Computer-Assisted Sleep Staging

Rajeev Agarwal

Introduction

When assessing the polysomnographic recordings, it has become the norm to break the data into 20 or 30 second epochs which are subsequently labeled as one of the sleep stages defined by a scoring standard. For a full nocturnal recording, this can involve a large number of epochs that are typically on the order of 1500. Manual scoring of large number of epochs is time-consuming and plagued by significant inter-scorer subjectivity. With the advent of digital PSG/EEG recording and the available massive computing power, it is surprising that major advances in the automation of sleep staging have not come to see the light of day in routine clinical polysomnography. This is not because of shortage of research in the area, but rather due to the rules and standards with which sleep is scored (see below). The concept of automating sleep staging by computer analysis is not new, researchers were developing automatic sleep staging (ASS) algorithms as early as the late sixties of the last century. In the earlier years many researchers used the Dement and Kleitman scoring criteria in their auto staging techniques. Approaches based on period or spectral analysis, multiple discriminant analysis, decision trees or hybrid techniques that combine analog and digital processing, to name a few, were investigated by many researchers. Performance compared to visual scoring appeared to be reasonably good however, no consistent methodology of data analysis was employed for fair comparisons i.e., varying epoch sizes, selected data, dynamic threshold adjustment for different data sets, grouping of stages and different scoring criterion were employed in the different methods.

In the literature a number of computer methods using the Rechtschaffen and Kales scoring standard have been proposed. Gaillard and Tissot showed 83% agreement with visual scoring for 1-minute epochs with an hybrid system. Using pattern-recognition techniques, Martin et al. showed 83% agreement with visual scoring on 30-second epochs of five healthy subjects. Methods based on wave detection and Bayesian analysis, interval histogram, expert systems, neural networks, fuzzy clustering, among others have shown reasonable performance, however almost all are based on limited data of normal subjects. More recently, several schemes that use pattern recognition techniques, cardio-respiratory parameters, frequency based automatic sleep stage and segmentation and clustering and have shown renewed promise. A departure from the traditional epoch-based sleep staging has led to some works that describe the continuity in the evolution of the sleep state. This is more in concert with the idea that sleep is a continuum rather than discrete states that change from one state to another depending on depth of sleep.

The brief review above, while certainly not exhaustive, provides some indication of the richness of the literature in ASS algorithms. This interest has not been restricted to the academics as a number of systems have been commercially available. (see also and reference therein). However, none have been found to be acceptable in the routine clinical practice for various reasons. It would seem that if computer methods were available during the time where computers were so much less powerful, then given the vast resources of today automatic sleep staging should not be a great challenge. Why, then, are there no reliable automatic sleep staging methods available or more importantly why have they gotten such a bad rap over the years? In the following we will attempt to describe some reason why this may be so. We also present an overview of a recently presented semi-automatic staging method that attempts to incorporate the “natural” evolution of the sleep state.

Why Has Automatic Sleep Staging Been So Elusive?

Before we can understand why automation of sleep staging has been so difficult, it is fruitful to understand what is it that we are trying to automate. Since the introduction of the report by Rechtschaffen and Kales in 1968, sleep staging is almost universally done on the basis of this report and has come to be known as the RK staging or the RK rules. The PSG is generally divided into 20 or 30 second epochs which are then visually classified – termed sleep staging - into one of RK stages by a sleep technologist. The resulting time-evolutionary description of sleep in terms of changes of stages is known as the hypnogram and is used for clinical assessment of sleep architecture. The RK rules provide definition of each of the six sleep stages that have come to be widely accepted. These stages are called the non-REM stages 1 through 4, REM stage and the wake stage.

The great advantage of this scoring system is that it provides a language of communication that has come to be widely accepted. Two individuals reviewing a full night of sleep that has been staged with RK rules will come to similar conclusions about the sleep structure even though they are looking at a compressed view of the PSG.

Sleep is a non-uniform biological state that has been artificially divided into various “states” or “stages” based on polysomnographic ( PSG) measurements that include the
brain state (EEG) and other physiological measurements such as electrooculogram (EOG) and electromyogram (EMG). More appropriately, sleep in fact is a continuum of a state that goes from light sleep to deep sleep that is periodically interrupted by *wake-like* rapid-eye movement (REM) sleep. This evolving pattern of light sleep to deep sleep typically cycles several times during the course of a night. Current practice of characterization or the demarcation of sleep in terms of RK “states or stages” can be considered a methodological simplification that allows the standardized analysis across laboratories and reviewers.

The widespread recognition and usage of the RK standards is clearly a huge advantage that has stood the test of time, however, there are some clear disadvantages. An important drawback (probably the most important for automatic staging) is the subjectiv or imprecise definition of the rules. Many of the rules are defined with qualitative description of stages. For example, the RK rules state that “Stage 1 is defined by relatively low voltage, mixed frequency EEG ... The transition from an alpha record to stage 1 is characterized by a decrease in the amount, amplitude and frequency of alpha activity.” The use of words such as “relative” or “decrease” implies judgment or interpretation by the human scorer in assigning a stage to each epoch. Not all scorers will interpret this in exactly the same manner. This is one of primary sources of reduced concordance between two different scorers of the same record— the well-known inter-scorer differences. Inter-scorer reliability studies for data from healthy subjects have shown great variations. Visual scoring of PSGs from two health subjects across 10 laboratories in Japan showed a 67%–75.3% agreement.

