Certainty about Uncertainty in Sleep Staging

Citation for published version (APA):
van Gorp, H., Huijben, I. A. M., Fonseca, P., van Sloun, R. J. G., Overeem, S., & van Gilst, M. M. (2022). Certainty about Uncertainty in Sleep Staging: a Theoretical Framework. *Sleep*, *45*(8), Article zsac134. https://doi.org/10.1093/sleep/zsac134

Document license:
TAVERNE

DOI:
10.1093/sleep/zsac134

Document status and date:
Published: 11/08/2022

Document Version:
Typeset version in publisher’s lay-out, without final page, issue and volume numbers

SLEEP, 2022, 1–8

https://doi.org/10.1093/sleep/zsac134
Advance Access Publication Date: 8 June 2022
Perspective

Perspective

Certainty about uncertainty in sleep staging: a theoretical framework

Hans van Gorp[1,2,*], Iris A. M. Huijben[1,3], Pedro Fonseca[1,2], Ruud J. G. van Sloun[1,2], Sebastiaan Overeem[1,4] and Merel M. van Gilst[1,4]

[1]Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, the Netherlands, [2]Personal Health, Philips Research, Eindhoven, the Netherlands, [3]Onera Health, Eindhoven, the Netherlands and [4]Sleep Medicine Centre, Kempenhaeghe Foundation, Eindhoven, the Netherlands

*Corresponding author. Hans van Gorp, Department of Electrical Engineering, Eindhoven University of Technology, Room Flux 7.104, PO Box 513, Eindhoven, 5600 MB, the Netherlands. Email: h.v.gorp@tue.nl.

Abstract

Sleep stage classification is an important tool for the diagnosis of sleep disorders. Because sleep staging has such a high impact on clinical outcome, it is important that it is done reliably. However, it is known that uncertainty exists in both expert scorers and automated models. On average, the agreement between human scorers is only 82.6%. In this study, we provide a theoretical framework to facilitate discussion and further analyses of uncertainty in sleep staging. To this end, we introduce two variants of uncertainty, known from statistics and the machine learning community: aleatoric and epistemic uncertainty. We discuss what these types of uncertainties are, why the distinction is useful, where they arise from in sleep staging, and provide recommendations on how this framework can improve sleep staging in the future.

Statement of Significance

Aleatoric and epistemic uncertainty are used to address two challenges in the annotation of sleep structure: the relatively limited inter-rater agreement between sleep-scoring experts, and the apparent upper bound on accuracy for automated sleep-staging models. The distinction between these types of uncertainty has not been formally introduced to sleep staging but holds several advantages. The framework highlights where resources should be spent to reduce uncertainty and compares the pros and cons of human vs automated scoring. Most importantly, it allows for experts working on sleep staging to look beyond only inter-rater agreement and accuracy toward a more probabilistic outlook.

Key words: sleep staging; hypnogram; inter-rater agreement; machine learning; uncertainty; aleatoric; epistemic

Introduction

Since the publication of the Rechtschaffen and Kales (R&K) manual in 1968, sleep staging has become standardized [1]. In the R&K manual, a polysomnography (PSG) measurement is divided into segments, or epochs, of 30 seconds in length. A human scorer then visually classifies each epoch into one of the seven sleep stages: wake, stage S1–4, REM (rapid eye movement), and movement time. While the R&K standard was widely used for nearly 40 years, three important criticisms were raised: the R&K standard is based only on young and healthy adult participants,

Submitted: 16 February, 2022; **Revised:** 12 May, 2022

there is no clinical difference between S3 and S4, and the transition rules leave a lot open to subjective interpretation.

In response to these critiques, the American Academy of Sleep Medicine (AASM) manual was introduced in 2007 [2]. The AASM manual uses only five sleep stages: Wake (W), REM, and non-REM sleep (N1, N2, and N3). S3 and S4 from the R&K rules were merged into N3, and "movement time" was disregarded. Moreover, clearer rules regarding the transitioning between sleep stages were established, and recommendations for sampling rates and filter settings for the PSG were given.

With the introduction of the AASM manual, interscorer agreement increased [3, 4]. Currently, the overall agreement between expert scorers using the AASM criteria is around 82.6%, with a Cohen's kappa of 0.76 [5, 6]. The agreement of a single scorer against a panel can vary. For example, Stephansen et al. [7] analyzed the individual scoring of six scorers against the group consensus for the Inter-scorer Reliability Cohort [8]. The biased accuracy had a lower limit of 74.1% and an upper limit of 85.4% (with a mean of 81.3%), showing that some scorers tend to agree more with the group consensus than others.

