RMS exceeds a constant threshold for 0.3-3 s.]

```
Bandpass filter signal from C3-M2 in the 11-15 Hz band
Calculate the RMS of the bandpass filtered signal with a time resolution of 25 ms using a
time window of 25 ms [# no overlap]
threshold ← 95^th percentile of RMS signal [# only S2+S3+S4]
if RMS > threshold and 0.3 s ≤ duration above threshold ≤ 3 s then
       Detect spindle
end if
```


Nature Methods: doi:10.1038/nmeth.2855

a5. Wamsley et al, 2012 [18]

[# Calculate the wavelet transform of C3-M2 and calculate the MA in sliding windows. Detect a spindle if the MA exceeds a constant threshold for 0.3-3 s.]

```
Calculate the continuous wavelet transform of C3-M2 using a complex morlet wavelet with
center frequency of 13.55 Hz
Extract the real part of the wavelet coefficients x_wave(t)
Calculate the moving average of x_wave(t), x_MA(t), using 0.1 s windows
threshold ← 4.5 × mean(x_MA(t)) [# only S2]
Find possible spindles when 0.3 s ≤ x_MA(t) above threshold ≤ 3 s
for i = 2:no of spindles do
       if time between spindle end_i-1 and spindle end_i < 1 s then
              Discard i^th spindle
       end if
end for
```

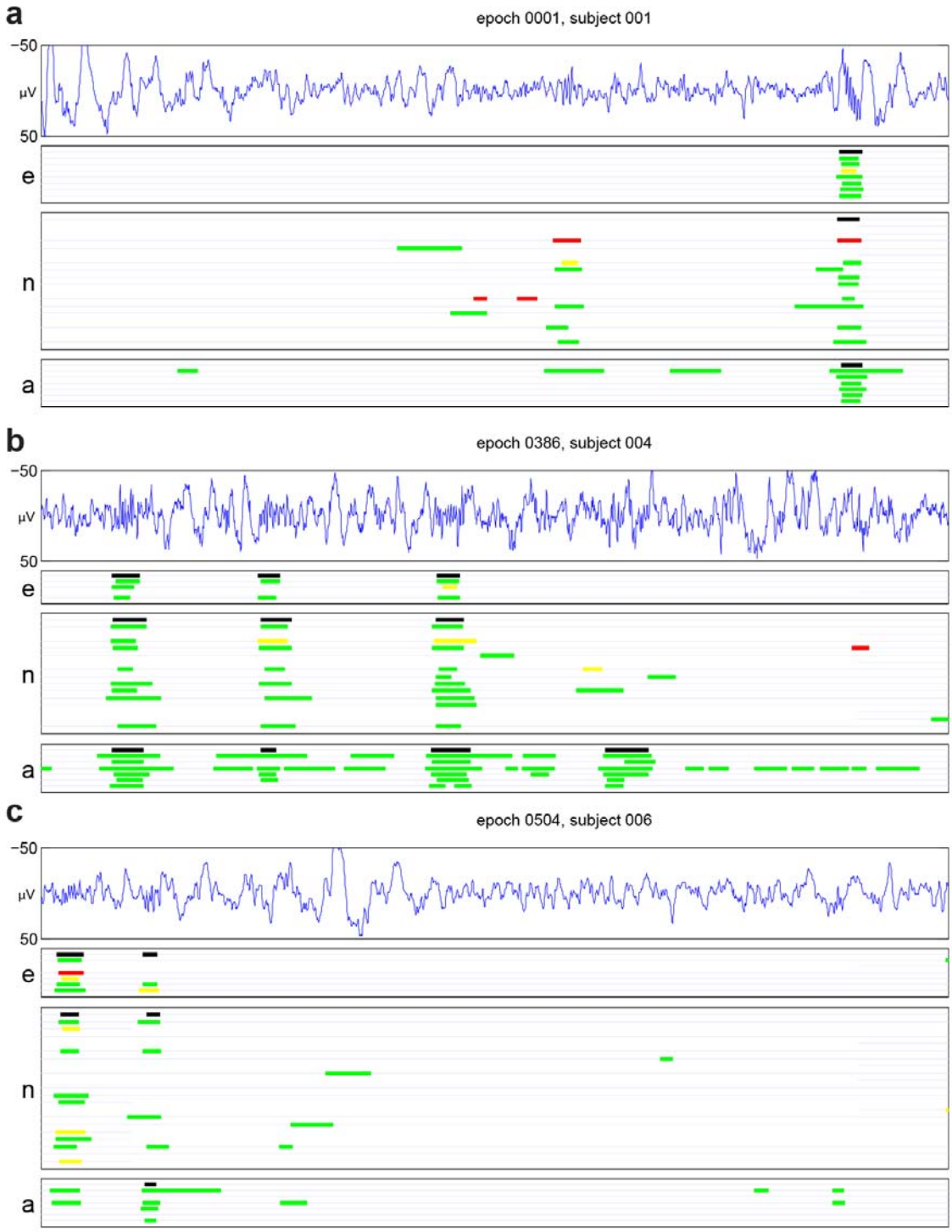
a6. Wendt et al, 2012 [45]

