Nature and Science of Sleep

Dovepress
open access to scientific and medical research

Open Access Full Text Article

EXPERT OPINION

The Future of Sleep Staging, Revisited

Neil Stanley

Independent Sleep Expert, Farnborough, Hampshire, UK

Correspondence: Neil Stanley, Email drrneilstanley@yahoo.co.uk

Abstract: In 1996, I published a paper entitled "The Future of Sleep Staging". At this time, paper and ink records were the standard way of recording sleep records. Computerised systems had only recently become commercially available. The original article was a response to those initial computer-based systems, pointing out the potential limitations of the systems. Now, digital sleep recording is ubiquitous and software and hardware capabilities have improved immeasurably. However, I will argue that despite 50 years of progress, there has not been an increase in the accuracy of sleep staging. I will propose that this is due to the limitations of the task that we have set the automatic analysis methods.

Keywords: sleep staging, Rechtschaffen and Kales, EEG, artificial intelligence

The Past

> To be ignorant of what occurred before you were born is to remain always a child. For what is the worth of human life, unless it is woven into the life of our ancestors by the records of history? Cicero, Orator

As Cicero alludes to above, we must know the past to understand the present and future of sleep staging. In 1968 the first manual standardising the scoring of sleep stages was published.[1] This manual, universally known as Rechtschaffen and Kales, or simply R&K, was compiled over several years by an ad hoc committee of 12 leading sleep researchers. The description of sleep given in R&K is thus a purely human construct that does not adequately describe the complexity of the human EEG (there is a perhaps apocryphal story that the door had to be locked on the committee until they agreed on the criteria for stages 3 and 4). While initially, R&K was a good tool for the recording technology of the time, it remained unchanged for far too long, "The guidelines of Rechtschaffen and Kales (R&K) were meant as a reference method. However, it became, unintentionally, a gold standard".[2]

The first notable feature of R&K is that it was solely designed for staging from paper traces. From Berger's first recordings of the EEG in 1929 until at least the mid-1990s, all sleep records were recorded and staged on paper. At the time of the genesis of R&K, signals were recorded onto paper running at 10mm/sec, giving a 30-s epoch on a 30cm sheet of paper. Impedances on all electrodes were checked and were all below 5kΩ at the start of the recording. Amplifiers were calibrated, so the deflection of the pens was correct in amplitude. Before starting the sleep recording, a bio-calibration was performed; the patient/subject was instructed to open eyes, close eyes, look left, look right, look up, look down, blink five times and finally grit their teeth. Recordings were always attended, mainly for simple, practical reasons, such as refilling the ink wells, uncrossing jammed pens, clearing paper jams, and ensuring the paper folded correctly. However, from a technical point of view, attended recordings also allowed real-time changes to the recording gains and sometimes the recording montage in the event of artefacts or electrode loss.

Another thing to be aware of is that R&K only described and illustrated a few data channels. This was because of the considerable expense of amplifiers at the time, limiting the number of routinely recorded traces. A typical sleep recording montage, using the 10–20 system, would be C_4-A_1, O_2-A_1, right and left electrooculogram (R_{EOG}, L_{EOG}) and submental electromyogram (EMG).[3] Electrodes for recording C_3 -A_2 and O_1-A_2 were often also applied to act as backups in the event of electrode loss.

Received: 3 February 2023
Accepted: 19 April 2023
Published: 2 May 2023

The accuracy of sleep staging using paper records was facilitated because the paper was printed with squares, with each large square horizontally representing 0.5 sec. Each large square was subdivided into five smaller squares; therefore, visually judging phenomena with 0.1-s accuracy was easy. The amplitude of the various waveforms depended on the electrode sensitivity used during the recording. These could vary for several reasons, but any change was marked on the trace. This meant it was easy to visually identify 0.5 sec or 75μV peak to peak (being mindful of changes in sensitivity) and even easier to measure amplitude and frequency using a ruler. Paper recordings also allowed easy identification of the lowest EMG level of the night and permitted easy reference to the bio-calibration and forward and back searching for various events. Paper recordings allowed two pages to be displayed simultaneously and where necessary, four or more pages were easily viewed with no loss of scale, facilitating the accurate identification of entry and exit of sleep stages.

In the sleep laboratory where I learnt to stage sleep (Neuroscience Division, Royal Airforce Institute of Aviation Medicine, Farnborough, UK), paper records were then independently scored by two proficient sleep stagers; their staging was then consolidated, i.e., where there was concordance of staging, this stage was retained. Where there was disagreement, these epochs were discussed between two experienced stagers (not necessarily the original scorers) to agree on the sleep stage to be assigned. In the event of a further lack of agreement, the issue was referred to a more experienced stager. A well-thumbed copy of R&K was always at hand during this process. Scorers were generally sleep staging with someone who had also been taught how to sleep stage within the same team; this meant that the concordance of staging was very high. There were, of course, individual differences in applying the R&K rules (I worked with someone who barely ever staged stage 4 and another who was very eager to give stage 4). However, the influence of such biases was minimised because of the double staging and the rectification process. Thus, the final agreed stages were as accurate to R&K as humanly possible. Although it should be noted that in 35 years of staging sleep, the author has only ever seen two people with sleep EEG traces that could be accurately staged merely by reference to the R&K criteria. Because members of a particular research group are taught to stage in the same way by their colleagues and learn how others in the same team stage in practice, they end up defining and adopting their own "norms" and means for interpreting R&K, leading to the intragroup accuracy of sleep scoring that was shown by several studies.[4,5]

