Dealing with Inter-Individual Differences in the Temporal Dynamics of Fatigue and Performance: Importance and Techniques

Hans P. A. Van Dongen, Greg Maislin, and David F. Dinges

Van Dongen HPA, Maislin G, Dinges DF. ***Dealing with inter-individual differences in the temporal dynamics of fatigue and performance: importance and techniques*****. Aviat Space Environ Med 2004; 75(3, Suppl.):A147–54.**

Inter-individual differences in performance impairment from sleep loss are substantial and consistent, as demonstrated and quantified here by means of the intraclass correlation coefficient (ICC) in two laboratory-based sleep deprivation studies. There is an urgent need, therefore, to consider inter-individual variability in biomathematical models of fatigue and performance, which currently treat individuals as being all the same. Traditional regression techniques do not handle inter-individual variability, but cutting-edge mixed-effects modeling techniques have recently become available to deal with inter-individual differences in the temporal dynamics of fatigue and performance. The standard two stage (STS), restricted maximum likelihood (REML), and non-linear mixed-effects modeling (NMEM) approaches to mixed-effects models are compared here using data from a chronic partial sleep deprivation experiment. Mixed-effects modeling can be incorporated in the two distinct steps (the direct and inverse problems) of biomathematical model development in order to deal with inter-individual differences. This paper demonstrates that inter-individual variability accounts for a large percentage of observed variance in neurobehavioral responses to sleep deprivation, and describes tools that model developers will need to produce a new generation of fatigue and performance models capable of incorporating inter-individual variability and useful for subject-specific prediction.

Keywords: sleep deprivation, performance, inter-individual differences, between-subject variance, within-subject variance, intraclass correlation coefficient, ICC, mixed-effects models, standard two stage, STS, restricted maximum likelihood, REML, nonlinear mixed-effects modeling, NMEM, biomathematical model development.

BIOMATHEMATICAL MODELS of fatigue and performance are based in part on the changes in fatigue and performance consistently observed as a consequence of sleep loss. For fundamental neurocognitive functions such as vigilance, working memory, situational awareness, and decision-making capability, degradation has been documented to be a systematic consequence of sleep loss (12,21). Across multiple days of total sleep deprivation, increases in fatigue and decrements in performance occur in response to increasing homeostatic sleep drive in interaction with circadian variation. This results in accumulation of deficits with increased circadian modulation over time, making the impairment worse during biological night (19,40). Even when wakefulness is extended by chronic sleep reduction rather than total sleep deprivation, considerable cumulative neurobehavioral deficits develop over time (2,41).

Virtually every experiment on human sleep deprivation has reported substantial inter-individual differences (often tenfold on the most sensitive performance metrics) in the magnitude of performance deficits due to sleep loss (12). These inter-individual differences can be illustrated by categorizing subjects based on the magnitude of their performance deficits from sleep loss. **Fig. 1** shows results from 40 h of sleep deprivation in our laboratory (11). As measured by performance lapses on a psychomotor vigilance task (13), some subjects were found to have much greater impairment from sleep loss than others (Fig. 1, left-hand panel). Thus, some subjects seemed to be relatively vulnerable to performance impairment due to sleep loss, while others seemed to be relatively resilient. Interestingly, vulnerable and resilient individuals did not differ significantly in the amount of sleep they felt they needed or routinely obtained, as assured by surveys and daily diaries. This suggests that inter-individual differences in vulnerability to sleep loss are not just determined by differences in sleep need (42).

Inter-individual differences in vulnerability to performance deficits from sleep loss are particularly problematic since individuals may not be fully aware of their susceptibility to impairment (41). Performance impairment associated with loss of sleep appears to manifest itself in the form of wake state instability (14). Individuals may start well on a given task, with no indication of vulnerability at that time, but their performance rapidly becomes more variable and unreliable as task demands continue. Thus, subjects may have little a priori insight into their capability to sustain optimal performance under conditions involving sleep loss. This can be illustrated by classifying the subjects in our 40-h sleep deprivation experiment as either resilient or vul-

From the Division of Sleep and Chronobiology, Department of Psychiatry, University of Pennsylvania School of Medicine, Philadelphia, PA (H. P. A. Van Dongen, D. F. Dinges), and Biomedical Statistical Consulting, Wynnewood, PA (G. Maislin).

Address reprint requests to: Hans P.A. Van Dongen, Ph.D., who is a Research Assistant Professor of Sleep and Chronobiology at the Unit for Experimental Psychiatry, University of Pennsylvania School of Medicine, 1019 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021; vdongen@mail.med.upenn.edu.