[# Band-pass filter EEG. The time varying threshold is determined as the envelope of the rectified signal with a given offset. Detect a spindle if the rectified filtered signal exceeds the time varying threshold and determine the beginning and end of the spindle based on the shape of the rectified filtered signal. Discard spindles if they are more likely to be alpha intrusions, artifacts, or do not meet the duration criteria. Detect spindles using a combination of two different sets of envelopes and offsets.]

```
Bandpass filter signals from C3-M2 and O1-M2 in the 11-16 Hz band
for i = 1 to 2 [# detector no.] do
       if i = 1 then
              f_passband ← 2.25 Hz and offset ← 3 µV
       else if i = 2 then
              f_passband ← 1 Hz and offset ← 8 µV
       end if
       Calculate envelope of rectified bandpass filtered C3-M2 using a lowpass filter
       with f_passband
       p_extrema ← points of local extrema of the envelope and its first derivative (ignore
       almost stationary points of inflection on the envelope)
       if rectified band-pass filtered C3-M2 > envelope + offset then
              Mark interval between surrounding p_extrema as SS candidate
       end if
       if SS candidate frequency ≤ 13 Hz and power of bandpass filtered O1-M2 > power of
       bandpass filtered C3-M2 then
              Remove SS candidate [# alpha intrusion]
       end if
       if SS candidate amplitude of any sample in rectified C3-M2 > 85 µV then
              Remove SS candidate [# artifact]
       end if
       if duration of SS candidate < 0.3 s or duration of SS candidate > 3 s then
              Remove SS candidate [# wrong duration]
       end if
       Return detection_i(n)
end for
if ∑_{i=1}^{2} detection_i(n) ≥ 1 then
       result(n) ← 1
else
       result(n) ← 0
end if
for j = 1:no of spindles do
       if duration_j > 3 s then
              Discard j^th spindle
       end if
end for
```


Nature Methods: doi:10.1038/nmeth.2855

Supplementary Figure 9: Example aggregation of the group consensus for experts (e), non-experts (n) and automated spindle detectors (a). Expert, non-expert and automated scores are averaged using the weighted average of the confidence scores at each sample point (green = 'Definitely' =1.0, yellow = 'Probably' = 0.75, red = 'Maybe'/'Guessing' = 0.5, no spindle = 0) and then scored as a spindle (black bars) if the average is greater than the group consensus threshold T_{gc}. T_{egc} = 0.25 for experts, T_{ngc} = 0.4 for non-experts, and T_{agc} = 0.5 is shown. A light line is used to indicate which portion of the epoch an individual annotator viewed.



Nature Methods: doi:10.1038/nmeth.2855

Supplementary Figure 10: Definitions.

Events (E)	- individual spindles in the gold standard dataset.
Detections (D)	- individual spindles annotated by humans or automated algorithms.
Sample	- one datapoint in the timeseries. (ie a signal with a sampling frequency of 128Hz has 128 samples per second).
Subject	- individual whose EEG is examined for spindle events.
Annotator/Detector	- human or automated algorithm that identifies spindle events.
T_{egc}	- threshold for the expert group consensus.
T_{ngc}	- threshold for the non-expert group consensus.
T_{agc}	- threshold for the auto detector group consensus.
True positives (TP)	- correct detection (matches event).
False positives (FP)	- incorrect detection (does not match event).
True negatives (TN)	- correct non-detection.
False negatives (FN)	- incorrect non-detection (event not detected).
Event-Detection pair	- identified D that overlaps with E.

Overlap of ED pair:

$$O_{ED} = \frac{E \cap D}{E \cup D}$$

$$O_{ED} > T_{Overlap} \xrightarrow{yields} TP$$

Recall (Sensitivity; 1- Miss Rate):

$$Recall = \frac{TP}{TP + FN}$$

Precision (Positive predictive value PPV; 1- False Discovery; Selectivity):

$$Precision = \frac{TP}{TP + FP}$$

F_1-score:

$$F1score = 2\frac{Precision * Recall}{Precision + Recall}$$

Specificity:

$$Specificity = \frac{TN}{TN + FP}$$

Negative Predictive Value (NPV):

$$NPV = \frac{TN}{TN + FN}$$

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{N} \; where \; N = TP + TN + FP + FN$$

Cohen's Kappa:

$$\kappa = \frac{\frac{TP + TN}{TP + TN + FP + FN} - \Pr(e)}{1 - \Pr(e)} \quad where \; \Pr(e) = \frac{TP + FN}{N}\frac{TP + FP}{N} + (1 - \frac{TP + FN}{N})(1 - \frac{TP + FP}{N})$$

Matthews Correlation Coefficient (MCC)

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

Nature Methods: doi:10.1038/nmeth.2855

Supplementary Figure 11: Pseudo-code for the intersection-union rule.