AASM Rules

The criteria for sleep staging in R&K were intended to reaffirm the partitioning of sleep into stages proposed by Dement and Kleitman (1957).[6] R&K also sought to revise and expand these criteria to consider the research done in the intervening 10 years. Rechtschaffen and Kales say in their concluding comments,

> This handbook should be viewed as a working instrument rather than a statute. Experience with the manual may suggest possible revisions. When these suggestions accumulate appreciably, it would seem in order to have a review of the manual.

In response to the lack of revision of R&K despite a massive increase in the number of sleep recordings being performed worldwide and envisaging the technological future, I wrote in 1996[7]

> Any new sleep staging criteria must be exhaustively comprehensive and because of this computers will probably be relied upon to a much greater degree, as the set of rules must be accurately, reliably and reproducibly applied in all centres.

However, R&K had become a de facto "gold standard" and despite the expressed desire that R&K be modified and various criticisms of R&K,[2] it was not until 2003 that such a revision was initiated by the board of directors of the AASM who approved a proposal to develop a new manual for sleep scoring. In 2004, a task force was set up to develop reference material and support the new manual. Although the new sleep scoring rules were published in 2007,[8] long after digital sleep recording had become widespread, the authors seemingly took little or no account of the fact that sleep was now predominantly being staged on screens, not paper.

The new AASM guidelines contained relatively few minor modifications to the rules of R&K. Interestingly, despite the technological advancements in signal processing and analysis in the intervening years, the new rules, rather than getting more complex, eg, introducing additional sleep stages to give a better description of sleep as envisaged in their introduction by R&K, were simplified. There were now three stages of NREM, not four as per R&K, which itself was a simplification of the five stages proposed by Loomis in 1935.[9]

These new rules were found to have a significant effect on sleep staging when compared to R&K.[10] Sleep latency and REM latency, total sleep time, and sleep efficiency were not affected by the classification standard; the time (in minutes and in per cent of total sleep time) spent in sleep stage 1 (S1/N1), stage 2 (S2/N2) and slow-wave sleep (S3+S4/N3) differed significantly between the R&K and the AASM classification. While Stage 1 (+2.8%) and slow-wave sleep (S3+S4 vs N3, +2.4%) significantly increased, stage 2 sleep decreased significantly according to AASM rules (–4.9%). Moreover, wake after sleep onset was significantly prolonged by approximately 4 min according to the AASM standard.

The new rules also led to changes in the inter-rater reliability of sleep staging, with the modification of the scoring rules improving IRR due to the integration of occipital, central and frontal leads. However, there was a decline in IRR specifically for N2 due to the new rule that cortical arousals with or without a concurrent increase in submental electromyogram are critical events for the end of N2.[11]

Because the new AASM guidelines produced statistically significant differences in the scoring of sleep stages, it is wrong to state that sleep staged according to the AASM guidelines is being staged according to R&K. Yet, this is repeatedly asserted in the literature.

R&K, and subsequently the AASM scoring rules, were written for a different time and way of sleep scoring. Because of this, I believe that the current AASM scoring manual, like R&K before, is the primary factor holding back accurate computerised sleep scoring or the implementation of Artificial Intelligence (AI)/Machine Learning (ML) scoring algorithms.

Yet, despite the failings of R&K and the fact that the AASM revision led to no improvement in staging accuracy, Lee et al 2021,[12] seem to envisage that there is still a necessity for improving the guidelines for manual scoring writing

> These results can serve as baseline data with which to judge whether the guidelines to be updated improve the interrater reliability.

But they do not state how such a further revision of the staging guidelines is to be achieved. However, while they claim that the inaccuracy of the human scorer is a reason to improve the guidelines, they claim that automatic systems are "accurate" even though they use the same guidelines.

This contrasts with my view that

> that humans are not up to the task of using such a criterion, but with the correct algorithms, computers could be.

As I wrote in 1996, the answer is not to rewrite the guidelines for human stagers but

> The solution must be to rewrite R&K for the computer age. This must be done as soon as possible, as failure to do so will result in automatic sleep systems being developed, all with various and different algorithms for staging sleep.

Unfortunately, this complete lack of standardisation between systems is exactly what has occurred. We have ended up with imperfect staging rules being imperfectly implemented, in numerous different ways, by dozens of companies, using proprietary algorithms, most of which are not in the public domain, all of which claim they are staging sleep according to the R&K/AASM guidelines. As foreseen in 1996, this lack of standardisation between systems is a real problem as we have no way of knowing how these systems are staging sleep. Bandyopadhyay & Goldstein,[13] in what seems to be an effort akin to closing the stable door after the horse has bolted, state

> With multiple commercial companies developing FDA cleared algorithms, there is a need to standardise commercial algorithms through certification by an accredited regulatory body.