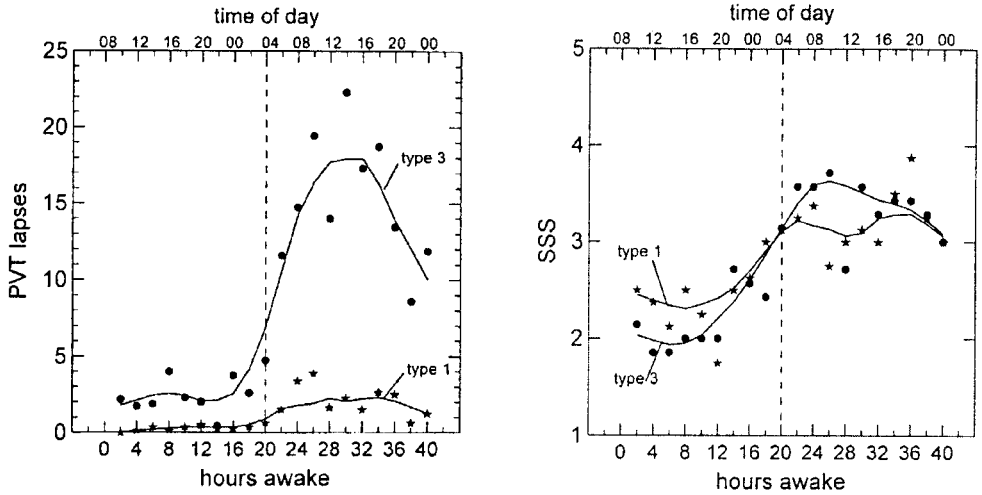

Fig. 1. Neurobehavioral performance lapses on a psychomotor vigilance task (PVT; left-hand panel) and subjective sleepiness score on the Stanford Sleepiness Scale (SSS; right-hand panel) during 40 h of total sleep deprivation in a laboratory environment (11). Stars show PVT performance lapses and SSS sleepiness scores for the eight subjects most resilient to sleep loss (type 1); dots show the data for the 7 subjects most vulnerable to sleep loss (type 3). Despite the considerable difference in psychomotor performance impairment, there was no statistically significant difference in the profile of sleepiness scores between these two groups.

nerable to neurobehavioral impairment from sleep loss. When the subjects filled out the Stanford Sleepiness Scale (22) prior to neurobehavioral performance testing, no differences were observed between resilient and vulnerable subjects in the temporal profiles of their sleepiness self-ratings (Fig. 1, right-hand panel). Thus, their differences in cognitive performance deficits were not reflected in their subjective experiences of sleepiness.

In safety-sensitive operations, performance deficits due to sleep loss elevate the risk of accidents (1,9). Nevertheless, a relatively small portion of individuals in populations at risk may be found to account for the bulk of neurobehavioral impairment (29). As people cannot be relied on to accurately assess their own risk level (see Fig. 1), predicting which individuals are most at risk for performance impairment at any given time would be an important capability of biomathematical models of fatigue and performance (10). Current models of fatigue and performance (28) do not address this issue. Recent developments in statistical methodology, driven by major increases in computer power, have created new possibilities for handling inter-individual differences in biomathematical models of the temporal dynamics of fatigue and performance. This paper summarizes evidence that substantial inter-individual differences exist, and provides an introduction into some techniques available to assess and handle inter-individual differences in temporal profiles of responses to sleep deprivation.

Stable Inter-individual Differences in Vulnerability to Sleep Loss

While inter-individual differences in vulnerability to performance impairment from sleep loss could reflect stochastic variance (i.e., noise), there is converging evidence that they reflect more predictable variability. The stability of inter-individual differences is apparent in the five studies that have been done on repeated exposure to sleep loss in the same subjects. The early studies of Wilkinson (44) and Webb and Levy (43) both reported substantial inter-individual differences in the effects of sleep deprivation that appeared to reliably reflect greater sensitivity of some subjects to the loss of sleep. Wilkinson (44) studied 12 subjects under multiple exposures to one night of total sleep deprivation. Based on the subjects' performance on a five-choice serial reaction test, Wilkinson observed that "impairment varied greatly. . . those men who showed a large effect of lack of sleep in the first [2-wk period] did the same in the second and third. This effect was significant. . . This consistency suggests that there are indeed real differences in the extent to which individuals are affected by lack of sleep" (44, p. 267). Webb and Levy (43) studied six subjects under multiple exposures to one night of total sleep deprivation. During each laboratory session, the subjects underwent six extensive test bouts containing a variety of performance tests. Webb and Levy noted that there were substantial individual differences in the effect of sleep deprivation. Furthermore, the individual differences in performance decrements from sleep loss appeared to be reliable: "examination of the individual scores of these tests revealed that, in each instance, they reflected greater sensitivity of some subjects to deprivation" (43, p. 56). Neither the Wilkinson study (44) nor the Webb and Levy study (43) actually quantified the stability of inter-individual differences in performance deficits from sleep loss.