RK sleep staging requires labeling each epoch as one of the six predefined stages. However, due to the continuum, the time of exact change of state is highly subjective and leaves significant room for inter-scorer as well as intra-scorer variations. Specifically, the transitional epochs (e.g., Stage 1 to Stage 2 or Stage 2 to stage 3, etc.), may be scored differently by different scorers. They are often scored differently by the same scorer on different occasions. Of course, in addition to being inherently subjective, visual/manual scoring of a full night of PSGs is a tedious and laborious task that easily lends itself to errors.

The RK rules were developed using data from young healthy college students and were intended to provide guidelines for normal adult sleep data. These same rules are often suitable to score the sleep for many pathological conditions. However, some departure from the rules is necessary in a number of pathologies. Clearly, ASS algorithms developed within the RK framework cannot be used in these instances, thus they are doomed for failure.

The above describes some of the fundamental difficulties that must be addressed by the researchers and developers of automatic sleep staging methods. The imprecise scoring rules must be translated to clear and precise rules for the computer programs. The encoding or the quantification of terms such as “relatively” or “decrease in amplitude” raises the obvious question how does one do this? How can amplitude be encoded? Should it be based on peak-to-peak amplitude or RMS amplitude? Most developers of ASS algorithms have attempted to interpret these measures in the spirit of the RK rules; however, there is no consensus on these criteria.

Given the above description, it is clear the level of agreement between automatic method and a human scorer is likely to be less that a 100%. Furthermore, comparison of automatic scoring against two independent manual scoring will not yield identical agreement – no two scorers are likely to agree 100% as discussed earlier. The best that we can expect with automatic scoring should be no better than the agreement between two individual manual scorers. Therefore, if two scorers agree on the average 85% to 90%, then our expectation from automatic scoring should be no better than, say 90%. Despite knowing that two scorers may not agree 100%, the sleep technologists still expect to get a 100% agreement with automatic scoring. In reality, the performance of automatic scoring should not and cannot have 100% agreement if the gold standard itself has significant discrepancies. This unrealistic demand from automatic staging schemes has been one of the main reasons for their lack of acceptance in the clinical settings.

Even though automatic staging has not found a common place in the clinical setting, there are still sufficient justifications for further development. One of the more important advantages of automatic scoring is the consistency of scoring. Because the scoring is based on fixed set of precise rules, every record will be scored objectively that can be repeated across laboratories.

We described above the idea of a continuum of a sleep state that has been artificially segmented into four non-REM sleep states (Stages 1-4) based on gradual emergence or changes in the EEG features. Because of this continuum, RK or any other staging scheme based on fixed number of stages is bound to be artificial and plagued by the transitional epochs. To minimize this effect, the following presents an overview of a recently presented new semi-automatic sleep staging method that attempts to identify pseudo-natural cluster or classes of PSG patterns (in the context of sleep related features) using self-organization techniques. By pseudo-natural, it is meant that the identified patterns are specific for the record under consideration whether it is normal or pathological. The patterns are not based on any *a priori* known or accepted notion. The approach is semi-automatic in the sense that the computer identified clusters of PSG patterns can be labeled as one of the predefined stages by the sleep technologist in accordance to his preference and experience. In this manner, the predefined stages can be one the RK stages or any other scoring standard of preference.

**Computer-Assisted Sleep Staging (CASS)**

Sleep staging is based on the premise that patterns repeat throughout the night. The differentiation between these patterns is based on generic measures and events such as amplitude and frequency, presence of spindles or eye movements, etc. The method we propose is based on the characterization of each epoch with these generic sleep related measures, the subsequent grouping of epochs having similar characteristics (“natural” groups of epochs), and finally using the input of the scorer in helping the system match these “natural” groups to the traditional R & K stages. The following summarizes the key steps:

1. **The PSG channels relevant to staging (EEG from central and occipital derivation, submental EMG and left and right EOGs) are simultaneously broken down into variable length epochs wherein the signals are relatively stationary or unchanging.** Over the course of the night this yields several thousand, 5-channel, variable-length segments of the PSG.
2. **Each segment is represented or encoded by a set of basic sleep-related features (amplitude, frequency, presence of spindles, and presence of REM, various spectral measures etc).**
3. By using the encoding in (2) the multi-channel segments of PSG are grouped into a maximum of 8-10 clusters using a self-organization scheme. This results in clusters of segments having similar characteristics, i.e., 8-10 classes of segment types.