Few people today who are designing automatic sleep scoring algorithms, championing the possibilities of AI/ML, etc., have experience staging sleep according to R&K on paper. Instead, they see sleep staging merely as an engineering/programming problem. Because of this view, they see it as a challenge to implement an algorithm replicating R&K/AASM.

The Human/Computer Accuracy Paradox

When comparing human and computer staging of sleep, papers frequently reference the sleep staging being performed by an "experienced" sleep stager, yet they do not define what defines an "experienced" stager. Being "experienced" is not

a quantifiable measure of quality. (I, for instance, have staged >25,000 sleep records both in research and clinically. I learnt to score sleep in 1982 on paper using a Grass 8–10 and Nihon Kohden. In the computer age, I have scored sleep using equipment from Oxford (Medilog 9000 and 9200), Temec, Nicolet, Compumedics, Embla, Alice. I last manually staged a sleep record in April 2017 using the Nox system). Thus, the "experience" of the stagers used in comparison studies is unknown and unquantifiable. Although many people claim a human-to-human accuracy of 75–85%, the actual figures are far more nuanced, as seen by the work of Ferri et al,[5] who showed an overall epoch-by-epoch agreement of between 60.9% and 96% in a study of nine different Italian sleep laboratories, comprising 17 specialists. For the five example records analysed, two of which were identical, they found significant intergroup differences in identifying stages 2, 3 and 4. An epoch-by-epoch "consensus" analysis was obtained from the nine groups using the most frequent value scored for each epoch. The intergroup agreement was between 47.5–100% for stage 0; 33.3–96.3% for stage 1; 39.9–99.1% for stage 2; 3–100% for stage 3/4 and 62.6–100% for REM stage. They also found that seven of the nine laboratories agreed more than 84% of the time; they concluded that intergroup ratings of 80% or better should be considered acceptable as evidence of the consistency with which staging could be performed. However, the variability inherent in the scoring of sleep records shows that R&K is not the "gold standard" to which it is so often referred.

However, this inaccuracy of human sleep staging has somehow been transformed into a measure of automatic/AI staging accuracy. The argument goes something like this; human scorers are XX% accurate at staging sleep; therefore, if automated/AI staging is close to or even better than XX% accuracy, this is satisfactory or even superior to human scoring. This is a widespread conceit in the literature; for instance, Stanus et al 1987[14] write, "Average agreement rates of both methods compared to expert visual scoring were very similar, although a few specifics occasionally appeared for partial sleep stages. The comparison of more than 40,000 sleep decisions (on 20-sec epochs) yielded 75% absolute reliability for normal controls and 70% for pathological cases. However, if the agreement rate obtained for routine visual scoring (82%) in our sleep laboratory is considered satisfactory, our system is then 90% satisfactory".

However, the question that seems never asked is against what standard do the human scorers agree 82%? If R&K/AASM are not the "Gold standard", then what is? Sleep staging has no definitive standard; there is nothing like the standard kilo or metre. There is no perfectly staged sleep record(s) against which to benchmark human scorers or automatic scoring algorithms. The accuracy of automatic sleep stagers is based on false logic, as it is compared to the standard of two human sleep scorers who only agree 75–85% of the time. From this, it is intuited that as long as the staging agrees with a single human sleep scorer at approximately the same rate, the algorithm is as good as human staging. This fails logically because we do not know the precise accuracy of the human scorer's staging, as there is no reference with which to compare. They could be 75–85% accurate to R&K/AASM, but equally, they may only be 50% accurate, 25% or even 0%. The accuracy of the human scorer is arbitrarily assumed, whereas, in practice, it is unknown and indeed unknowable.

Rechtschaffen and Kales, in the foreword to their manual, stated,

> An evaluation of how much such standardisation contributes to the reliability of scoring will have to await the development of experience with the system and empirical testing.

This has never been done, but in my original paper, I proposed a solution to this problem

> It is imperative to find out how human scorers actually stage sleep records and to identify important and false cue utilisation.

I argued that in order to be useful, it is imperative that the records scored must include examples of all known deviations from the rules of R&K, eg, sleep in geriatrics, drug-induced sleep, sleep in depressed patients and poor quality recordings.

This process would

> allow rules for sleep scoring to be drawn up which are set and a good descriptor of underlying phenomenon.

Attempts at such a standardised database have been implemented, eg the SIESTA project[15,16] and the development of the European Data Format,[17] but, however worthy these endeavours, their impact has been somewhat limited.

Technology Means That Humans are Even More Inaccurate

Recording and staging sleep on paper is now a thing of the distant past; the first commercially available computer-based sleep systems started to appear in the mid-1980s. The new systems' initial attraction was their ability to store a seemingly vast amount of data in a minimal space rather than 1/3 of a mile-long paper traces that were sometimes required to be stored for 5 years. But with the advent of computer systems came the idea that they could also be used to stage sleep automatically. In departments with a long tradition of staging sleep, this was a mere novelty and human staging was still the default sleep staging method.