Leproult et al. (25) studied eight subjects twice under similar constant routine conditions. These conditions entailed 27 h of continuous wakefulness with constant bed rest, semi-constant light intensity, and controlled caloric intake via intravenous glucose infusion. Continuous wakefulness was verified post-hoc by investigation of waking electroencephalogram (EEG) recordings. During the constant routine, subjects were tested hourly on a selective attention task, a sustained attention task, and a visual analog scale for global vigor. These measures were smoothed using a three-point

moving average, after which the minimum and maximum in the temporal profiles were determined. To assess the magnitude of impairment in each of the two exposures to sleep deprivation, the difference between maximum and minimum in the temporal profile was expressed as a percentage of the minimum (for the two attention tasks) or the maximum (for global vigor). Parametric and nonparametric correlations over subjects between the magnitudes of impairment in the first vs. the second exposure to sleep deprivation were used as quantitative measures for the stability of inter-individual differences in the response to sleep deprivation. For minimum global vigor, the correlation coefficients were reported to be 0.95 (parametric) and 0.90 (nonparametric); for maximum reaction time on the attention tasks, the correlation coefficients were 0.93 (parametric) and 0.88 (nonparametric).

Although the Leproult et al. (25) investigation is indicative of predictable inter-individual differences in responses to sleep deprivation, the use of correlation analysis is not optimal for evaluating the degree to which responses to sleep deprivation reflect systematic inter-individual variability. Correlation statistics only implicitly recognize the partitioning of total variance into between-subject variance and within-subject variance. Thus, it is often overlooked that the magnitude of the correlation coefficient varies for populations with different amounts of between-subject variance. Moreover, correlation statistics do not easily generalize to a partitioning of total variance into multiple components, each of which reflect a distinct source of variance, and also do not readily distinguish between random effects and those fixed by design within an experimental paradigm. If there are order effects or other systematic changes over time in the data (which cannot be ruled out in reference 25), correlation statistics can be considerably inflated (cf. 26). Therefore, correlation statistics cannot be relied on to quantify inter-individual variability. This is noteworthy because none of the three studies discussed thus far (25,43,44) revealed whether the inter-individual variability in the data is large enough to be of concern for biomathematical models of fatigue and performance.

Alternative approaches involving random-effects and mixed-effects analyses of variance (ANOVA) and variance components analyses overcome these difficulties (36). In the simplest case, an intraclass correlation coefficient (ICC; 18) can be computed as the ratio of between-subject variance to the sum of between-subject variance and within-subject variance. It thus provides a direct assessment of the proportion of variance in the data explained by inter-individual variability. As such, the ICC is seen not as a fixed characteristic of a test or measuring instrument, but as depending on the population of subjects being sampled (16). This interpretation of the ICC as the proportion of total sample variance attributable to between-subject variance is often useful as a summary measure. Mixed-effects ANOVA can subsequently be employed to remove between-subject variance explainable by known sources, such as age and gender, resulting in an adjusted ICC. The adjusted ICC would be interpreted as the proportion of variance attributable to between-subject variance after removing the variance explained by differences in these known sources of inter-individual variability. More general variance components models can be specified and estimated by restricted maximum likelihood (REML) methodology (7,33) using widely available software (e.g., 35).

In the next section, we present results using the ICC to quantify the stability of inter-individual differences. The only two studies to date using this statistic (38,39) will be discussed as providing quantitative evidence of trait-like inter-individual differences in vulnerability to performance impairment from sleep loss.

Trait-Like Inter-Individual Differences Quantified by Means of the Intraclass Correlation Coefficient

As introduced above, a statistically suitable approach to assessing trait-like inter-individual variability in performance impairment, over multiple exposures to sleep loss per subject, involves considering two distinct components of the variance in the data: within-subject variance and between-subject variance. The within-subject variance reflects the changes in performance impairment across the repeated exposures to sleep loss for each of the individual subjects. The between-subject variance reflects inter-individual differences in performance impairment not accounted for by variability within subjects; that is, the between-subject variance reflects systematic inter-individual differences over the repeated exposures to sleep loss. Inter-individual differences are shown to be systematic over repeated exposures to sleep loss, therefore, when the between-subject variance is found to be relatively large and the within-subject variance is found to be relatively small. The variance in the data can be partitioned into the between-subject and within-subject variance by way of a random-effects ANOVA (36).