[# Locate event spindles and detection spindles. E is the set of event spindles in temporal order and E(i) is the i^{th} event spindle of the total I event spindles. D is the set of detection spindles in temporal order and D(j) is the j^{th} detection spindle of the total J detection spindles.]

```
for i = 1:I do
       if E(i) ∩ D > 0 do
              Obtain the detection numbers that contributes to the intersection [j_1,…,j_N]
              for n = 1:N do
                     O_ED(i,j) ← E(i) ∩ D(j_n) / E(i) ∪ D(j_n)   [# Calculate the intersection union score for the
                     intersecting event and detection]
              end for
       end if
end for

for i = 1:I and for j = 1:J do
       if O_ED(i,j) > Threshold do
              TP_candidate(i,j) ← 1
       else
              TP_candidate(i,j) ← 0
       end if
end for

for i = 1:I do [# first round of matching]
       if ∑TP_candidate(i,:) > 0 do
              idx ← the j with max O_ED with i^th event [# in case of an exact tie choose the
       lowest j]
              Event_match(i,idx) ← 1
       elseif ∑O_ED(i,:) = 0 do
              FN_no intersection(i) ← 1 [# no detection is intersecting with event i]
       end if
end for

 for j = 1:J do
       if ∑TP_candidate(:,j) > 0 do
              idx ← the i with max O_ED with j^th detection [# in case of an exact tie choose
              the lowest i]
              Detection_match(idx,j) ← 1
       elseif ∑O_ED(:,j) = 0 do
              FP_no intersection(i) ← 1 [# no event is intersecting with detection j]
       end if
end for
Best_match = Event_match + Detection_match

for i = 1:I and for j = 1:J do
       if Best_match(i,j) = 2 do
              TP(i,j) = 1 [# max O_ED from both perspectives]
              Best_match(i,:) = 0
              Best_match(:,j) = 0
       end if
end for
```


Nature Methods: doi:10.1038/nmeth.2855

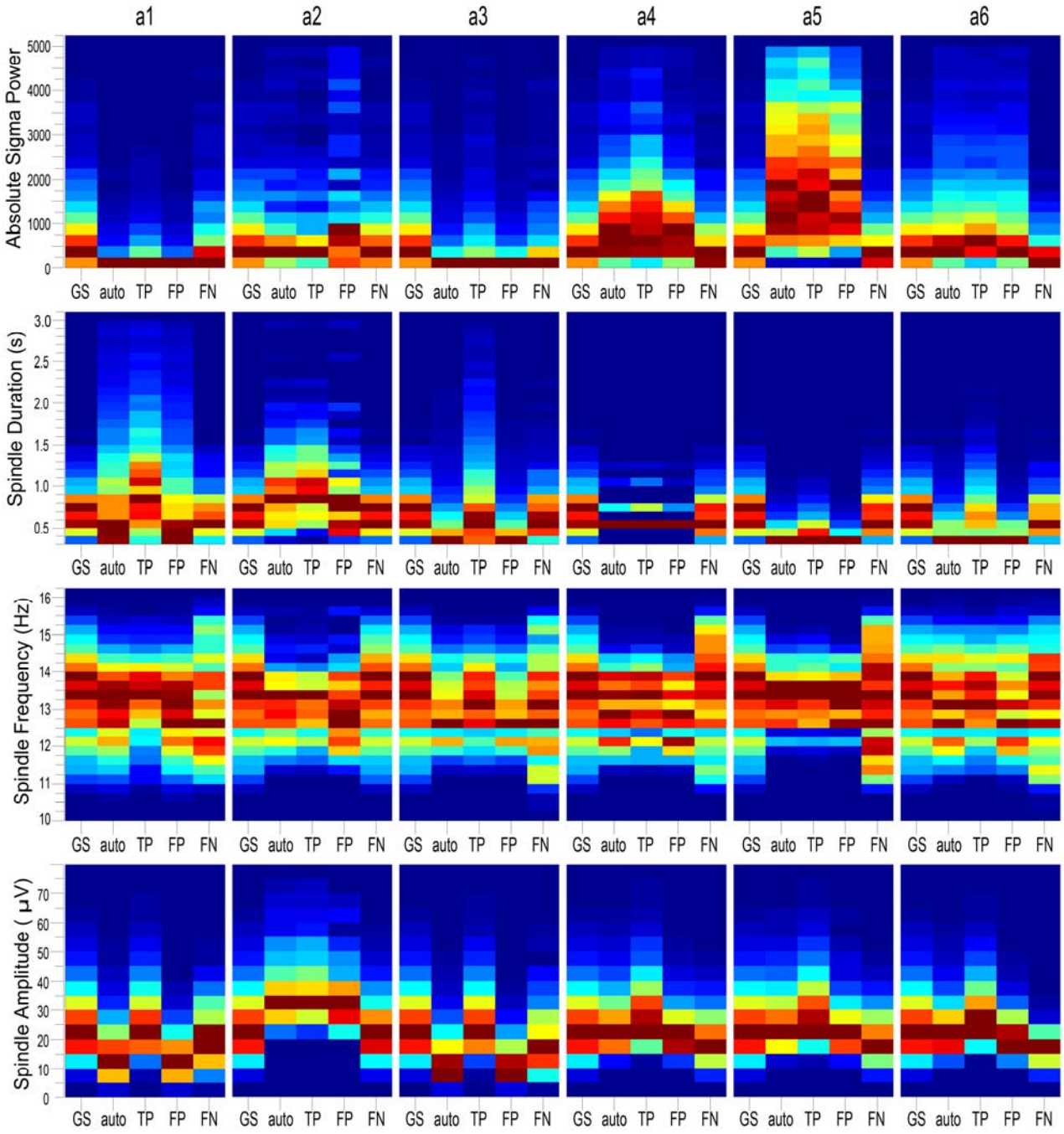
```
if ∑Best_match ≠ 0 do [# second round of matching]
       Create a new O_ED2 with values only where Best_match = 1 and create a corresponding
       TP_candidate2
       for i = 1:I do
              if ∑TP_candidate2(i,:) > 0 do
                     idx ← the j with max O_ED2 with i^th event [# in case of an exact tie
                     choose the lowest j]
                     Event_match2(i,idx) ← 1
              end if
       end for
       for j = 1:J do
              if ∑TP_candidate2(:,j) > 0 do
                     idx ← the i with max O_ED2 with j^th detection [# in case of an exact
                     tie choose the lowest i]
                     Detection_match2(idx,j) ← 1
              end if
       end for
       Best_match2 = Event_match2 + Detection_match2
       for i = 1:I and for j = 1:J do
              if Best_match2(i,j) = 2 do
                     TP(i,j) = 1 [# max O_ED from only one perspective]
              end if
       end for
end if

for i = 1:I do
       if ∑TP(i,:) = 0 do
              FN(i) ← 1
       end if
end for

for j = 1:J do
       if ∑TP(:,j) = 0 do
              FP(j) ← 1
       end if
end for
```

