Predicting duration of sleep from the three process model of regulation of alertness

Torbjörn Åkerstedt, Simon Folkard

Abstract

Objectives—Irregular working hours severely disturb sleep and wakefulness. This paper presents a modification of the quantitative (computerised) three process model of regulation of alertness to predict duration of sleep in connection with irregular sleep patterns.

Methods—The model uses a circadian “C” (sinusoidal) and homeostatic “S” (exponential) component (the duration of previous periods awake and asleep), which are summed to yield predicted alertness (on a scale of 1–16). It assumes that waking from sleep will occur at a given alertness level (S’ + C) when recuperation is complete. Variables of electroencephalographic duration of sleep from two studies of irregular sleep were used to model the S and C variables in a regression approach to maximise prediction. The model performance was cross validated against published field and laboratory data.

Results—The model parameters were defined with a high degree of precision R² = 0.99 and the validation yielded similar values R² = 0.98–0.95, depending on the acrophase. The paper also describes a simplified graphical version of the computation model seen as a two dimensional duration of sleep nomogram.

Conclusion—The model seems to predict group means for duration of sleep with high precision and may serve as a tool for evaluating work and rest schedules to reduce risks of sleep disturbances.

Keywords: sleep; prediction; shift work

Disturbance of sleeping and waking is a serious problem for people who have to displace sleep or work unconventional hours. Severe subjective sleepiness and sleep related electroencephalographic changes characterise, for example, night working lorry drivers1 and train drivers.2 In process operators bouts of sleep actually occur during work.3 Also, duration of sleep is reduced on night and morning shifts4 and sleep latency is increased before the morning shift5 and during daytime and early evening in general.6 As may be expected, sleepiness at work leads to accidents, including major events such as nuclear disasters.8

To explain and predict variations in alertness related to irregular sleeping patterns we have constructed a quantitative model that uses the times of sleep and waking to predict alertness.9 The work was inspired by the two process model of sleep regulation, which had shown that duration of sleep and slow wave activity could be described by a combination of a homeostatic (previous time awake and asleep) and a circadian influence.10 With subjective alertness data from several experiments of altered sleep patterns we found that alertness was predictable from a circadian and a homeostatic component—plus a component for sleep inertia (described later). The output of the model has been validated against subjective ratings, performance, and electroencephalographic and electrooculographic measures of sleepiness, and has shown considerable accuracy.9,11 It has also been shown to predict sleep latency.12

The time of falling asleep and waking immediately before the interval to be predicted were entered into the model. The model then uses the time asleep, the time since waking, and the time of day to compute predicted sleepiness. Clearly, it would greatly enhance the usefulness of the model if a key input variable such as duration of sleep could also be included in the prediction. As sleep latency is also predictable, the only major component that would need to be included would be the decision to go to bed. The practical implications of such a modification of the model may be considerable, particularly in relation to scheduling and planning of work and rest patterns.

Admittedly, a model that predicts duration of sleep already exists in the two process model of sleep regulation.9,13 However, that model has not been used to predict electroencephalographic measurements of duration of sleep in applied situations and cannot be easily integrated into our alertness model as it is based on variables such as temperature and electroencephalographic slow wave activity. Also, it would be conceptually elegant if the end of sleep were found to be closely linked to the level of alertness during sleep. Thus, we think there is a need for a practical predictor of duration of sleep that may be integrated with predictions of alertness and sleep latency. The purpose of this paper was to modify the three process model of regulation of alertness to predict duration of sleep and to validate the prediction from the actual data from shift work.
Predicting duration of sleep from the three process model of regulation of alertness

Method

As already mentioned, the model was derived with subjective alertness data from several experiments of altered sleep patterns. Modelling these data we found that alertness was, indeed, predictable from three parameters: S, C, and W. Parameter C represents sleepiness due to circadian influences and has a sinusoidal form with an afternoon peak (fig 1). Parameter S is an exponential function of the time since waking, is high on waking, then falls rapidly initially and gradually approaches a lower asymptote. At the start of sleep S is reversed and called S' and recovery occurs in an exponential fashion that initially increases very rapidly but subsequently levels off towards an upper asymptote. Total recovery is usually accomplished in eight hours. The final component (not in fig 1) is the wake up process W, or sleep inertia. This component is not part of the present modelling and will not be further discussed. The times of rising and going to bed for the period investigated were put into the model.