Based on the between-subject variance (σ_{bs}^2) and the within-subject variance (σ_{ws}^2), the ICC is defined as follows to quantify trait-like inter-individual variability:

$$\mathrm{ICC} = \frac{\sigma_{bs}^2}{\sigma_{bs}^2 + \sigma_{ws}^2}$$

The greater the between-subject variance relative to the within-subject variance, the closer to 1.0 the value of the ICC is. Landis and Koch (24) defined benchmark values characterizing ICC values in the following ranges, which here reflect increasing stability of observed inter-individual differences: 0.0–0.2 (slight); 0.2–0.4 (fair); 0.4–0.6 (moderate); 0.6–0.8 (substantial); and 0.8–1.0 (almost perfect). An F test or Wald Z test can be applied to examine the statistical significance of ICC values (relative to zero), although significance testing is usually only of passing interest as the salient issue is the magnitude of between-subject variance relative to other sources of variance.

The ICC proper is a statistically valid measure for quantifying the stability of inter-individual differences over repeated exposures to sleep loss. In order to interpret ICC values, however, it is important to also consider the within- and between-subject variances sepa-

rately. The within-subject variance may encompass measurement error, random fluctuations, as well as systematic differences in the circumstances from one sleep deprivation challenge to the next. The within-subject variance may also include differential learning curves for cognitive performance tasks, and changes in sensitivity to sleep deprivation over repeated exposures. On the other hand, the between-subject variance represents trait-like variance, both in the vulnerability to impairment from sleep loss as well as in potential baseline differences such as aptitude differences. In addition to true trait variance, the between-subject variance may also reflect state differences among individuals that remain constant over the repeated exposures to sleep loss; these may include consistent differences in sleep timing and duration prior to the repeated sleep deprivation experiments, consistent differences in food intake during each exposure to sleep deprivation, etc. In order for the between-subject variance to reliably approximate the true trait variance, therefore, consistent state variance must be minimized by standardizing demand characteristics, controlling environmental factors, and satiating any pre-existing sleep debt, in repeated exposures to sleep deprivation.

Using the ICC, we quantified the stability of inter-individual differences in neurobehavioral performance responses to sleep loss in two different experiments (38,39) involving repeated exposure to sleep deprivation under strictly controlled laboratory circumstances. In the first experiment (39), 10 subjects were subjected to 40 h of total sleep deprivation on 2 occasions each. Subjects were tested every 2 h on a 10-min psychomotor vigilance task (13). For each of the two exposures, which differed in the degree of environmental stimulation subjects experienced (8), vigilance decrements were computed as the number of psychomotor vigilance lapses from 12:00 until 20:00 at the end of the 40-h sleep deprivation period (i.e., response to sleep loss) minus the number of lapses at the same clock times 24 h earlier (i.e., baseline). The results are shown in **Fig. 2**. Substantial inter-individual differences were observed in vigilance decrements, while the response to sleep deprivation was remarkably similar within subjects across the two exposures. The value of the ICC was 0.58, indicating that a substantial portion of the variance (i.e., 58%) in the data could be attributed to trait-like inter-individual differences in the response to sleep deprivation ($F_{9,9} = 4.61$, $p = 0.016$).

We further investigated the trait aspect of inter-individual differences in vulnerability to neurobehavioral performance deficits from sleep loss in a second laboratory experiment, in which 21 subjects were exposed to 36 h of sleep deprivation on 3 occasions (38,42). The level of comparability between the exposures to sleep deprivation exceeded that of the first study, and included laboratory isolation, strictly controlled food intake, restriction of light exposure to less than 50 lux, fixed ambient temperature, and continuous waking behavioral monitoring. The study also required subjects to satiate their sleep debt by staying in bed 12 h per day (22:00–10:00) in the week prior to two of the three exposures to sleep deprivation, and to restrict their sleep to 6 h time in bed per day (04:00–10:00) in the week prior to the remaining exposure to sleep deprivation (in randomized, balanced order). Every 2 h during wakefulness, subjects were tested on a 1-h neurobehavioral performance battery, which included a 6.5-min digit-symbol substitution task and a 20-min psychomotor vigilance task.