Moreover, the interscorer agreement is not uniform across all sleep stages. For example, in the study performed by Rosenberg and van Hout, REM stage agreement was as high as 90.5%, while agreement on N1 was only 63.0%. Interestingly, no correlation was found between the amount of agreement in an epoch and whether or not said epoch was a transition epoch (i.e. least one of its neighbors was scored differently). Furthermore, a large group of varied scorers performed just as well as a small group of highly trained experts. Leading Rosenberg and van Hout to conclude that a basic understanding of the AASM rules is enough to produce competence, and a search for excellence may not be fruitful.

Next to human scorers, machine learning (ML) models that classify PSGs (or surrogate measurements) into sleep stages, have recently received great interest and many advancements have been made with regard to classification performance [7, 9–13]. Currently, the state-of-the-art accuracy compared to expert scoring is around 85% [14]. To arrive at this estimate, Fiorillo et al. reviewed 14 studies, the vast majority of which only compared to a single annotation per sleep recording. Only 2 out of 14 studies performed validation on multi-annotated data, including Stephansen et al. [7]. Comparing automated scoring systems against individual annotations has limitations due to the aforementioned interscorer disagreement and comparing to panels of scorers might be more robust to this effect [15, 16].

Moreover, it seems that ML models are reaching their upper limit in accuracy. For example, in a recent comparative study, it was found that different state-of-the-art models achieved similar performance on a pediatric sleep staging challenge, even making similar mistakes [17]. This begs the question, should a continued push for accuracy really be the main goal of automated sleep staging? If the inter-rater agreement between human scorers is only 82.6%, does it then make sense to strive for a higher and higher accuracy? We believe that, while accuracy is indeed very important, ML models should also explore the ambiguity in the way humans score a PSG.

In our opinion, both the limited inter-rater agreement of human scorers and the apparent upper limit on the accuracy of ML models can better be understood through the lens of uncertainty analysis. To that end, we shall introduce two variants of uncertainty known from statistics and the ML community: aleatoric and epistemic uncertainty [18, 19]. While epistemic uncertainty can be resolved through additional training or more diverse data collection, aleatoric uncertainty is inherent to a specific measurement setup and cannot be reduced. By splitting uncertainty in these two forms, we can explain why both humans will never reach 100% inter-rater agreement and ML models will never reach 100% accuracy. Because while they are both able to lower their epistemic uncertainty through additional training and expertise, their aleatoric uncertainty will never diminish.

This point of view may seem defeatist at first: simply blame aleatoric uncertainty for the residual inter-rater disagreement and for the imperfect accuracy of ML models, to then do nothing about it. However, we believe that this view on uncertainty in sleep staging is actually extremely helpful. It allows expert scorers and ML professionals to focus on the uncertainties that they have and what steps they can take to reduce them. For example, while aleatoric uncertainty cannot be reduced given a specific measurement setup, different measurement setups might result in lower aleatoric uncertainty.

The rest of the study will be structured as follows. First, we shall elaborate on the concepts of aleatoric and epistemic uncertainty. Second, we will explore the different sources of aleatoric uncertainty. Third, we will analyze epistemic uncertainty in both human scorers and ML model. Fourth, we will give recommendations about how this view of uncertainty can help improve sleep medicine in the future.

Uncertainty

For this study, we define uncertainty as "a state of limited information where it is impossible to exactly describe some phenomenon." Leveraging this definition for sleep stage classification, uncertainty refers to the lack of information, such that it is impossible to choose the correct sleep stage 100% of the time. In other words, there is ambiguity or variability in what sleep stage should be chosen.

We will refer to both the human scorer and automated classifier as a model. For the study, a model is a system or process that can be used to determine to which class the data belongs. This process can be described by decision boundaries or rules, for example, an epoch is scored as N3 if more than 20% consists of slow wave activity.

An example of a classification model is shown in Figure 1. It shows a binary classification problem, for example, wake vs sleep. In light blue and yellow, we show fictional data distributions of the two classes, which are generally unknown in practice. However, getting an idea of these distributions by estimating them, helps us in understanding and interpreting the recorded data. We show such estimation in dark blue and yellow, with the edges representing the decision boundaries for the corresponding classes. Situations may arise where decision boundaries are fuzzy, for example, in the green area, due to the presence of non-discriminative features or ambiguous classification rules. We can now see four different cases of (un)certainty emerging: (A) no uncertainty, (B) aleatoric uncertainty, (C) epistemic uncertainty, and (D) data point outside of the current task.