With the introduction of the AASM guidelines, it was stated[18] that

> No visual based scoring system will ever be perfect, as all methods are limited by the physiology of the human eye and visual cortex, individual differences in scoring experience, and the ability to detect events viewed using a 30-second epoch.

Yet, while this statement is relevant to staging on paper, it takes no cognisance of staging on the screens of computerised systems. Initially, when they became commercially available, computerised systems only had a small screen that was less than ideal for staging (indeed, for a time, we printed out the digital records onto paper using a high-speed Mingograf printer, we were blessed with a lot of storage space). However, over time staging on such a small screen became the norm. However, one main issue is that the computer monitors, due to their size, do not present traces as though they were a 30cm page. Even with the introduction of larger monitors, this issue has not been addressed. Events such as K-complexes, sleep spindles, etc., are thus not represented as described in the AASM scoring manual and, therefore, the visual judgement of 0.5 sec or 75uV peak to peak is challenging. This is exacerbated by the almost infinite changes the scorer can make to how the data is presented. Quick and easy measurement of these events by a ruler is practically impossible. Rather than replicate the look of the pen and ink on a 30cm page of squared paper, at best, screens only usually provide 0.5-sec grid lines, although these are typically optional and do not provide any vertical grid lines. Perhaps how far removed we are from R&K can be illustrated by the fact that in the relevant chapter of the most recent edition of Principles and Practice of Sleep Medicine,[19] examples of sleep stages are shown with neither the x nor y-axis defined.

The inaccuracy inherent in scoring from screens is further compounded by the use of 4k monitors and very high sampling rates, which means that the scorer is now presented with traces that look less and less like the examples given in R&K. (Because we have not seen waveforms/events before we do not know if they are a unique phenomenon, a hardware recording artefact or a meaningful population-wide brain phenomenon).

Computerised PSG systems can also record a large number of data channels. The price of amplifiers is no longer a factor; a scorer may be presented, or choose to be presented, with data from electrode derivations that are not described in the R&K/AASM criteria. Again, this moves any staging away from that set out in R&K/AASM.

It is also possible for the user to modify the sleep scoring criteria and thus move even further away from the "Gold standard" and introduce even more inaccuracy.

Automatic Staging Inaccuracy

Automatic sleep scoring first became commercially available in the mid-1980s.[20–22] There have also been numerous algorithms proposed since R&K for measuring/describing sleep. In the early 1990s, several analyses unrelated to R&K, eg Cyclic Alternating Pattern (CAP),[23–25] wavelets,[26,27] and neural networks,[28,29] were reported to be able to analyze sleep accurately. However, these techniques have failed to gain acceptance outside the research laboratory.

Essentially, the AASM scoring manual makes it impossible for a computer to mimic human sleep staging accurately and thus achieve accurate computerised sleep scoring. The problems with R&K/AASM for automatic sleep staging were identified by Hirshkowitz and Moore (1994)[30] when they asked, "How can a computer be expected to agree with human stagers if two human stagers do not agree with each other?"

For example, the rules for assigning stage 3 (delta >75 μV peak to peak, <2 cps >20%, <50% of epoch) and stage 4 (>50% delta) were the most unambiguous in the R&K. Yet, the wide range of values found in the Ferri et al[5] study illustrates that even simple rules are applied inaccurately or inconsistently. If the rules of R & K were fit for purpose, then squared paper and a ruler would be sufficient to stage slow-wave sleep accurately. However, experience dictates that

inter-rater reliability becomes more variable with the increasing complexity of sleep recording. The lack of agreement between stagers using R & K for straightforward scoring does not augur well for consistency when patients with a significant physical pathology and greatly fragmented sleep are evaluated. Thus, if it is difficult for two human stagers to agree on interpreting a set of EEG data, it will be almost impossible for a software engineer to design a computerised version of the rules to be superior. Designers of automatic sleep staging systems are restricted to making their machines mimic AASM as closely as possible. Yet they have condemned themselves to produce software that scores sleep inaccurately by doing so. R&K was written for human stagers when computers were essentially non-existent in sleep science. However, the descriptions of sleep stages seem highly amenable to automation; R&K was never written with computers in mind. This can be most clearly seen with the R&K definitions for stages 3 and 4; the definition given in R&K should be easy to convert to a rule an algorithm can run. However, such basic criteria mean that any such algorithm will identify any waveform, however spurious, that fits the criteria. As the old saying goes, "Garbage in, Garbage out" and with regard to automatic sleep staging, the limitations inherent in the R&K/AASM scoring rules can, in my opinion, be regarded, although completely understandably, as "garbage", not fit for purpose.