4. Clusters of segments are translated into clusters of fixed-size staging epochs either 20 or 30 seconds. Conventional fixed-size staging epochs are grouped according to the dominant cluster type within the epoch. For example, if a 30s epoch has a 18s segment belonging to cluster A, a 8s segment belonging to cluster B and a 4s segment belonging to cluster C, the epoch is labeled as belonging to cluster A. Note that at this point we have not made any correspondence between the content of each cluster and sleep stages. We have simply grouped epochs according to “similar” segments.

5. The user is shown a few sample epochs from each cluster and asked to score these epochs according to the RK rules. Subsequently, these sample epochs are used by the computer to learn the user preference and score the remaining epoch in each cluster. We have, thus, incorporated the user preference in staging. It may be that two different clusters will be assigned to the same stage (for instance epochs with low amplitude EEG, low EMG, and rapid eye movements may be labeled REM, as may be a cluster of epochs with low EEG and low EMG but no rapid eye movements). Figure 1 is an example of an epoch not having all the characteristics of the REM stage, which the user nevertheless decides to classify as a REM epoch.

6. Post-processing is then performed to fine-tune sleep staging. It may for instance be necessary to reclassify an epoch because of the neighboring epochs, if the properties of the epoch allow such a reclassification. For example, a stage 3 epoch in the midst of a block of stage 2 epochs may be reclassified as stage 2 if the epoch properties warrant a change.

7. For reasons discussed earlier, it is rare that results of automatic methods can be used without visual review. The final step requires the visual review of the automatically generated hypnogram where the reviewer can change the score of any epoch with which she disagrees. This process yields a second hypnogram that we have termed “rescored” hypnogram.

Fig. 1. An epoch that has considerable EMG activity but user may still decide this is a REM epoch.
Fig. 2. Comparison between the hypnograms of a normal subject scored (A) manually, (B) using the presented computer-assisted methodology and (C) by a sleep technologist revising the computer-generated hypnogram (B), i.e. correcting epochs with which she did not agree. Altogether, 19.9% of epochs were corrected. All three hypnograms provide a similar picture of the sleep macro-structure.

Fig. 3. Illustration of the computer-assisted scoring of PSG of a subject with REM behavior disorder and mild alpha-intrusion by two different sleep technologists. (A) manually scored by a sleep technologist; (B) computer-assisted scoring by sleep technologist 1; (C) computer-assisted scoring by sleep technologist 2. Both computer-assisted hypnograms provide a similar sleep profile with an agreement of 80% that is on the same order as accepted manual inter-scorer agreements.

Results

Figure 2 shows an example of a night of sleep as originally scored by visual analysis (A) and by the computer-assisted system (B). Figure 1C shows the computer-assisted hypnogram that has been reviewed or “rescored” by visual analysis. The difference between B and C is what the operator felt needed correction after computer-assisted staging. It should be noted that in both the computer scored (B) and the rescored (C) hypnogram the overall profile of the sleep architecture (thick light gray line) remain very similar to that of the manual scoring (A). The level of agreement between A and B is 79.1% and it is 80.1% between B and C. Figure 3 is another example of a record with REM behavior disorder and mild alpha intrusion. This example illustrates the consistency of two different computer-scored hypnograms (by two different scorers). Figure 3A is the manually scored hypnogram, 3B is a computer scored hypnogram by scorer 1 and 3C is a computer scored hypnogram by scorer 2. Both computer-assisted hypnograms provide a similar sleep profile with an agreement of 80% that is on the same order as accepted manual inter-scorer agreements.
hypnogram by scorer 2. The level of agreement between the manual scoring and the computer scored hypnogram is 75.6% and 75.2%, respectively. However, the two computer-scored hypnograms have an agreement of 80%, which is on the same order as the expected inter-scorer agreement. In both examples, it is clear that a very similar impression of the night of sleep is reached from observation of all three hypnograms (thick light gray lines). In an evaluation study, it was determined that it took on average twice as long to manually score a hypnogram compared with using the computer-assisted sleep staging method followed by “rescoring” or correction. Clearly, there is an advantage in terms of efficiency for computer-assisted sleep staging. The final result is a hypnogram that is fully validated by a human operator.

Conclusion

Rechtschaffen and Kales state clearly in their manual that the rules are adequate at the time of their definition, but that they should not be cast in stone and should be adapted to new observations and new situations. However, after four decades of use, the RK rules have not gone through any changes. It is clear that in the current context of computer-based digital recording, the restriction to fixed size epochs and qualitative descriptions are an obstacle in exploiting the vast powers of computers. However, until a new set of scoring schemes based on more precise definitions can be created, we must continue to work within the context of RK. As described here, if we can understand what can be realistically expected of automatic staging, it is still possible to take advantage of computer power to do a significant part of the work of sleep staging. We have described the Computer-Assisted Sleep Staging (CASS) method that is not based on any fixed set of rules, but rather adapts to user preferences and the record under considerations. In approaches where user preferences can be incorporated within the scoring systems, it is possible to bridge the gap between complete automation and the necessary human interpretation for certain pathological conditions that cannot be handled by fixed rules. It is indeed possible to automate the sleep staging conundrum, however, one must have realistic expectations from a process where the gold standard is not clearly defined.

References