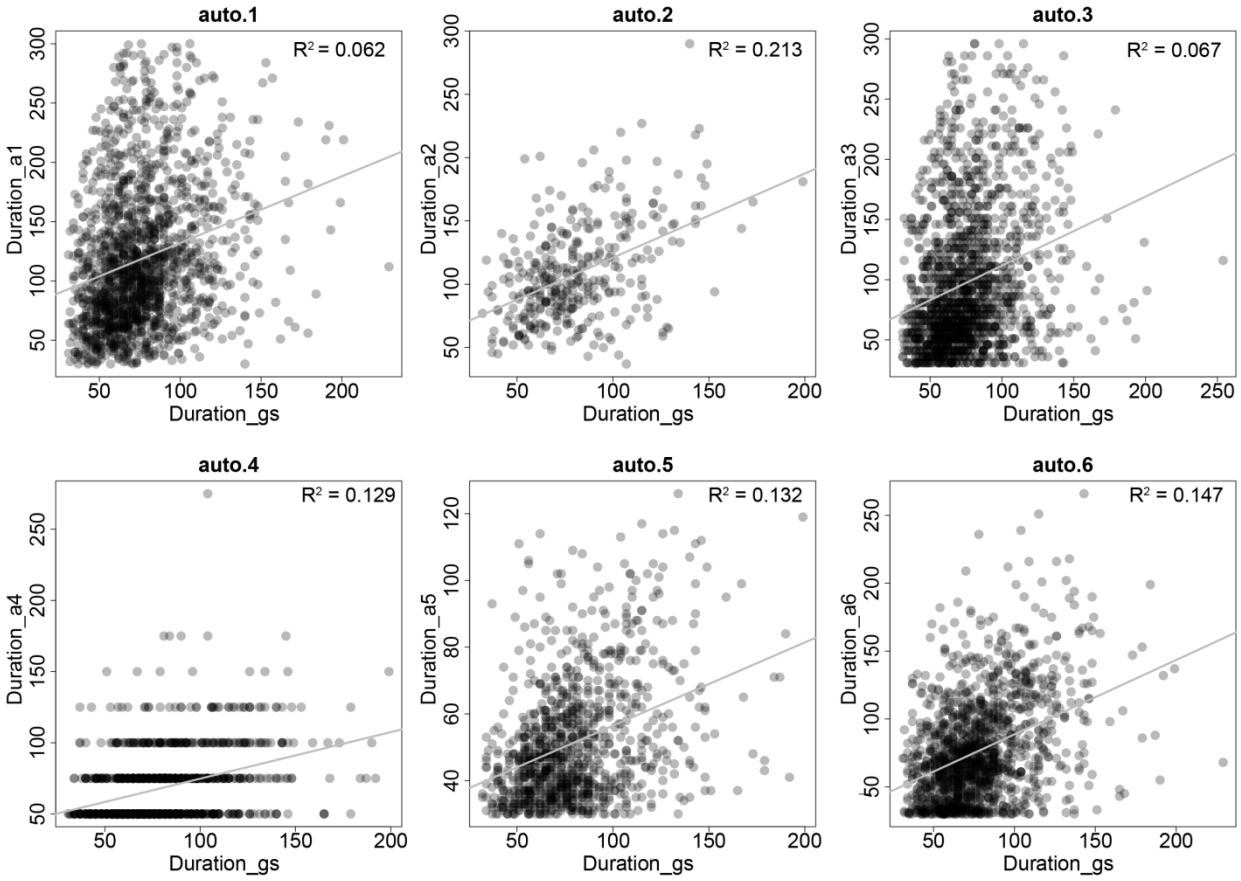

Nature Methods: doi:10.1038/nmeth.2855

Supplementary Figure 12: Spindle characteristics comparison between automatic detections to the gold standard. Absolute sigma power, spindle duration, oscillation frequency, and maximum peak-to-peak amplitude are plotted for each automated detector (a1-a6, indicated at top). In each vertical panel the distribution of the spindle characteristic is plotted for 5 'groups' - spindles in the gold standard (GS), all detections for each detector (auto), as well as the true positives (TP), false positives (FP), and false negatives (FN) for that specific detector. Color scaling indicates the relative distribution of values within one grouping (red = maximum, blue = minimum), in order to compare the relative distribution between groupings. Each detector has a unique pattern of differences between its spindle detections and the gold standard.



Nature Methods: doi:10.1038/nmeth.2855

Supplementary Figure 13: Correlation of spindle duration in the gold standard (Duration_gs) versus automated detectors (Duration_a1-6). Linear regression line and coefficient of determination (R^2) is shown for each automated detector (a1-a6). Each dot in the plots is one spindle. Comparison is made against the gold standard (expert group consensus with T_{egc} = 0.25).



Nature Methods: doi:10.1038/nmeth.2855

Supplementary Table 1: Performance measurements of experts (e), non-expert groups (ng), automated (a) and automated groups (ag). The ag and ng consensus threshold varies from 0.0 to 1.0. Individual a and e are listed by ID number. Additional performance measurements are listed in the by-sample analysis because true negatives can be counted in the by-sample, but not by-event analysis. Performance is compared against the gold standard (T_{egc} = 0.25).