Simply by virtue of its capabilities in acquiring, filtering and processing raw data, a computer should be much more accurate and error-free than a human scorer in its identification sleep stages, according to R&K/AASM. The new system's high sampling rates also mean that automatic staging algorithms analyse information that was impossible to obtain when R&K and the current AASM manuals were written. This can be illustrated by the work of Hori et al 1994,[31] who classified the EEG patterns during the Sleep Onset Period (SOP) into 9 EEG stages. If there is this degree of complexity in the short Sleep Onset Period, imagine the actual complexity of sleep across all the AASM sleep stages. Rather than discussing the clinical relevance of these nine "stages", the revision of R&K made sleep less complex. Sleep EEG is vastly more complex than AASM staging allows. At the same time, a computer has to analyse all this information; it reduces the complexity of what it "sees" to match the AASM criteria. Surely, the point of technology is to be able to do complex things better/faster than humans. Even when we make the staging rules simpler, technology still cannot perform sleep staging better than a human.

The human sleep scorer's ability to pick out the various frequencies in a 30-s epoch of data is much less than a computer working to the same rules. It is, therefore, possible that the difference between human and computer scoring lies in the fact that a computer applies R&K/AASM in a completely rigid and reproducible manner. As such, it is possible that in some instances, eg, stages 3 and 4, the automatic scoring is a better descriptor of the rules of R & K than a human stager. It is not surprising that there are differences; the human scorer may miss important information, or it may lead to an over/underestimate of the amounts of specific frequencies in an epoch, whereas a computer working to a rigid set of rules cannot do this. However, due to the nature of R&K/AASM, the computer's answer may be wrong, even though the way it achieves the result will appear to be consistent and reproducible.

Why Automatic Sleep Staging is Inaccurate

The main problem with the current AASM guidelines, as with R&K before them, is that they are not universal rules. R&K said their manual was applicable for adult humans, yet 80% of the examples of EEG traces given were from normal young males. Because the ASSM manual is based on R&K, it appears that the rules merely describe young, healthy males. (It is also interesting that the AASM scoring manual does not give examples of PSG traces). Such subjects are not representative of the general population. They are even less typical of insomniac and sleep disorder patients, who are the ones that are routinely having their EEGs recorded for diagnostic purposes. Significant interindividual differences are found in EEG; these can be compounded by factors such as age, health, medications, drugs of abuse, etc. Therefore, a sleep stager may only ever see a few records entirely representative of R&K/AASM. The remainder of the time, the expert sleep stager must fit the data to the rules in the best possible manner, with varying degrees of success. With every sleep record staged, it is necessary to continuously interpret the scoring rules according to the scorer's experience; this is done in an unquantifiable and sometimes idiosyncratic manner. However proficient or "experienced" a particular sleep expert may be, the ambiguity and lack of universality of R&K inevitably lead to differences in the accuracy and consistency of identifying sleep stages between human stagers. (Although, as mentioned, double scoring the data and reaching a consensus can, to a degree, ameliorate this inaccuracy). However, an algorithm cannot deviate from the

criteria by which it is programmed, and therefore cannot adapt to changes seen in EEG, for instance, in depressed subjects who have reduced alpha;[32] fluoxetine users that have prominent eye movements in NREM;[33] the low amplitude delta seen in the elderly;[34] or alpha-delta sleep.[35] How can the algorithm accurately stage if it cannot apply modified criteria the way a human sleep stager could?

As I have pointed out, the accuracy of automatic staging needs to be assessed in various populations, eg, across age groups and in different pathologies. This has not been done with any current or proposed automatic/AI/ML scoring technology (however, it is also true that no assessment of this kind has been conducted with human sleep scorers either). Therefore, it is correct to hypothesise that it is impossible to reliably measure the accuracy of automatic/AI/ML staging in a diverse selection of subjects/patients if such data does not exist for human scoring. Automatic staging based on the R&K/AASM rules cannot be expected to accurately score sleep in the different age groups or pathologies seen in the sleep clinic or research laboratory.

The same problem would be encountered in the way computers identify and remove artefacts caused by patient sweating, patient movement or possible contamination of signals due to the ubiquity of electrical devices such as phones, tablets, or headphones in the bedroom as well as the signals from transmitting devices, such as wireless routers, keyboards, and monitors.

If automatic scoring/AI/ML systems are not using the EEG channels given in AASM, then they will, by definition, produce different results from a human stager. Unless, of course, the human is also scoring using these non-ASSM channels, but if this were the case, then they are not staging according to the AASM criteria. Therefore, if non-standard EEG channels are used, it is inevitable that automatic scoring/AI/ML will be different from a human scorer. Analysing data from numerous channels again moves the analysis further from the R&K/AASM rules. More importantly, if data from non-standard channels are used, this inevitably means that the human scorer is an imperfect comparator. This is particularly true of systems that utilise merely a single channel of EEG.

Another reason for the disagreement between human stagers and automatic staging is when recording on paper; the fold in the paper delineates each 30-s epoch. Whereas with digital systems, the epoch is defined as 30-s from the start of the recording period. This could potentially mean a mismatch between the exact timing of the waveforms seen in a particular 30-s epoch, eg, the first k-complex or sleep spindle. This could influence the epoch in which the start or end of a specific sleep stage is given.

Also of importance is how do digital systems use filtering. Low and high pass filters can reduce or eliminate specific frequencies but equally filtering out a signal too much can cause artefact. How is this implemented in each system? And what effect does this filtering have on any automatic staging derived from the filtered signals?