The predicted alertness is expressed as the arithmetic sum of the two (W presently excluded) functions above (S + C, fig 1). The scale of the model ranges normally from 1–16, but in practice 3 corresponds to extreme sleepiness and 14 to high alertness and 7 to a sleepiness threshold based on electroencephalographic alpha activity and electrooculographic slow eye movements. In figure 1 S + C shows predicted alertness when wakefulness is extended by eight hours (to 24 hours), as is frequently the case with a first night shift. This particular prediction assumes that waking occurs at 07 00 in the morning after an eight hour sleep, and thereafter no sleep occurs (due to the night shift) until 07 00 the next day. The combined effect of S + C (long time awake and the circadian down swing) yields a fall of alertness during the night, with a trough in the early morning. After sleep starts, the sleep recovery of factor S, together with the circadian up swing causes a rapid increase in (latent) alertness during sleep.

To model the end of sleep we assumed that waking from sleep will occur whenever sufficient alertness is reached—that is, when S' + C' reaches a critical end of sleep level (C' = C during sleep, fig 2). We, furthermore, assumed that this critical end of sleep level should be equal or close to the asymptote of S'—the level that homeostatic alertness approaches, without circadian interference. As, however, an asymptote cannot, by definition, be reached, we assumed the critical end of sleep level was 0·1 unit below the asymptote. Figure 2 shows the principle of calculating the end of sleep. Parameter C' has been drawn tangential to the end of sleep level. Note that C' is drawn inverted, hanging from the end of sleep level. Thus, waking occurs whenever S' reaches C—that is, when S' + C' = end of sleep level. In figure 2, “a” shows the very short sleep predicted for a bedtime at 11 00 after rising at 07 00. Index “b” indicates a sleep initiated at 23 00 after 16 hours awake. As C' rapidly withts after the afternoon peak, S' will not intersect C' until after the trough is reached and C' increases again. Index “c” indicates the start of sleep at 11 00 after 28 hours of wakefulness and index “d” a sleep at 19 00 after 36 hours of wakefulness.

Initial tests of the model indicated that it strongly exaggerated sleep truncation during the daytime—that is, the circadian component was far too strong. To obtain optimum parameters we therefore modelled actual data with different C' values. Data from two previous laboratory experiments of sleep manipulation with subsequent uninterrupted sleep were used. Both were carried out under isolation from environmental time cues and with strict control of activity (sedentary) and food intake. In both studies sleep was recorded with portable equipment with electroencephalographic electrodes CzOz/02P4. Sleep was scored conventionally.

The first experiment was the same as was
used for the original experiment on the end of sleep in the Borbely-Daan model. It involved going to bed at 23 00, or delaying bedtime to: 03 00, 07 00, 11 00, 15 00, 19 00, and 23 00. One week of "wash out" was scheduled between the parts of the experiment. Note that time awake increased progressively. The subjects (eight men, 29–45 years old) rose from sleep at 07 00 after an eight hour sleep, reported to the sleep lab at 19 00, had electrodes applied, started the study at 22 00 and continued with waking activity until the scheduled bedtime. The effect on duration of sleep was highly significant (repeated measures ANOVA). Sleep gradually shortened from 23 00 to 11 00, and then started to increase to a maximum at 19 00 (fig 3).

The second experiment involved eight men (age 20–47) and four experimental parts. It lasted for 28 hours, from 07 00 to 11 00 the next day. The subjects had either no intervening night sleep or eight hours of sleep (23 00–07 00), four hours of sleep (03 00–07 00), two hours of sleep (05 00–07 00), or no sleep at all. The no sleep condition was not used as it replicated one of the parts in the previous study. The sleep before the experiment was a normal baseline sleep between 23 00 and 07 00. The effect on duration of sleep was highly significant (repeated measures ANOVA). Duration of sleep thus ranged from two hours (after a full night’s sleep) to 4.5 hours (after no night sleep).