	by-Event			by-Sample									
	F_1-score	Recall	Precision	F_1-score	Recall	Precision	PPV	NPV	MCC	Cohen Kappa	MissRate	False Discovery	Specificity	Accuracy
a1	0.28	0.80	0.17	0.20	0.74	0.11	0.11	0.99	0.24	0.15	0.26	0.89	0.81	0.80
a2	0.28	0.17	0.72	0.26	0.17	0.52	0.52	0.97	0.29	0.25	0.83	0.48	0.99	0.97
a3	0.21	0.83	0.12	0.20	0.71	0.11	0.11	0.99	0.23	0.15	0.29	0.89	0.81	0.81
a4	0.50	0.55	0.46	0.43	0.43	0.42	0.42	0.98	0.41	0.41	0.57	0.58	0.98	0.96
a5	0.52	0.51	0.54	0.42	0.33	0.56	0.56	0.98	0.42	0.40	0.67	0.44	0.99	0.97
a6	0.41	0.71	0.29	0.39	0.57	0.30	0.30	0.99	0.39	0.37	0.43	0.70	0.96	0.94
ag0.0	0.19	0.88	0.11	0.16	0.88	0.09	0.09	0.99	0.21	0.10	0.12	0.91	0.69	0.69
ag0.1	0.19	0.88	0.11	0.16	0.88	0.09	0.09	0.99	0.21	0.10	0.12	0.91	0.69	0.69
ag0.2	0.36	0.84	0.23	0.35	0.75	0.23	0.23	0.99	0.38	0.32	0.25	0.77	0.92	0.91
ag0.3	0.36	0.84	0.23	0.35	0.75	0.23	0.23	0.99	0.38	0.32	0.25	0.77	0.92	0.91
ag0.4	0.52	0.69	0.42	0.48	0.56	0.43	0.43	0.99	0.47	0.46	0.44	0.57	0.97	0.96
ag0.5	0.54	0.51	0.56	0.45	0.38	0.57	0.57	0.98	0.45	0.44	0.62	0.43	0.99	0.97
ag0.6	0.54	0.51	0.56	0.45	0.38	0.57	0.57	0.98	0.45	0.44	0.62	0.43	0.99	0.97
ag0.7	0.45	0.34	0.69	0.35	0.23	0.70	0.70	0.97	0.39	0.33	0.77	0.30	1.00	0.97
ag0.8	0.45	0.34	0.69	0.35	0.23	0.70	0.70	0.97	0.39	0.33	0.77	0.30	1.00	0.97
ag0.9	0.22	0.13	0.83	0.16	0.09	0.84	0.84	0.97	0.26	0.15	0.91	0.16	1.00	0.97
ag1.0	0.22	0.13	0.83	0.16	0.09	0.84	0.84	0.97	0.26	0.15	0.91	0.16	1.00	0.97
ng0.0	0.27	0.82	0.16	0.20	0.82	0.11	0.11	0.99	0.25	0.15	0.18	0.89	0.78	0.78
ng0.1	0.41	0.87	0.26	0.32	0.85	0.20	0.20	0.99	0.38	0.28	0.15	0.80	0.88	0.88
ng0.2	0.56	0.86	0.42	0.48	0.80	0.35	0.35	0.99	0.50	0.46	0.20	0.65	0.95	0.94
ng0.3	0.66	0.76	0.59	0.59	0.66	0.54	0.54	0.99	0.58	0.58	0.34	0.46	0.98	0.97
ng0.4	0.67	0.62	0.73	0.59	0.51	0.69	0.69	0.98	0.58	0.57	0.49	0.31	0.99	0.98
ng0.5	0.58	0.45	0.83	0.48	0.35	0.80	0.80	0.98	0.52	0.47	0.65	0.20	1.00	0.98
ng0.6	0.43	0.28	0.88	0.33	0.20	0.86	0.86	0.97	0.41	0.32	0.80	0.14	1.00	0.97
ng0.7	0.26	0.15	0.91	0.19	0.10	0.90	0.90	0.97	0.30	0.18	0.90	0.10	1.00	0.97