While many studies have been published on automatic sleep staging since my paper in 1996, it is unnecessary to review them. The simple fact is that in 1971, Sith and Karacen[36] reported an 83% accuracy for their analysis against human scoring. In 1973, with an "entire programme that consisted of 1664 (3200 octal) 16-bit words", Gaillard and Tissot[37] reported a mean accuracy against visual scoring of 83.71%. Fifty years later, after all this time and effort, accuracy for various commercially available devices was, according to the review by Fiorillo et al[38] reported to be "on average around 85%", ie, there has been zero progress in the accuracy of automatic staging despite the unimagined increase in computing power and sophistication and complexity of staging algorithms. A simple automatic sleep staging based on R&K/AASM cannot work; commercial interests and marketing drive the hyperbole about automatic sleep staging and, I believe, ignores the reality of sleep science.

The "Impossibility" of Sleep Staging Using AI/ML

Although neophiles seem to consider AI/ML the future, AI sleep staging has been under development since at least the mid-1990s[39] and yet, as Bandyopadhyay and Goldstein,[13] in their review of the literature show, the accuracy of AI/ML sleep staging is still only approx. 85% (a figure strangely similar to that reported using non-AI automatic scoring).[38] However, it is interesting that the authors do not comment on this level of accuracy or its clinical significance.

However, AI/ML may provide a way forward despite this inaccuracy, but not how its advocates imagine. Again, they are trying to compare AI/ML with R&K. If AI/ML can make its own rules, how can this be compared to a human

working to a staging criterion, however, inaccurately? However, AI/ML potentially has the possibility of helping us get a better definition of sleep; in 1996, I wrote

> Rather than having an expert committee come up with a set of rules for humans that are then translated into an automatic staging algorithm, we need to do precisely the opposite; we need to find out what a computer can measure

Once this is done I proposed that an expert committee would decide if what can be measured by the computer was actually of any descriptive utility in staging sleep. I recommended that

> The computer would need to measure sleep from numerous types of populations.

Once the experts had decided on which waveforms provided utility in staging the

> It will be possible to marry the human experience with a computer's ability. The performance of the computerised rules must be equal to that of the human scorer and would additionally include criteria that humans are unable to apply.

Sleep medicine does not need computers to be as good as humans; it needs them to be better. As an outcome of the above process

> A comprehensive, reproducible set of staging algorithms would result and be adopted as the standard. All automatic sleep stagers would then be required to use this standard.

Simply using AI/ML to learn about EEG and then deciding the clinical relevance of what is found could 1) move forward sleep science and 2) lead to reproducible and accurate automatic sleep staging. It says a lot about inertia in the field that, in 2022, Bandyopadhyay and Goldstein[13] echo my thoughts

> Most of the existing datasets using polysomnogram data are research datasets collected from a subgroup meeting certain inclusion criteria. Hence, they are not generalisable and not representative of what the clinician encounters in real practice.

In order to ensure this they propose that

> There is an acute need for larger-scale research trials which can corroborate machine algorithm generated measures to clinically significant outcomes. This prompts the need for research datasets with heterogeneity in signals, patient demographics, sleep disorders, and clinical outcomes.

The Future

As Hirshkowitz and Moore wrote in 1994:[30]

> We clearly need more properly conducted research that tests automated system performance. Until a competent literature amasses in the appropriate arena, confidence remains the "bottom line" issue.

It is, I believe, impossible to automate the accurate staging of sleep according to R&K/AASM as all AI/ML-based systems that use R&K/AASM as the basis for their rules have an inherent and, I believe, fatal weakness built-in. The future of sleep staging looks very similar to the future I envisioned in 1996.

> The research required to draw up a new set of rules will be time-consuming and expensive, but that should not discourage the sleep community from rising to the challenge. The potential benefits outweigh the problems.

My final call to action was that

> Rewriting R&K for the computerised future of the 21st century cannot be seen as too difficult to attempt. It has to be done and must be started now.

The lack of progress in developing sleep scoring rules in light of the possibilities of new technology coupled with the multitude of commercially available systems using a myriad of different algorithms now available means that it will be

more challenging than it has ever been to bring about standardisation of sleep scoring and the optimisation of scoring accuracy.

However, there is a belief from some that progress has been made; for instance, "In our view, automatic sleep staging with PSG on healthy people has basically been solved, not only for adults but also for children" (although the authors give no indication why healthy people would undergo PSG examination).[40] Yet despite this seemingly positive statement, the authors admit that we are, in fact, no closer to having accurate automatic sleep staging in the clinic,

> It is not sufficient to have high-quality PSG scoring of healthy people only. In a clinical setting, the tools applied should be equally capable when confronted with non-textbook sleep phenotypes

This is important in cases where

> The sleep EEG may either be masked by disease-related artifacts, or where the sleep EEG itself may be so drastically changed by the patient's condition.

They conclude that to be more widely adopted clinically,

> Automatic sleep scoring should be as robust as manual scoring.