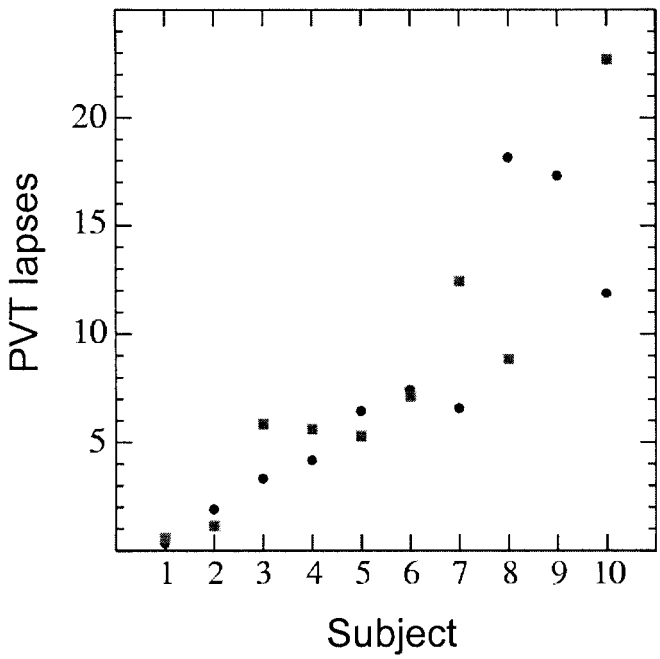

Fig. 2. Relative performance decrement after 40 h of total sleep deprivation for 10 subjects, on 2 separate occasions (circles vs. squares). The 10 subjects are shown on the abscissa (numbered 1 through 10). The ordinate shows the number of performance lapses on the psychomotor vigilance task (PVT) from 12:00 until 20:00 after one night of sleep deprivation, minus the baseline number of performance lapses at the same clock times 24 h earlier. Note that the data for the second sleep deprivation session of subject 9 are missing.

Neurobehavioral deficits on the performance tasks in the second study were compared between the two identical sleep deprivation sessions (i.e., the two which were preceded by sleep satiation in the week prior to deprivation). For each of the two sessions, subjects' performance deficits were computed as the average of the test results during the last 24 h of the 36-h sleep deprivation period. ICC values were computed while correcting for any order effects in the exposures to sleep deprivation; this was relatively easily accomplished in a mixed-effects analysis of variance (ANOVA) used to calculate the ICC. For the digit-symbol substitution task, the ICC was found to be 0.82 ($Z = 2.74$, $p = 0.003$); and for the psychomotor vigilance task, the ICC was 0.69 ($Z = 2.40$, $p = 0.008$).

These results indicate that there are substantial trait-like inter-individual differences in vulnerability to sleep loss. It should be noted that performance aptitude differences among individuals may have contributed to the between-subject variance in addition to the true trait variance. Therefore, the ICC estimates should be interpreted with care for cognitive tests, such as the digit-symbol substitution task, which are particularly prone to variability in aptitude. It would seem reasonable to consider the ICC value of 0.69 found for psychomotor

vigilance, which involves virtually no aptitude differences among healthy individuals, as a good (i.e., lower-limit) quantitative estimate of the trait-like inter-individual differences in vulnerability to performance impairment from sleep loss.

The between-subject variance for psychomotor vigilance, corrected for order effects, was 286.09. This between-subject variance (and hence the value of the ICC) is specific for our sample of healthy adults (age 29.5 ± 5.3; nine women) and may vary for other populations. The within-subject variance corrected for order effects was 131.27. The within-subject variance is theoretically dependent only on the experimental design and not the population from which the sample was obtained. The within-subject standard deviation (which was 11.46) may be interpreted as the standard deviation around the known expected value of an individual, and so has potential for use in individualized predictions. In contrast, the standard deviation for predictions made for randomly selected individuals (from our population) is equal to the square root of the sum of the variance components (i.e., 20.43). Thus, prediction error variance conditional on subject-specific expected values is much smaller than prediction error variance in the population at large.

In summary, the ICC and its associated variance components can be used to quantify the stability of inter-individual differences in repeated experiments. Our two studies (involving a total of 31 subjects) confirmed that the inter-individual differences in vulnerability to performance impairment from sleep loss are substantial and trait-like. The ICC values, which are measures of the systematic inter-individual variability relative to the total variability in the data set, revealed that (much) more than 50% of the variance in the neurobehavioral response to sleep deprivation is due to systematic inter-individual differences. Ignoring these inter-individual differences would mean ignoring the larger portion of the variance in the data. Thus, it is important to take inter-individual differences into account in biomathematical models of fatigue and performance. In the next section, therefore, we discuss statistical techniques to deal with inter-individual differences in time series.