Aleatoric uncertainty (from the Latin *alea*, meaning chance or die) refers to uncertainty that arises from the random or

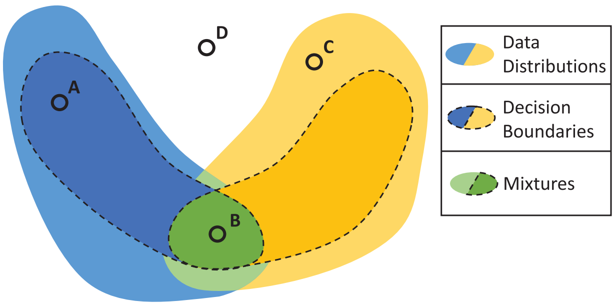

Figure 1. Aleatoric vs epistemic uncertainty. Four cases of uncertainty for a binary classification model, for example, wake vs sleep. Shown are the data distributions as well as the decision boundaries. (A) no uncertainty, (B) aleatoric uncertainty, (C) epistemic uncertainty, and (D) data point outside of the current task.

ambiguous nature of data and their measurements. Because of this, no model will be able to get rid of aleatoric uncertainty once a specific measurement has been performed. For example, in sleep, this happens when multiple features of different sleep stages are present in an epoch, and even the most trained experts disagree on proper classification. Moreover, artifacts and noise could hamper classification. This disagreement/uncertainty will not diminish through additional training of the experts, the ambiguity is simply inherent.

Note that aleatoric uncertainty is fixed once a specific measurement setup is chosen. As an example, a patient's age is introducing aleatoric uncertainty if the model does not have access to this information. On the other hand, when the model does take age into account (i.e. it is "measured" in the chosen setup), it is not a source of aleatoric uncertainty anymore.

Epistemic uncertainty (from the Greek *episteme*, meaning knowledge) refers to uncertainty that arises from a lack of knowledge about the data or the optimal model. This type of uncertainty can be mitigated, by supplying the model with more experience, additional training data, or in the case of ML by leveraging higher-capacity models. In sleep, epistemic uncertainty may arise when an inexperienced scorer tries to score a particularly difficult epoch. However, by consulting the AASM rule book or more senior colleagues, this (epistemic) uncertainty can be reduced.

While Figure 1 shows a binary classification example, we can extend this uncertainty view to 5-class sleep staging with both human and computer models. The four cases of (un)certainty, as introduced in Figure 1, are also shown in Figure 2 with a sleep staging example. Here we can see how two human scorers and one ML model might react under the different cases. For example, in the case of aleatoric uncertainty (B), the models might all disagree on the class, but they all do agree that it is either N2 or N3, and not say "Wake." Moreover, if the computer model has not been programmed to deal with faulty measurements (D), it will still provide an output, while humans will see that this does not make sense.

We would like to stress that aleatoric and epistemic uncertainty are distinct from systematic and random errors. Systematic and random error refer to (measurement) errors made due to, for example, wrongly calibrated equipment, while the aleatoric–epistemic uncertainty view considers the entire pipeline of sleep stage classification, from PSG measurement to human decision making. Under this lens, random and systemic errors are parts of aleatoric uncertainty.

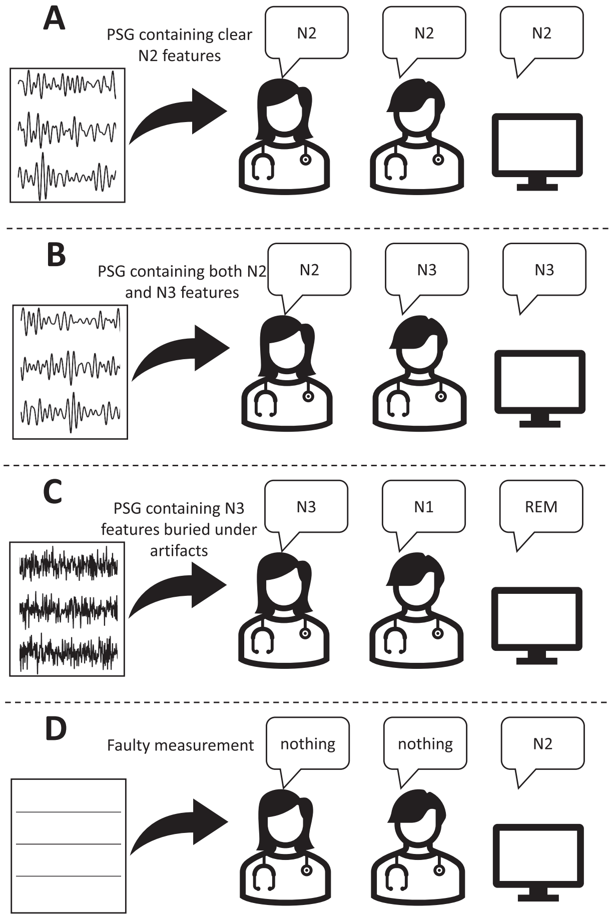
A
PSG containing clear N2 features
N2
N2
N2
B
PSG containing both N2 and N3 features
N2
N3
N3
C
PSG containing N3 features buried under artifacts
N3
N1
REM
D
Faulty measurement
nothing
nothing
N2

Figure 2. Examples of uncertainty in sleep. Examples of the different cases of uncertainty in sleep shown here are how two hypothetical human scorers and one machine learning model would score the PSG data. (A) no uncertainty, (B) high aleatoric uncertainty, (C) epistemic uncertainty, and (D) data point outside of the current task.