The unavoidable fact is after 50 years of attempting to automate sleep scoring and 27 years after my previous call to action, we are, despite the massive increase in computer sophistication, the myriad of publications and marketing hype, no closer to producing an accurate automatic staging algorithm. The simple fact is that this will never happen as long as we try to get computers to replicate R&K/AASM. We must stop accepting "around 85%" accuracy, against a false standard, ie the arbitrarily defined accuracy of human stagers, as good enough. If we, as a field, continue to believe that a model "achieves excellent performance" when it shows accuracy against manual staging using FP_z-C_z channel of 84.6% and using P_z-O_z channel of 82.3%[41] then we can and will never make progress on this vital issue.

Disclosure

The author reports no conflicts of interest in this work.

References

1. Rechtschaffen A, Kales A. *A Manual of Standardised Terminology. Techniques and Scoring System for Sleep Stages of Human Subjects*. Washington: Public Health Service, USA Government Printing Office; 1968.
2. Himanen SL, Hasan J. Limitations of Rechtschaffen and Kales. *Sleep Med Rev*. 2000;4(2):149–167. doi:10.1053/smrv.1999.0086
3. Jasper H. Report of the committee on methods of clinical examination in electroencephalography. *Electroencephalogr Clin Neurophysiol*. 1958;10:370–375. doi:10.1016/0013-4694(58)90053-1
4. Kubicki S, Hermann WM, Holler L. Critical comments on the rules by Rechtschaffen and Kales concerning the visual evalution of EEG sleep records. In: Kubicki S, Hermann WM, editors. *Methods of Sleep Research*. Stuttgart: Gustav Fischer; 1985:19–35.
5. Ferri R, Ferri P, Colognola RM, Petrella MA, Musumeci SA, Bergonzi P. Comparison between the results of an automatic and a visual scoring of sleep EEG recordings. *Sleep*. 1989;12(4):354–362. doi:10.1093/sleep/12.4.354
6. Dement W, Kleitman N. Cyclic variations in EEG during sleep and their relation to eye movements, body motility, and dreaming. *Electroencephalogr Clin Neurophysiol*. 1957;9(4):673–690. doi:10.1016/0013-4694(57)90088-3
7. Stanley, Neil. The future of sleep staging.*Hum Psychopharmacol*. 1996;11(3):253–256.
8. Iber C, Ancoli-Israel S, Chesson A, Quan SF. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology, and Technical Specification*. 1st ed. Westchester, IL: American Academy of Sleep Medicine; 2007.
9. Loomis AL, Harvey EN, Hobart G. Further observations on the potential rhythms of the cerebral cortex during sleep. *Science*. 1935;82 (2122):198–200. doi:10.1126/science.82.2122.198
10. Moser D, Anderer P, Gruber G, et al. Sleep classification according to AASM and Rechtschaffen & Kales: effects on sleep scoring parameters. *Sleep*. 2009;32(2):139–149. doi:10.1093/sleep/32.2.139
11. Danker-hopfe H, Anderer P, Zeitlhofer J, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *J Sleep Res*. 2009;18(1):74–84. doi:10.1111/j.1365-2869.2008.00700.x
12. Lee YJ, Lee JY, Cho JH, Choi JH. Inter-rater reliability of sleep stage scoring: a meta-analysis. *J Clin Sleep Med*. 2021;2021:193–202. doi:10.5664/jcsm.9538
13. Bandyopadhyay A, Goldstein C. Clinical applications of artificial intelligence in sleep medicine: a sleep clinician's perspective. *Sleep Breath*. 2022;9:1–7. doi:10.1007/s11325-022-02593-4
14. Stanus E, Lacroix B, Kerkhofs M, Mendlewicz J. Automated sleep scoring: a comparative reliability study of two algorithms. *Electroencephalogr Clin Neurophysiol*. 1987;66(4):448–456. doi:10.1016/0013-4694(87)90214-8