ng0.8	0.11	0.06	0.93	0.07	0.04	0.94	0.94	0.97	0.18	0.07	0.96	0.06	1.00	0.97
ng0.9	0.02	0.01	0.94	0.01	0.00	0.98	0.98	0.97	0.07	0.01	1.00	0.02	1.00	0.97
e01	0.80	0.67	1.00	0.73	0.61	0.90	0.90	0.99	0.74	0.72	0.39	0.10	1.00	0.99
e02	0.65	0.53	0.84	0.62	0.50	0.83	0.83	0.98	0.63	0.61	0.50	0.17	1.00	0.98
e03	0.72	0.59	0.92	0.63	0.49	0.87	0.87	0.98	0.65	0.62	0.51	0.13	1.00	0.98
e04	0.71	0.60	0.89	0.66	0.54	0.87	0.87	0.99	0.68	0.66	0.46	0.13	1.00	0.98
e05	0.74	0.68	0.81	0.71	0.66	0.76	0.76	0.99	0.70	0.70	0.34	0.24	0.99	0.98
e06	0.76	0.64	0.92	0.75	0.68	0.85	0.85	0.99	0.75	0.75	0.32	0.15	1.00	0.99
e07	0.78	0.69	0.91	0.68	0.54	0.92	0.92	0.99	0.70	0.68	0.46	0.08	1.00	0.99
e08	0.75	0.70	0.81	0.63	0.65	0.61	0.61	0.99	0.62	0.62	0.35	0.39	0.99	0.98
e09	0.69	0.71	0.67	0.63	0.65	0.61	0.61	0.99	0.62	0.62	0.35	0.39	0.99	0.98
e10	0.66	0.60	0.74	0.66	0.68	0.65	0.65	0.99	0.65	0.65	0.32	0.35	0.99	0.98
e11	0.77	0.81	0.73	0.72	0.73	0.72	0.72	0.99	0.72	0.72	0.27	0.28	0.99	0.99
e12	0.79	0.97	0.66	0.73	0.96	0.59	0.59	1.00	0.75	0.72	0.04	0.41	0.98	0.98
e13	0.82	0.85	0.80	0.76	0.69	0.83	0.83	0.99	0.75	0.75	0.31	0.17	1.00	0.99
e14	0.76	0.71	0.82	0.73	0.68	0.78	0.78	0.99	0.72	0.72	0.32	0.22	0.99	0.98
e15	0.77	0.75	0.79	0.72	0.69	0.76	0.76	0.99	0.72	0.72	0.31	0.24	1.00	0.99
e16	0.81	0.89	0.75	0.76	0.79	0.73	0.73	0.99	0.75	0.75	0.21	0.27	0.99	0.98
e17	0.62	0.46	0.92	0.60	0.46	0.88	0.88	0.98	0.63	0.59	0.54	0.12	1.00	0.98
e18	0.65	0.96	0.49	0.57	0.89	0.42	0.42	1.00	0.60	0.55	0.11	0.58	0.96	0.96
e19	0.75	0.68	0.85	0.68	0.57	0.84	0.84	0.98	0.68	0.67	0.43	0.16	1.00	0.98
e20	0.80	0.93	0.70	0.78	0.87	0.70	0.70	1.00	0.77	0.77	0.13	0.30	0.99	0.98
e21	0.71	0.90	0.59	0.63	0.87	0.49	0.49	0.99	0.63	0.61	0.13	0.51	0.96	0.96
e22	0.85	0.90	0.81	0.79	0.89	0.71	0.71	0.99	0.78	0.78	0.11	0.29	0.98	0.98
e23	0.78	0.89	0.69	0.71	0.88	0.59	0.59	1.00	0.71	0.69	0.12	0.41	0.98	0.97
e24	0.76	0.65	0.92	0.71	0.61	0.85	0.85	0.99	0.71	0.70	0.39	0.15	1.00	0.99
e.average	0.75	0.74	0.79	0.69	0.69	0.74	0.74	0.99	0.69	0.68	0.31	0.26	0.99	0.98