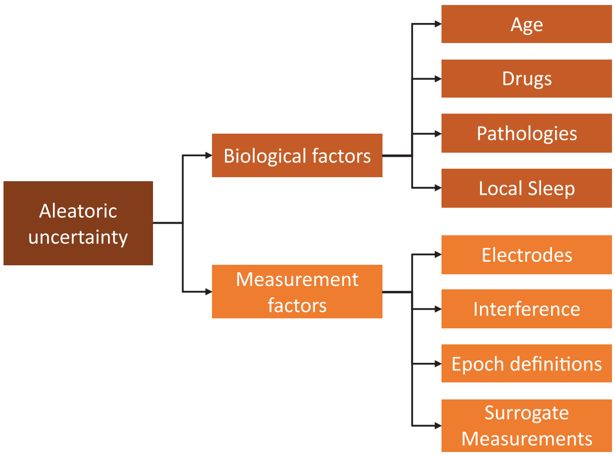

Figure 3. Sources of aleatoric uncertainty in sleep split into biological factors (i.e. relating to the participant themselves) and measurement factors (i.e. relating to the way we measure sleep).

Sources of Aleatoric Uncertainty

Several sources of aleatoric uncertainty in sleep staging can be identified as shown in Figure 3. Note that the type of model (human or ML) has no impact on aleatoric uncertainty, it is entirely inherent to the biological process of sleep (see section “Biological factors”) and the way that it is measured (see section “Measurement factors”).

Biological factors

In this section, we show how several biological factors will affect the measurement and scoring of sleep stages. As we alluded to in the previous section these factors induce aleatoric uncertainty if the model is not leveraging measurements of these factors. If the (only) measurement we perform is a PSG, then the model cannot control for these factors. However, if we perform additional measurements, such as asking somebody’s age or medicine usage a model may adjust for these factors while scoring. Possibly remaining uncertainties in sleep staging, related to these factors, have then become part of epistemic uncertainty.

Age. The age of the participant has an important effect on the structure of electroencephalogram (EEG) signals. Notably, as the neural connectivity decreases with age, it is less likely that neurons will fire synchronously. Because an EEG measurement never records single neurons, but always the aggregate behavior, the more out of phase the neurons fire, the lower the amplitude of the waves that are measured. For example, slow wave amplitudes are much lower in older participants than they are in younger participants which can lead to disagreement on how to apply the rules for N3 sleep in participants across varying age groups.

Drugs. Drugs that act on the central nervous system (CNS) (e.g. alcohol, opioids, sleep medication, antidepressants, and antipsychotics) will affect the PSG measurement. Due to this, features of sleep might present differently or not at all.

Pathologies. Similar to drugs, certain pathologies will affect the workings of the CNS or the connection from the CNS to other responses, for example, heart rate. Think, for example, of Parkinson’s disease or epilepsy. Moreover, brain lesions or trauma can also affect the PSG recording. Because pathologies can fog the features of sleep in the PSG, correct application of the scoring rules may be hampered.

Local sleep. Recently, the concept of “local sleep” has emerged [20–22], where instead of thinking about sleep as being a homogeneous process across the entire brain, different brain regions can be in different sleep stages. Contradictory features of different sleep stages can therefore be observed at the same time, making it hard to identify the correct sleep stage. One can even ponder the question if there is indeed such a thing as “one correct sleep stage” over the entire brain.

Measurement factors

Electrodes. The placing of electrodes is crucial to acquire the right information to do sleep staging. For example, alpha activity is typically higher in the occipital measurements, whereas delta activity is higher in the frontal electrodes. In a perfect world, we would therefore measure the EEG over the entire scalp, but we can only use a limited number of electrodes at the same time. To accommodate the placement of electrodes, the AASM manual specifies which standard locations to use. However, sleep staging can still be performed with single lead measurements. In this case, however, crucial information can be lost about the EEG activity on other regions of the scalp.

Because the electrodes are the interface between skin and wires, and the amplitude of EEG waves is so low, it is of critical importance that a good conductivity is reached, which is often done with the application of gels. The type, quality, and amount of the gel can have big consequences for the quality of the measured signals.