15. Klosh G, Kemp B, Penzel T, et al. The SIESTA project polygraphic and clinical database. *IEEE Eng Med Biol Mag*. 2001;20(3):51–57. doi:10.1109/51.932725
16. Penzel T, Glos M, Garcia C, Schoebel C, Fietze I. The SIESTA database and the SIESTA sleep analyser. Annual International Conference of the IEEE Engineering in Medicine and Biology Society; IEEE; 2011:8323–8326. doi:10.1109/iembs.2011.6092052
17. Kemp B, Värri A, Rosa AC, Nielsen KD, Gade J. A simple format for exchange of digitised polygraphic recordings. *Electroencephalogr Clin Neurophysiol*. 1992;82(5):391–393. doi:10.1016/0013-4694(92)90009-7
18. Silber MH, Ancoli-Israel S, Bonnet MH, et al. The visual scoring of sleep in adults. *J Clin Sleep Med*. 2007;3(02):121–131. doi:10.5664/jcsm.26814
19. Keenan S, Hirshkositz M. Sleep Stage Scoring. In: Kryger MH, Roth T, Goldstein CA, editors. *Kryger's Principles and Practice of Sleep Medicine*. Philadelphia: Elsevier Health Sciences; 2021:1839–1847.
20. Höller L, Riemer H. Comparison of visual analysis and automatic sleep stage scoring (Oxford Medilog 9000 System). *Eur Neurol*. 1986;25(Suppl. 2):36–45. doi:10.1159/000116080
21. Hoelscher TJ, Erwin CW, Marsh GR, Webb MD, Radtke RA, Anne L. Ambulatory sleep monitoring with the Oxford-Medilog 9000: technical acceptability, patient acceptance, and clinical indications. *Sleep*. 1987;10(6):606–607. doi:10.1093/sleep/10.6.606
22. Hoelscher TJ, McCall WV, Powell J, Marsh GR, Erwin CW. Two methods of scoring sleep with the Oxford Medilog 9000: comparison to conventional paper scoring. *Sleep*. 1989;12(2):133–139. doi:10.1093/sleep/12.2.133
23. Terzano MG, Parrino L. Clinical applications of cyclic alternating pattern. *Physiol Behav*. 1993;54(4):807–813. doi:10.1016/0031-9384(93)90096-X
24. Parrino L, Boselli M, Spaggiari MC, Smerieri A, Terzano MG. Cyclic alternating pattern (CAP) in normal sleep: polysomnographic parameters in different age groups. *Electroencephalogr Clin Neurophysiol*. 1998;107(6):439–450. doi:10.1016/S0013-4694(98)00108-4
25. Terzano MG, Parrino L, Sherieri A, et al. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. *Sleep Med*. 2001;2(6):537–553. doi:10.1016/S1389-9457(01)00149-6
26. Jobert M, Tismer C, Poiseau E, Schulz H. Wavelets—a new tool in sleep biosignal analysis. *J Sleep Res*. 1994;3(4):223–232. doi:10.1111/j.1365-2869.1994.tb00135.x
27. Oropesa E, Cycon HL, Jobert M. *Sleep Stage Classification Using Wavelet Transform and Neural Network*. Berkeley, California: International Computer Science Institute; 1999.
28. Pardey J, Roberts S, Tarassenko L, Stradling J. A new approach to the analysis of the human sleep/wakefulness continuum. *J Sleep Res*. 1996;5(4):201–210. doi:10.1111/j.1365-2869.1996.00201.x
29. Roberts S, Tarassenko L. New method of automated sleep quantification. *Med Biol Eng Comput*. 1992;30(5):509–517. doi:10.1007/BF02457830
30. Hirshkowitz M, Moore CA. Issues in computerised polysomnography. *Sleep*. 1994;17(2):105–112. doi:10.1093/sleep/17.2.105a
31. Hori T, Hayashi M, Morikawa T. Topographical EEG changes and the hypnagogic experience. In: Ogilvie RD, Harsh JR, editors. *Sleep Onset: Normal and Abnormal Processes*. Washington: American Psychological Association;1994:237–253. doi:10.1037/10166-000
32. Kan DP, Lee PF. Decrease alpha waves in depression: an electroencephalogram (EEG) study. In: 2015 International Conference on BioSignal Analysis, Processing and Systems (ICBAPS); IEEE; 2015:156–161. doi:10.1109/icbaps.2015.7292237
33. Schenck CH, Mahowald MW, Kim SW, O'Connor KA, Hurwitz TD. Prominent eye movements during NREM sleep and REM sleep behavior disorder associated with fluoxetine treatment of depression and obsessive-compulsive disorder. *Sleep*. 1992;15(3):226–235. doi:10.1093/sleep/15.3.226
34. Smith JR, Karacan I, Yang M. Ontogeny of delta activity during human sleep. *Electroencephalogr Clin Neurophysiol*. 1977;43(2):229–237. doi:10.1016/0013-4694(77)90130-4
35. Hauri P, Hawkins DR. Alpha-delta sleep. *Electroencephalogr Clin Neurophysiol*. 1973;34(3):233–237. doi:10.1016/0013-4694(73)90250-2
36. Smith JR, Karacan I. EEG sleep stage scoring by an automatic hybrid system. *Electroencephalogr Clin Neurophysiol*. 1971;31(3):231–237. doi:10.1016/0013-4694(71)90092-7
37. Gaillard JM, Tissot R. Principles of automatic analysis of sleep records with a hybrid system. *Comput Biomed Res*. 1973;6(1):1–3. doi:10.1016/0010-4809(73)90059-1
38. Fiorillo L, Puiatti A, Papandrea M, et al. Automated sleep scoring: a review of the latest approaches. *Sleep Med Rev*. 2019;48:101204. doi:10.1016/j.smrv.2019.07.007
39. Kubat M, Pfurtscheller G, Flotzinger D. AI-based approach to automatic sleep classification. *Biol Cybern*. 1994;70(5):443–448. doi:10.1007/BF00203237
40. Phan H, Mikkelsen K. Automatic sleep staging of EEG signals: recent development, challenges, and future directions. *Physiol Meas*. 2022;43:04TR01. doi:10.1088/1361-6579/ac6049
41. Huang J, Ren L, Zhou X, Yan K. An improved neural network based on SENet for sleep stage classification. *IEEE J Biomed Health Inform*. 2022;26(10):4948–4956. doi:10.1109/JBHI.2022.3157262