Nature Methods: doi:10.1038/nmeth.2855

Supplementary Table 2: Leave-one-out performance measurements of individual experts (e1-e24). Experts are compared to an expert group consensus (T_{egc} = 0.25) that does not include their own spindle annotations. Performance is reported using by-event agreement ($T_{overlap}$ = 0.2) and by-sample agreement.

	by-Event			by-Sample			
	F_1-score	Precision	Recall	F_1-score	Precision	Recall	CohenKappa
e1	0.73	1.00	0.57	0.62	0.86	0.48	0.61
e2	0.62	0.84	0.49	0.59	0.83	0.45	0.58
e3	0.63	0.81	0.51	0.52	0.75	0.40	0.51
e4	0.65	0.85	0.53	0.59	0.82	0.47	0.58
e5	0.67	0.75	0.61	0.62	0.68	0.57	0.61
e6	0.69	0.88	0.57	0.66	0.77	0.58	0.65
e7	0.71	0.84	0.61	0.60	0.83	0.47	0.59
e8	0.70	0.73	0.68	0.57	0.53	0.60	0.55
e9	0.62	0.59	0.65	0.55	0.53	0.57	0.54
e10	0.56	0.66	0.48	0.54	0.54	0.54	0.52
e11	0.67	0.63	0.72	0.63	0.63	0.62	0.62
e12	0.67	0.53	0.92	0.59	0.44	0.89	0.57
e13	0.74	0.69	0.78	0.65	0.72	0.59	0.64
e14	0.68	0.75	0.63	0.64	0.70	0.59	0.63
e15	0.70	0.78	0.64	0.62	0.72	0.55	0.62
e16	0.80	0.73	0.89	0.73	0.69	0.77	0.71
e17	0.57	0.90	0.42	0.54	0.83	0.40	0.53
e18	0.58	0.42	0.94	0.49	0.34	0.84	0.47
e19	0.65	0.71	0.60	0.56	0.69	0.48	0.55
e20	0.75	0.64	0.90	0.73	0.65	0.83	0.72
e21	0.55	0.41	0.83	0.46	0.32	0.77	0.43
e22	0.82	0.78	0.87	0.74	0.66	0.84	0.73
e23	0.71	0.61	0.84	0.61	0.49	0.82	0.60
e24	0.71	0.86	0.61	0.64	0.77	0.55	0.63
e.average	0.67	0.73	0.68	0.60	0.66	0.61	0.59

Nature Methods: doi:10.1038/nmeth.2855

Supplementary Table 3: Sleep spindle characteristics in the gold standard dataset. Also see **Figure 2**.

	Mean	SD	Median	Min	Max
Duration (s)	0.75	0.27	0.71	0.31	2.54
Frequency (Hz)	13.31	1.04	13.939	10.48	16.13
Max Peak-to-Peak Amplitude (μV)	27.01	11.02	25.04	4.65	77.92
Symmetry (Percent-to-Peak Amplitude)	0.49	0.21	0.48	0.02	0.99

Nature Methods: doi:10.1038/nmeth.2855

Supplementary Table 4: Inter-detector spindle counts. The total number of spindles found by each detector is indicated in the grey box on the diagonal line. The number of true positive (TP) spindles found by each detector is indicated in the gold standard column (first column). The number of spindles found in common between two detectors is indicated by the intersecting rows and columns. (gs - gold standard (T_{egc} = 0.25), ngc - non-expert group consensus (T_{ngc} = 0.4), agc - automated group consensus (T_{agc} = 0.5).

	gs	ngc	agc	a1	a2	a3	a4	a5	a6
gs	1987								
ngc	1226	1669							
agc	1020	871	1815						
a1	1587	1243	1565	9428					
a2	347	327	464	429	479				
a3	1643	1302	1674	5097	436	13784			
a4	1085	933	1644	1841	452	2083	2362		
a5	1016	865	1557	1517	425	1685	1518	1893	
a6	1411	1203	1586	2804	441	3204	1719	1463	4820

Nature Methods: doi:10.1038/nmeth.2855

Supplementary Table 5: Algorithm over-fitting: optimal parameters and performance. 'Parameter' is the detection parameter for each automated spindle detector that was varied. 'Published parameter value' is the values published elsewhere, and used for testing in this study. 'Over-fit parameter value' is the value of that parameter that produced the maximum F_1-score when the detector was optimized against the gold standard. F_1-score is reported using an event-by-event analysis. Because we have not cross-validated the performance estimates, these F_1-score values are susceptible to over-fitting our data, and therefore are an estimate of the maximum possible performance, and may be an over-estimate of performance in subsequent datasets. However, this data is provided for reference, as the optimal parameter values we obtained may be useful for future studies.