Interference. Several sources of interference and noise can be present in a PSG measurement, for example, mains interference, electrical noise interference, and baseline wander of the electrodes. To counteract this noise, the AASM specifies the use of a couple of filters. Notably, a high-pass filter at 0.3 Hz and a low-pass filter at 35 Hz for EEG, which do not act on the frequencies of interest (0.5–20 Hz).

Nevertheless, additional artifacts and noise sources can be present, for example, movement artifacts and electrical noise interference. Because the electrical response of movement artifacts is much larger than the EEG, it will bury the signal and make it impossible to score. Electrical noise interference is caused by other nearby wires, as they act like antennas and get electrically linked with each other. Even when filtering out this electrical noise interference, it can still have a negative impact on the signal quality, and therefore hamper correct scoring.

Epoch definitions. According to the AASM standard, an entire epoch is scored as one sleep stage. Conventionally, this epoch is 30 seconds long, but other window lengths can be used in ML models. The transitioning of one sleep stage to another is, however, unlikely to always happen exactly at the boundaries of 30 seconds windows (if such “hard” transitioning indeed occurs). By scoring an entire epoch as one sleep stage, we are thus aggregating features of (potentially) multiple sleep stages. Additionally, not only the epoch length matters, but where we define the first boundary. A constant offset to all boundaries by, for example, 10 seconds can result in a different scoring. Moreover, if a sleep stage is significantly shorter than the window length, we can miss it (or must ignore it according to AASM) in our analysis [23]. We believe that the aggregation of sleep stages over an epoch contributes to a significant source of aleatoric uncertainty.

Surrogate measurements. While the clinical standard for quantitatively measuring sleep is still considered a PSG, it is nowadays also possible to use surrogate measurement techniques (e.g. photoplethysmography (PPG) or actigraphy). While impossible for a human to score, ML models can be used to acquire hypnograms from these surrogates [24–26]. This is done by teaching the model to mimic the AASM scoring on a synchronously acquired PSG.

The reason that surrogate measurements can be used for sleep scoring is because there is a correlation between sleep stages and the autonomic activity of, for example, the heart. However, the exact mapping from the CNS to autonomic responses is unknown and potentially participant-dependent. All the biological factors mentioned above can have a detrimental impact on the mapping between surrogate measurements and sleep stages. For example, older participants will show a decrease in the autonomic expression of parasympathetic activity. Moreover, some medications blunt cardiac expression of autonomic activity, and certain disorders will also sever this link. Due to fact that it can be so hard to quantify what the link is between autonomic activity and sleep stages and also to quantify the quality of this connection, one can expect to have lower accuracies and higher uncertainties when using surrogate measurements.

Sources of Epistemic Uncertainty

Several sources of epistemic uncertainty in sleep staging can be identified as shown in Figure 4. In this section, we shall discuss each source, and how it relates to both human scorers and ML models.

Relation to prior belief/context

A human scorer is (implicitly) biased by its prior beliefs when assigning sleep stages, for example, due to the demographics of the patient, expected disorder(s), or at what point in the night the data was recorded. The prior in ML models is often implicit in the parameters and architecture, making it difficult for designers to know whether the (learned) prior is correct. A well-known example of a prior of ML models is the tendency to favor classification for N2 sleep, as that sleep stage is the most common during the night and therefore over-represented in the dataset used for training the model.

Quantifying the effects of prior beliefs can be difficult, both in human scorers and ML models. It is therefore important to note what type of patients a human scorer typically sees, or on which dataset an ML model has been trained. The prior belief over sleep stages may be invalid for a patient that does not fit the group of patients on which the model was trained. Additional training or retraining could be a solution for when a model encounters a new type of patient. For example, both human and ML models can be retrained for the specific hospital at which they are used, to ensure that it has the correct prior beliefs for the patients' group(s) typically seen there.

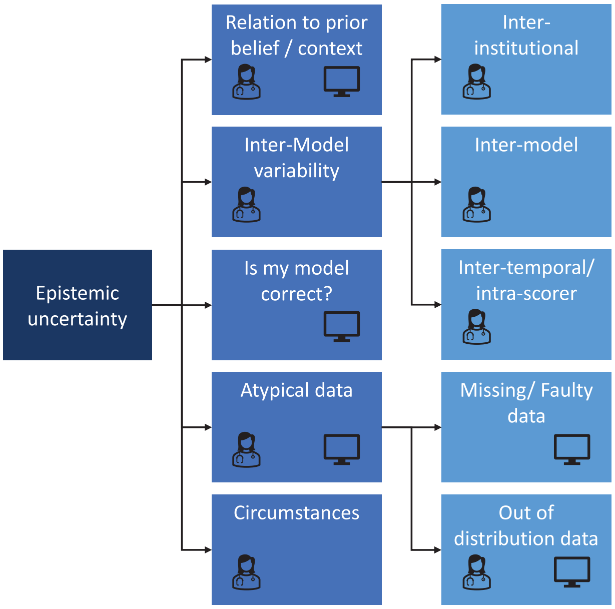

Figure 4. List of sources of epistemic uncertainty in sleep. Icons of a human scorer or computer model are shown for each of the sources when they are relevant, for example, "relation to prior belief/context" applies to both, while "circumstances" only apply to humans.