Statistical Modeling of Time Series with Trait-Like Inter-Individual Differences

The classical approach to the analysis of longitudinal data involves the use of repeated-measures ANOVA (45). Both the univariate approach with and without adjustment for unequal variances and covariances (23) and the often preferred multivariate approach to repeated measures (30) suffer from the same limitation in that they ignore inter-individual variability in time trajectories. They both provide explicit models for population expected values; however, all deviations around the expected values are assumed in residual error distributions. As a consequence, when inter-individual variability exists, these methods result in confounded estimates of standard errors, providing unreliable claims regarding statistical significance.

If inter-individual differences are not of primary interest, investigators may be tempted to fit responses to time using simple or multiple linear regression models in order to obtain population mean response curves, or to compare population mean responses over time among two or more experimental conditions. However, such an approach can lead to confounded results threatened by both inflated Type I (false positive) and Type II (false negative) errors (17). Ignoring the inter-individual variability can increase Type I error because within-subject variability is pooled with between-subject variability, falsely reducing the estimate of residual error that is used in statistical tests. This problem is particularly acute when there are relatively few subjects with many time points. Type II error can be inflated when consistent within-subject trends are undetected and not accounted for after pooling across subjects. This kind of Type II error inflation is analogous to that occurring when a two-sample *t*-test is used when a paired-samples *t*-test should be used. Thus, when data from different groups of subjects or different experimental conditions are compared, it is critical to take inter-individual variability into account.

There is a relatively recent but well-developed literature concerning mixed-effects modeling of longitudinal data, explicitly accounting for inter-individual differences (5,27,47). A conceptually simple and, in approximation, valid approach for analyses of changes over time is two-stage (random-effects) regression analysis or "standard two stage" (STS; 17,20). The first stage consists of obtaining least-squares estimates of simple linear regression slopes and intercepts for each subject independently. The slopes are interpreted as subject-specific estimates of the average change per unit time. The intercepts are interpreted as smoothed estimates of baseline values. The need for intercept estimation may be eliminated by modeling changes from time zero. The derived slopes are used in a subsequent analysis, the second stage, that can take a variety of forms. For example, mean slopes can be compared among experimental conditions with or without adjusting for baseline values and other factors (e.g., age and gender) in an analysis of covariance (ANCOVA).

A more efficient approach is to estimate all individual subjects' slopes simultaneously using restricted maximum likelihood (REML) estimation in an explicit mixed-effects model. The efficiency arises by partitioning between-subject variance and within-subject variance during the estimation procedure. REML estimation is preferred over standard maximum likelihood (ML) estimation for mixed-effects models because ML produces biased estimators of variance components (7). The magnitude of the bias increases as the number of fixed effects increases relative to the number of data points. This phenomenon is analogous to the well-known bias in the ML estimate of population variance (15). In essence, the REML method deals with this problem by considering linear combinations of the observed values whose expectations are zero. These "error contrasts" are free of any fixed effects in the model. Thus, REML estimates of variances and covariances are unbiased in balanced experimental designs. REML estimates of variance components have the same asymptotic distributional properties as their ML counterparts. In prac-

tice, therefore, unless there are many fixed effects, REML and ML produce very similar estimates of the variance components in the data.

The STS and the REML approaches have advantages and disadvantages. Aside from conceptual and computational simplicity, the STS approach often permits the use of valid small-sample statistics. Thus, STS can be relied on, whereas REML cannot, to produce accurate p-values when there are very few subjects, so long as the assumptions of the small-sample model (e.g., normality) are met. Furthermore, STS may be more robust than REML when particular subjects have extreme slopes or extreme residual errors. However, STS fails to account for the covariance between slopes and intercepts. Also, it assumes equal weighting of subjects' slopes in the second stage, which is not appropriate if the number of data points or the layout of the time values varies widely among subjects. Finally, STS disguises residual error, pooling it with between-subject variance and biasing the latter upward. If residual variance is small or the numerical values of variance components are not themselves of interest, however, this is not usually a problem. The primary drawback of REML is slow or difficult numerical convergence when some of the variance components are small. When both approaches are applied to balanced data sets, mean slope values are found to be identical. However, the standard deviation of subject-specific slopes tends to be a bit smaller for REML analysis than for STS analysis, since in REML analysis any extreme slopes are assumed to be due to both extremities in subject-specific effects and in residual error effects. This reduction in the standard deviation of slopes over subjects results in incrementally more statistical power in, for example, group comparisons of mean slopes in REML analysis.