The end of sleep was estimated to occur at the point at which S’ + C reached the adjusted asymptote of S’ (= 14.2). In a series of analyses to predict duration of sleep, the amplitude of C was varied systematically from 2.5 to 0.1 and the acrophase (the time of maximum alertness) from 15:8 to 21:8 hours. Thereafter, a series of regression analyses were carried out with the mean actual values as the dependent variable and the predictions from the model as the independent variables. The selection of a solution depended upon a one to one relation between the predicted and actual values—that is, a zero intercept and a slope = 1, with a high R².

**Results**

**MODELLING**

When carrying out the analyses it was noticed that all points could be estimated with a high precision except for the one at 15 00 after 32 hours of wakefulness, which also had a very large SD. It seems that half the subjects woke up after only three hours of sleep, whereas the other half slept for 11 hours. The reason for this large variation may be that at this point the assumed acrophase occurs only two hours later so its influence on the end of sleep might be expected to be at a maximum. Small individual differences in the position of the acrophase may determine whether sleep will be ended or continued for a long time. To reduce these effects we decided not to use the 15 00 values in our modelling.

Figure 3 shows the highly significant variation across the data, with a maximum duration of sleep for a bedtime at 19 00 after 36 hours awake and a minimum at 11 00 after four hours awake. This pattern is closely mirrored by the best prediction from the model (explained later).

The best prediction was determined from figure 4. It shows that the intercept (which should be zero) was large and negative (duration of sleep is underestimated) at low amplitudes and even bigger with increasing amplitude. Thus, high amplitudes consistently underestimated sleep by several hours, unless corrected for by a large constant (intercept). The acrophase modified the effect of amplitude such that the intercept increased with delay. The slope (which should = 1) fell as the amplitude increased and as the acrophase was delayed. The optimum solution for the criteria is indicated by the intersection of the vertical line from the x axis at zero and the horizontal line from the y axis at 1. This yields an amplitude of 0.6 and an acrophase of 16.48, with a regression function of: y = -0.02 + 0.99x, where y is the actual duration of sleep and x is the predicted duration of sleep.

The amount of explained variance for the optimum function was R² = 0.99. This
amount of variance could also be explained by the amplitudes from 0-4 to 1-1 with the acrophase still at 16 48. The same high level could be maintained also for amplitudes 0-4-1·0 with the 17 48 acrophase and for amplitudes 0-4 and 0-5 with the 18 48 acrophase. Other solutions fell below 90% and became marginal for late acrophases and high amplitudes.

**VALIDATION**

To validate predictions of the natural end of sleep is difficult as very few experimental studies have allowed sleep to end spontaneously as part of the design. Some information may, however, be gained from studies of shift workers, who sleep in their home environment and often end their sleep spontaneously (without alarm clocks), at least for sleep after a night or evening shift. The morning shift usually needed to be woken by artificial means and was not included in these validations.

In a study of three night work with 24 hour ambulatory polysomnography we used data from 25 male workers (22–55 years old) on a rotating three shift system. Twenty four hour polysomnography was carried out during the second 24 hour period on each shift. For the second afternoon shift the bedtime was 23 27, preceded by work between 14 00 and 22 00, and by sleep between 23 40 and 08 18 (no napping occurred). For the second night shift the bedtime was 06 40, preceded by a 25 minute nap at 15 04, the previous sleep was between 06 27 and 12 33, and work between 22 00 and 06 00. The subjects slept in their own bedrooms for as long as they wanted to. The duration of sleep differed significantly between shifts and was similar to what has been found in other electroencephalographic studies of shift workers who sleep at home. The model used an acrophase of 16 48 (or for comparison with a late acrophase: 20 48), and its prediction came relatively close to the actual values. Thus, after the night shift the model predicted a duration of 5-3 or 6-8 hours, depending on the acrophase (16 48 or 20 48) v the actual value of 5-2 hours. After the evening shift (sleep start at 23 37) the values were 7-9 or 10-0 hours v 7-5 hours.