Inter-model variability

Another source of epistemic uncertainty is the fact that no two models are the same. It is important to stress here is that we consider a model at one fixed point in time, thus even inter-temporal change can cause epistemic uncertainty. We distinguish three inter-model variabilities, from macro to micro levels.

Inter-institutional variability. The culture of an institution can have a big impact on how the AASM rules are interpreted and applied by human scorers. For example, one hospital might interpret the AASM rules a lot more strictly than another one. The effect of culture and workplace environment could be alleviated through programs, such as the Inter-Scorer Reliability Assessment System, through which the scoring of different institutions is brought closer together.

Inter-model variability. Not only can there be differences among institutions, but individuals within those institutions can also differ. This reflects someone's scoring style; their own signature in sleep staging. By comparing work with colleagues and aligning workflows, these differences can be alleviated.

Inter-temporal variability. Over time, models can also change. If you show the same PSG to a human scorer today and tomorrow, you will probably get a different scoring. Moreover, larger differences can appear if you increase the time difference between the moments of scoring.

Model correctness

The question of model correctness for human scorers comes down whether they have learned and implemented the AASM rules correctly. As shown by Rosenberg and van Hout [5], a basic understanding of the AASM rules is already enough to produce competence, and additional training may not be fruitful. In other words, human scorers are correct most of the time, and this source of epistemic uncertainty does not really exist for them.

The problem of model correctness or optimal model choice is much more a challenge in ML. We have to ask ourselves, out of the infinite set of models out there, have I chosen the correct one, or at least a good one? A systematized way to search for the best model would be neural architecture search [27]. However, to the best of our knowledge, neural architecture search has so far never been used for automatic sleep stage classification models. Even the question: Is my model at least "good enough?" can be difficult to answer for computer models. As ML models can easily fall for pitfalls such as overfitting and shortcut learning. Moreover, we can only empirically test if our model generalizes to an unseen dataset, but we can never prove it for all unseen data. Lastly, even the definition of "good enough" is ambiguous.

Atypical data

Missing/faulty data. While it is easy for a human to detect when a measurement has gone wrong, for example, a disconnected electrode, this can be difficult for an automated model to do. Especially if no special care is taken to give the automated model the ability to say "I do not know" or "Something went wrong"; in these cases, the automated model will still give us an answer even if it should not.

Out-of-distribution data. Much harder for both humans and ML models is the ability to detect out-of-distribution (OOD) data. With OOD data, we here refer to measurements that are unlike any the model has seen before. For a human, this can encompass data that is so different from what they usually see, and where they do not know how best to apply the AASM rules, while for an ML model, this entails data that differs too much from the training data. It is critical that the model is able to recognize when data are OOD, so that they may ask for a second opinion. This way, they could even learn from the example and reduce their epistemic uncertainty the next time that they encounter this type of data.

Circumstances

Circumstances, such as the mental and physical well-being of the expert scorer, can have a detrimental effect on performance. While ML models are not affected by such factors, human performance will deteriorate due to, for example, sleepiness, stress, and time pressure. To reduce this source of epistemic uncertainty, it is important that scorers can do their work without (too much) time pressure or stress placed upon them.

Discussion

We have reviewed two types of uncertainty in sleep staging: aleatoric and epistemic uncertainty. While the aleatoric sources of uncertainty are measurement setup-dependent, but model-independent, the epistemic sources are model-dependent, and thus differ between human scorers and ML models. This study outlines the different sources of uncertainty in sleep staging and how they may be alleviated.

As a consequence of aleatoric uncertainty, human scorers will never reach 100% inter-rater agreement, and ML models will never reach 100% accuracy with respect to a set of expert labels. It, therefore, does not make sense to solely focus on and promote these metrics. Instead, it may be more fruitful to analyze and capture existing uncertainties in sleep staging, and where they come from. If the sources of uncertainty are well understood, it becomes easier to estimate how (un)certain a given sleep scoring is. Such uncertainty estimates can greatly increase the confidence of the end-user in the hypnogram, especially in the case of automated scoring. In addition, uncertainty estimates in sleep diagnostics could be extended to other outcomes of PSG besides sleep staging, for example, applied to apnea-hypopnea index (AHI) estimates. In the Supplementary Material, we briefly review what work has already been done and provide some future perspectives for both expert scorers and automated models.