The assumption of linear changes using both the STS and REML approaches is not as restrictive as it appears. Individual changes from baseline may be characterized using a model that allows for curvilinear changes, for instance, by introducing a curvature parameter θ for time. Consider the following model:

$$\Delta_{it} = B_i \cdot t^{\theta} + \varepsilon_{it}$$

where i represents the different subjects, t denotes time, and the residual errors ϵ_{it} are assumed to be mutually independent and normally distributed with zero average. The changes Δ_{it} in the response over time are modeled, rather than the response over time itself, to remove the necessity of estimating regression intercepts. The B_i values reflect subject-specific growth rates (slopes) assumed to arise from normal distributions with condition-specific mean values. In this model, the time exponent θ represents curvature in the growth curves. If θ is equal to 1, the model for each subject reduces to simple linear regression (with no intercept). If $0 < \theta < 1$ or $\theta > 1$, then the growth curves are decelerating or accelerating over time, respectively. For any response variable, an optimum curvature θ may be obtained by finding the value that minimizes average mean square error (MSE). For example, the optimum value of θ may be obtained from a grid search in which θ is systematically varied from 0.1 to 3.0 in increments of 0.1. For each value of θ, the MSEs for the subject-specific regressions are then averaged in STS, or the residual variance is examined in REML. The key point is that conditional on establishing the value of θ, the mixed-effects model remains linear permitting application of STS or REML as described above.

Alternatively, nonlinear mixed-effects modeling (NMEM; 6) may be performed. This method is described in sufficient detail in Olofsen et al. (32) and will not be further explicated here. In the NMEM approach, the optimum curvature parameter θ is estimated simultaneously with the subject-specific slopes using maximum likelihood estimation. The chief challenge of this approach is to obtain numerical convergence in the maximum likelihood optimization routine. However, the NMEM method is more flexible than the REML and STS methods with regard to the types of (nonlinear) equations it can handle.

To illustrate and compare the three approaches, STS, REML, and NMEM, we consider data from an experiment in which a total of 35 healthy subjects (age 27.7 $\pm$ 5.4) spent 20 d inside a laboratory. After 3 baseline days with 8 h time-in-bed (23:30–07:30), sleep was restricted over 14 d to 4 h time-in-bed per day (03:30–07:30) for 13 subjects; to 6 h time-in-bed per day (01:30–07:30) for 13 subjects; and to 8 h time-in-bed per day (23:30–07:30) for 9 subjects. Neurobehavioral performance was tested every 2 h during wakefulness, and included a 10-min psychomotor vigilance test (PVT; 13). Daily averages (09:30–23:30) were computed for the number of PVT lapses (reaction times $\geq$ 500 ms) per test bout (41).

Table I displays summary statistics of the (nonlinear) slopes B_i for daily average PVT lapses over the 14 d of sleep restriction. These were computed using STS, REML, and NMEM for each subject in the three conditions; the value of θ that minimized average MSE was 0.7753. The substantial inter-individual variability in the response to the experiment is evident in the table—notice the range (minimum to maximum) and standard deviation of B_i values within each condition of chronic sleep restriction. There is considerable agreement in the subject-specific B_i values, as well as the mean slope values for each condition, among the three approaches. There is a slight difference in the means for the 8-h time-in-bed condition between the STS method and the REML and NMEM methods. This is caused by a single missing value in the temporal profile for one subject in the 8-h time-in-bed condition that the STS does not properly weigh. The standard deviations from the STS method are slightly greater, as the statistical theory predicts. It is noteworthy that both maximum and minimum extreme values for B_i obtained using STS are farther removed from the within-group mean values than are those obtained using REML and NMEM. The extreme values obtained using STS are slightly biased upwards due to intertwining of between-subject variance and error variance. For REML and NMEM, the extremes are attenuated appropriately toward the within-group mean values, as a consequence of partitioning between-subject and within-subject variance in these approaches.

Table I provides evidence that the use of an optimal value for time curvature θ permits the application of

TABLE I. SUMMARY STATISTICS FOR THE SUBJECT-SPECIFIC REGRESSION COEFFICIENTS B_i ESTIMATED WITH THREE MIXED-EFFECTS MODELING METHODS.