Detector	Parameter*	Published parameter value	Over-fit parameter value	Maximum F_1-score
a1	Threshold ratio for slow and fast spindles	T_{slow}x1 T_{fast}x1	T_{slow}x3 T_{fast}x1.5	0.49
a2	Lower and upper threshold ratio	T_{lower} = 2 T_{upper} = 8	T_{lower} = 2 T_{upper} = 6	0.51
a3	Threshold ratio	1.5	4.5	0.33
a4	Threshold percentile	95^{th}	94.5^{th}	0.50
a5	Threshold ratio	4.5	4	0.53
a6	Low-pass cut-off frequency and off-set	$F_{lowpass}$ = 1 Hz offset = 8	$F_{lowpass}$ = 0.2 Hz offset = 7.75	0.48

*Parameter description:

a1) The algorithm detects slow and fast spindles separately by following the same principles. The envelope of the rectified signal has to exceed a threshold for a spindle to be detected. We multiplied each of these thresholds by values of 0-5 with 0.5 increments (thus the original implementation would have a threshold multiplied by 1x in both cases). In total we investigated the performance of the algorithm with 121 different combinations of thresholds.

a2) The algorithm uses a lower (2x) and an upper (8x) threshold ratio to detect spindles. We tried with lower threshold ratios of 2-5 with 0.5 increments and upper threshold ratios of 1-10. In total we investigated the performance of the algorithm with 70 different combinations of thresholds and found that reducing the upper threshold ratio compared to the published one increased the performance.

a3) The algorithm uses a threshold that is calculated as the mean standard deviation of the band-pass filtered signal times a constant (1.5x). We tried with constants ranging 0.5-10 with 0.5 increments. In total we investigated the performance of the algorithm with 20 different constants.

a4) The algorithm uses a percentile of the RMS signal to determine the threshold (95^{th}). We varied the percentile from 90-99.5 with 0.5 increments. In total we investigated 20 different percentiles and reached optimal results very close to the results with the published parameters.

a5) The algorithm uses the signal mean, times a constant (4.5x) to define the threshold. We varied the constant from 0.5-10 with 0.5 increments. In total we tested 20 different constants.

a6) The original implementation uses a combination of two detectors whereas the over-fitting attempt only uses one detector. The algorithm uses low-pass filters (2.25 and 1 Hz) with off-sets (3 and 8) to define the time varying threshold that the rectified band-pass filtered signal needs to exceed. We tested 11 low-pass filters and 21 off-sets for a total of 231 different combinations.

Nature Methods: doi:10.1038/nmeth.2855

Supplementary Table 6: By-subject spindle density estimates. Mean is the group average of the average spindle density (spindles / min) of the 110 sleeping subjects in the study. Also see **Figure 5b**.

	mean	SD	median	min	max
a1	12.22	2.66	12.26	4.17	22.70
a2	0.61	0.62	0.52	0.00	2.87
a3	16.31	12.80	10.17	0.26	45.65
a4	3.09	1.36	2.87	0.00	7.04
a5	2.44	1.47	2.35	0.00	7.83
a6	7.17	5.12	5.86	0.26	23.22
autogroup0.5	2.40	1.50	2.35	0.00	7.57
non-expertgroup0.4	2.15	1.60	1.83	0.00	7.30
gold standard	2.32	2.01	1.70	0.00	9.91

Nature Methods: doi:10.1038/nmeth.2855

Supplementary Table 7: By-subject spindle duration estimates. Mean is the group average of the average length of spindles (seconds) of the 110 sleeping subjects in the study. Also see **Figure 5c**.

	mean	SD	median	min	max
a1	1.04	0.18	1.03	0.58	1.49
a2	1.02	0.28	1.03	0.43	1.63
a3	0.60	0.17	0.55	0.41	1.31
a4	0.63	0.07	0.63	0.50	0.83
a5	0.46	0.07	0.45	0.31	0.68
a6	0.55	0.11	0.54	0.32	0.98
autogroup0.5	0.53	0.09	0.52	0.32	0.79
non-expertgroup0.4	0.66	0.14	0.62	0.43	1.34
gold standard	0.74	0.15	0.73	0.43	1.40

Nature Methods: doi:10.1038/nmeth.2855

Supplementary Table 8: Demographics of subjects in the gold standard EEG dataset.

	mean	SD	min	max
Age	56.92	7.81	42.23	71.84
Body Mass Index	31.41	7.43	18.30	60.18
Apnea-Hypopnea Index	5.72	8.23	0.00	35.70
Total Sleep Time	375.47	58.41	202.00	515.50
Stage N2 Minutes	256.66	49.51	79.00	376.00
Leg-Movement Index	26.54	28.33	0.00	121.40
Sex (% Male)	0.47			

Nature Methods: doi:10.1038/nmeth.2855