All current studies of the inter-rater agreement only investigate the agreement on an individual level between scorers, or between the individual and the majority vote of the entire group of scorers. To the best of our knowledge, no studies exist that explore the inter-rater agreement between panels of scorers, even though, for example, the AASM gold standard for scoring is based on a consensus formed by a panel of expert scorers. It would be an interesting opportunity for future research to investigate the inter-panel agreement. Intuitively, one would expect that agreement between panels would be higher, as a group of scorers can collectively lower their epistemic uncertainty.

Furthermore, because of the substantial inter-rater disagreement between human scorers, it is greatly beneficial to compare automated scoring methods to panels of scorers, and not just single scorers. One possible approach to do this is to compare the prediction of the automated model to the majority vote of a panel of human scorers, as is done in the studies [7] and [15]. To supplement such a comparison, the authors of the study [7] also compare each scorer individually against the majority vote of the panel. This way, they establish that their automated model agrees with the majority vote more, than each separate scorer agrees with the majority vote. The authors of the study [15] apply such an approach also to statistics calculated from sleep staging, such as AHI. However, the authors of the study [16] raise the criticism that such averaging is limited because the distribution of these statistics can be skewed and suffer from outliers. Instead, the authors of the study [16] model how each statistic is distributed and then apply statistical testing to see if the automated predictions come from the same population.

Recently, advances in Bayesian ML have enabled for one model to yield multiple likely predictions, instead of just one point estimate (see the Supplementary Material). However, none of the aforementioned methods delves into how to compare multiple automated predictions against the multiple ground truths of a panel. Such comparisons would be very interesting, and should not just involve comparing majority votes. While comparing majority votes gives insight into the performance of point estimates, it does not establish whether the model has captured all scoring "styles" present in the dataset, and whether the aleatoric uncertainty of statistics such as the AHI is properly estimated. We will leave the question of how to compare multiple automated predictions to that of a panel to future research.

Throughout this study, we have made the assumption that the current sleep stages as defined by the AASM rulebook are in fact the ground truth. However, alternative sleep representations have also been proposed. For example, the odds ratio product [28], data-driven clusters, or continuous representations (see Hermans et al. [29] for an overview). These alternative sleep staging methods are promising in the regard that they might prove to be better representations of the structure of sleep and provide physicians with more information. However, additional research is needed to prove the clinical application of these methods. Moreover, interpretability, especially for data-driven clustering, remains a concern. This interpretability issue is often resolved by comparing the new sleep stages with the original AASM sleep stages. The main message of this study: the utility of the distinction between aleatoric and epistemic uncertainty stays applicable to these alternative staging methods. It is beneficial to explore different models working on the same data, as this lies at the heart of reducing epistemic uncertainty in sleep staging.

Conclusion

In this study, we offer a perspective on sleep stage classification through the lens of aleatoric and epistemic uncertainty analysis. We discussed the sources of these two types of uncertainty for both human scorers and automated models. The sources of aleatoric uncertainty do not depend on the choice of model, as it is inherent to the measurement and patient. On the other hand, the sources of epistemic uncertainty do differ between human scorers and automated models, but they can still be united into broad categories, allowing us to see the pros and cons of both types of scorers. Moreover, we recommend that instead of trying to fruitlessly achieve the highest possible accuracy or increase inter-rater agreement through additional training, we should create tools that accurately reflect the model's uncertainty. If well calibrated, the models of the future could accurately split aleatoric and epistemic uncertainty, giving us certainty about uncertainty in sleep staging.

Supplementary Material

Supplementary material is available at *SLEEP* online.

Funding

At the time of writing, H.G., P.F., and R.J.G.S. were employed and/or affiliated with Royal Philips, a commercial company and manufacturer of consumer and medical electronic devices, commercializing products in the area of sleep diagnostics and sleep therapy. Philips had no role in the study design, decision to publish, or preparation of the manuscript. P.F. reports personal fees from Philips Research during the conduct of the study; personal fees from Philips Research, outside the submitted work. This work was performed within the IMPULSE framework of the Eindhoven MedTech Innovation Center (e/MTIC, incorporating Eindhoven University of Technology, Philips Research, and Sleep Medicine Center, Kempenhaeghe Foundation), including a PPS supplement from the Dutch Ministry of Economic Affairs and Climate Policy. Additionally, the work of I.A.M.H. was supported by Onera Health and the project "OP-SLEEP". The project "OP-SLEEP" is made possible by the European Regional Development Fund, in the context of OPZuid.