In another study 11 male train drivers (age 27–58) took part. Sleep in the day differed significantly from night sleep and both were relatively close to those obtained by Foret and Lantin. The previous sleep data were entered into the model and the sleep latency predicted (with an acrophase at 16 48 or 20 48). For night sleep after day work, the bedtime occurred at 23 07 (after a previous sleep between 23 10 and 07 10, without naps) and we predicted 8·2–10·2 hours (for acrophases of 16 48 and 20 48 respectively) v the actual 8·5 hours. For day sleep after night work (at 07 36 and after a previous sleep between 23 15 and 09 10, plus a 1·5 hour nap at 16 00), the model predicted 4·7–6·2 hours v the actual 4·5 hours.

In a third study by Gilberg et al sleep in the laboratory started at 11 00 after rising from the previous night’s sleep at 07 00. The spontaneously ended duration of sleep was 110 minutes whereas the model predicts 120 minutes for the early acrophase and 210 minutes for the late one (normal day workers, 25–45 years of age, were used).

A regression of predicted on actual results from the three validation studies (five points) yielded $R^2 = 0·99$ (with an intercept of 0·1 hour) for the 16 48 acrophase and $R^2 = 0·95$ for the late one (with an intercept of 0·6 hours).

**Discussion**

The modification of the alertness model to predict duration of sleep could be made with a high degree of fit to the data used to derive it. The results also seem to predict duration of sleep reasonably well under conditions of irregular working hours, although daytime duration of sleep was slightly underestimated. This may have been due to daytime noise levels, or other factors like stress, climate, drugs, etc. Drugs do not belong in the present model of alertness and sleep regulation but will have to be considered when interpreting the data. Still, the predictive power of the model seems to be considerable, even though further validation is needed, particularly for more unusual combinations of sleep loss and time of day. At present, however, very little other polysomnographical data on spontaneous duration of sleep is available for analysis. Indeed, the total dominance of scheduled, as opposed to spontaneous, ending of sleep in experimental sleep research is quite remarkable.

Apart from sheer disturbances from the environment we must also consider other factors outside the model that will systematically influence sleep duration. One such is the ultradian rhythm, which causes a bimodal sleep pattern with an increased tendency to sleep during the afternoon. There is, however, very little knowledge available about the influence of the factor on sleep duration. Another influence is individual differences, such as diurnal type, somatotype, long or short sleepers, age, or sex. Modifying the model for such influences will probably make it possible to predict individual data. At present, it should be emphasised that the model should be applied only to group mean data.
Figure 5: Duration of sleep nomogram (see text). Each curve represents predicted duration of sleep for each bedtime, after a given number of hours awake. The acrophase was set to 16.48. The arrows illustrate the use of the nomogram to predict duration of sleep at 07.00 after 24 hours of wakefulness.

For the technical aspects of the model we think that the end of sleep theory is intuitively attractive—recovery of alertness (S') gradually approaching a threshold of complete recuperation, unless interrupted by circadian interference. However, one might question the choice of the adjusted asymptote. The specific value selected was simply based on the need to have a value lower, but close to the asymptote of S'. Considerably lower values could have been selected, but that would have required adjustment of the parameters S' and C'. This seemed a pointless exercise as no increased precision would be gained. In the same vein one might also discuss the choice of modifying only C', and not S'. The reason was simply that the concept behind the end of sleep model postulated the circadian influence as the modifier of the recuperative process (S'), imposing severe restrictions to duration of sleep at the acrophase and none at the trough. Thus, the problem of excessively short daytime sleep (compared with actual data) was handled by modifying the amplitude and the acrophase of C'. Incidentally, removal of C' would yield a duration of sleep of 10-5 hours after a bedtime at 23.00 (v the 8-5 hours with C') and 11 hours after a bedtime at 07.00 after 24 hours awake (v the 5-3 hours with C').