Disclosure Statement

S.O. received an unrestricted researchgrant from UCB Pharma and participated in advisory boards forUCB Pharma, Jazz Pharmaceuticals, and Bioproject, all paid to theinstitution and all unrelated to the present work.

References

1. Rechtschaffen A, Kales A, eds. *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. Washington, D.C.: Government Printing Office, 1968. (NIH publication no. 204.)
2. Iber C, Ancoli-Israel S, Chesson AL Jr, Quan SF; for the American Academy of Sleep Medicine. *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*. 1st ed. Westchester, IL: American Academy of Sleep Medicine, 2007.
3. Danker-Hopfe H, *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.
4. Ruehland W, *et al*. The 2007 AASM recommendations for EEG electrode placement in polysomnography: impact on sleep and cortical arousal scoring. *Sleep*. 2011;**34**(1):73–81. doi:10.1093/sleep/34.1.73
5. Rosenberg R, *et al*. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. *J Clin Sleep Med*. 2013;**9**(1):81–87.
6. Lee YJ, *et al*. Inter-rater reliability of sleep stage scoring: a meta-analysis. *J Clin Sleep Med*. 2021;**18**(1):193–202.
7. Stephansen J, *et al*. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nat Commun*. 2018;**9**(1):5229.
8. Kuna ST, *et al*. Agreement in computer-assisted manual scoring of polysomnograms across sleep centers. *Sleep*. 2013;**36**(4):583–589. doi:10.5665/sleep.2550
9. Chriskos P, *et al*. Automatic sleep staging employing convolutional neural networks and cortical connectivity images. *IEEE Trans Neural Networks Learn Syst*. 2020;**31**(1):113–123.
10. Mousavi S, *et al*. SleepEEGNet: automated sleep stage scoring with sequence to sequence deep learning approach. *PLoS One*. 2019;**14**(5):e0216456.
11. Michielli N, *et al*. Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals. *Comput Biol Med*. 2019;**106**:71–81.
12. Yildirim O, *et al*. A deep learning model for automated sleep stages classification using PSG signals. *Int J Environ Res Public Health*. 2019;**16**(4):599.
13. Phan H, *et al*. XSleepNet: multi-view sequential model for automatic sleep staging. *IEEE Trans Pattern Anal Mach Intell*. 2021:1.
14. Fiorillo L, *et al*. Automated sleep scoring: a review of the latest approaches. *Sleep Med Rev*. 2019;**48**:101204.
15. Malhotra A, *et al*. Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. *Sleep*. 2013;**36**(4):573–582. doi:10.5665/sleep.2548
16. Punjabi NM, *et al*. Computer-assisted automated scoring of polysomnograms using the somnolyzer system. *Sleep*. 2015;**38**(10):1555–1566. doi:10.5665/sleep.5046
17. Phan H, *et al*. Pediatric automatic sleep staging: a comparative study of state-of-the-art deep learning methods. *IEEE Trans Biomed Eng*. 2022.
18. Hüllermeier E, *et al*. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach Learn*. 2021;**110**(3):457–506.
19. Indrayan A. Aleatory and epistemic uncertainties can completely derail medical research results. *J Postgrad Med*. 2020;**66**(2):94–98.
20. Emrick J, *et al*. Different simultaneous sleep states in the hippocampus and neocortex. *Sleep*. 2016;**39**(12):2201–2209. doi:10.5665/sleep.6326
21. Krueger J, *et al*. Local sleep. *Sleep Med Rev*. 2019;**43**:14–21.
22. Stålesen Ramfjord L, *et al*. Local sleep and wakefulness—the concept and its potential for the understanding and treatment of insomnia disorder. *Somnologie*. 2020;**24**(2):116–120.
23. Perslev M, *et al*. U-Sleep: resilient high-frequency sleep staging. *NPJ Digital Med*. 2021;**4**(1):72.

24. Fonseca P, *et al*. Validation of photoplethysmography-based sleep staging compared with polysomnography in healthy middle-aged adults. *Sleep*. 2017;**40**(7). doi:10.1093/sleep/zsx097
25. Korkalainen H, *et al*. Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea. *Sleep*. 2020;**43**(11). doi:10.1093/sleep/zsaa098
26. Imtiaz S. A systematic review of sensing technologies for wearable sleep staging. *Sensors*. 2021;**21**(5):1562.
27. Elsken T, *et al*. Neural architecture search: a survey. *J Mach Learn Res*. 2019;**20**(1):1–21.
28. Younes M, *et al*. Odds ratio product of sleep EEG as a continuous measure of sleep state. *Sleep*. 2015;**38**(4):641–654. doi:10.5665/sleep.4588
29. Hermans LW, *et al*. Representations of temporal sleep dynamics: review and synthesis of the literature. *Sleep Med Rev*. 2022;**63**:101611.