The fact that the amplitude of C' turned out to be much lower (0-6) than that of C (2-5) could be interpreted to mean that the circadian influence during waking is stronger than during sleep, at least in terms of ending sleep. It is an interesting question whether these differences possibly correspond to the sleep inertia seen after waking? and whether the dissipation of sleep inertia then corresponds to the recovery of a normal circadian amplitude (from C' to C).

An alternative explanation of the lower amplitude of C' may lie in the manner in which we have modelled duration of sleep. For the sake of parsimony, we have assumed that C has no influence on the initial level of S' at the start of a sleep episode, and that this level depends only on the value of S at the end of the previous period of wakefulness. It is clear from this paper that these assumptions resulted in the fairly accurate modelling of many results. However, if both S and C were to influence the starting level of S', this might well result in a rather larger estimated amplitude of C'. This issue will have to be considered in subsequent studies.

The acrophase had, for obvious reasons, a strong influence on the end of sleep, with overestimation increasing with later acrophases. Clearly, the optimum acrophase for the end of sleep should be at about 16.48. This is the originally established acrophase and could probably be marginally improved by a higher resolution of the analyses. On the other hand several other factors, such as the ultradian influence and diurnal effects, need to be evaluated before determining the acrophase with higher precision.

It should be emphasised that the present alertness model was inspired by the three process model of sleep regulation and may be seen as an extension of that model to provide practical tools for predicting duration of sleep, alertness, and sleep latency in connection with irregular sleep patterns. Indeed, our work confirms many of the ideas of the Borbély-Daan model, with completely different data sets, and focusing on very different types of variables (subjective alertness). In fact, the parameters in our model are remarkably very close to those of the original one. For the construction of the end of sleep event, the Borbély-Daan model ends sleep when S during sleep reaches a skewed circadian threshold (H)—derived from rectal temperature. As discussed above, in our model sleep ends when S' + C' add up to an end of sleep level, which is the adjusted asymptote of S'. Both solutions are very similar mathematically. The fact that all components of the alertness model are derived from alertness ratings makes it possible to view the end of sleep as a result of sufficient alertness having been recovered for sleep to end at a particular circadian phase. This integration of alertness and sleep may make the relation between wake and sleep easier to analyse and perhaps the regulatory processes of circadian and homeostatic sleep easier to understand.

One further attempt to bring the alertness model closer to the user is to simplify its use. In our paper, the predictions from the model were obtained through an interactive computer program developed for Macintosh. However, to simplify and to increase accessibility we have developed a “barefoot" version of it—the duration of sleep nomogram. It is envisaged that such an instrument may be
used to train shift workers to optimize sleeping strategies or to guide administrators towards more physiological scheduling of shifts.

The nomogram was constructed with the program (with C set to 16.48) to simulate duration of sleep for each time of day and after 0-36 hours awake. Figure 5 shows the results and contains the predictions based on time of day and hours since the preceding major sleep period. Naps and shortened major sleep periods are not included yet. Each curve represents the circadian variation of duration of sleep (plotted at bedtime) after a certain number of hours awake. To use the nomogram, the subject first identifies the point in time (see arrows for example) for which a prediction of duration of sleep is desired. From this point a vertical line is drawn upwards to the curve that represents the number of previous hours awake (for that point in time). From this intersection a horizontal line is drawn to the left and predicted duration of sleep is read off on the y axis. In the example, a prediction is sought for 07:00 at the end of the first night shift. The point is thus preceded by about 24 hours of wakefulness (if no naps have occurred). The prediction lands on a duration of sleep of 5-3 hours for the early acrophase.

It is concluded that the three process model of regulation of alertness is accurately able to predict changes in duration of sleep due to altered sleep patterns. It is suggested that this aspect of the model may be used as a tool for estimating the risk of shortened sleep induced by irregular working schedules, and in turn, the suitability of the schedule.