Condition*	Method	Mean	SD	Min	Max
4 h TIB	STS	1.9269	1.3493	0.0638	3.8784
4 h TIB	REML	1.9269	1.3245	0.0981	3.8425
4 h TIB	NMEM	1.9268	1.3255	0.0965	3.8440
6 h TIB	STS	1.2897	1.6999	−0.5383	5.5608
6 h TIB	REML	1.2897	1.6686	−0.5046	5.4822
6 h TIB	NMEM	1.2896	1.6700	−0.5062	5.4855
8 h TIB	STS	0.3345	0.6851	−0.3228	2.0517
8 h TIB	REML	0.3353	0.6721	−0.3107	2.0201
8 h TIB	NMEM	0.3352	0.6726	−0.3113	2.0215

* There were 35 subjects exposed to 14 d of sleep restriction. The subjects were randomized to one of three conditions: 4 h time-in-bed per day (4 h TIB), 6 h time-in-bed per day (6 h TIB), or 8 h time-in-bed per day (8 h TIB). Daily averages of psychomotor vigilance performance lapses were subjected to mixed-effects modeling using three different methods: standard two stage (STS), restricted maximum likelihood (REML), and nonlinear mixed-effects modeling (NMEM). For each experimental condition, the mean, standard deviation (SD), minimum (Min), and maximum (Max) are shown for the regression coefficients B_i (as per day increases in lapses per test bout). There are clear inter-individual differences within each condition, the the three methods yield very similar subject-specific results.

linear mixed-effects models that properly account for inter-individual variability in non-linear changes over time. Alternatively, nonlinear mixed-effects models can be used to accomplish this. Mixed-effects model approaches may have great utility in the development of predictive biomathematical models of fatigue and performance, as discussed in the next section.

Predictive Biomathematical Modeling of Inter-Individual Differences in Temporal Profiles

The development of biomathematical models to predict temporal changes in fatigue and performance involves two conceptually distinct steps (4): mathematical simulations to identify and verify the dynamic properties (i.e., equation types) of the biomathematical model ("direct problem"); and statistical fitting to assess the static properties (i.e., model parameters) of the biomathematical model ("inverse problem"). In an application of the Box iterative scheme for statistical model building (3), these two steps would be alternated to iteratively improve the biomathematical model as new data become available. Although this scheme may not have been used frequently in the development of existing models of fatigue and performance, it is essential to include the inverse problem in the development procedures. Without this step, the uncertainty in inferences based on the biomathematical models, and these models' sensitivity to errors in model specification and errors in parameter estimation cannot be properly evaluated (4). Both the direct problem and the inverse problem are more challenging when the phenomenon to be modeled displays large inter-individual differences. As demonstrated in the present paper (see also 10,42), models of fatigue and performance must take inter-individual differences into account in order to accurately describe and predict individuals' responses to sleep loss.

Mixed-effects models, and the NMEM approach in particular, are useful in the direct problem and in the inverse problem of biomathematical model building. For the inverse problem, the NMEM method can be applied to estimate model parameters more accurately as well as more informatively, because the population distributions of the model parameters are estimated explicitly*. An enticing example of NMEM methodology applied to the direct problem is found in Olofsen et al. (32), where it is described how the NMEM paradigm can be employed for subject-specific one-day-ahead predictions of performance deficits. In this kind of application, prior knowledge about the population distributions of parameter estimates (established using NMEM in the inverse problem) is incorporated in the biomathematical model and integrated with sequentially accumulating subject-specific information in order to individualize predictions of performance deficits in an on-going manner. Powerful computers and designated software (31,34,46) have recently become available for the computation of nonlinear mixed-effects models. Thus, the next generation of fatigue and performance models may take full advantage of mixed-effects modeling to deal with inter-individual differences.

Conclusions

In two experiments, we have demonstrated and quantified (using the intraclass correlation coefficient) the substantial, trait-like inter-individual differences in performance deficits resulting from up to 40 h of sleep loss. These inter-individual differences are so important (explaining more than 50% of total variance) that it is not reasonable to consider all individuals to be equal in their responses to sleep loss; thus, biomathematical models of sleep loss must be revised to include inter-individual differences. We have shown that for analyzing and modeling longitudinal data, inter-individual

*The NMEM approach for the inverse problem is also featured in Van Dongen (37), but in this work by necessity the models are treated as black boxes with fixed parameters. Thus, the statistical evaluations in Van Dongen provide an informative overview of current models' capabilities, but do not give direct insight into the models' sensitivities to model specification and parameter estimation errors, nor do these evaluations reveal precisely how the model equations and parameter estimates can be improved.

differences can be dealt with by means of mixed-effects models. It is feasible, therefore, to take inter-individual differences into account in predictive biomathematical models of fatigue and performance. New research aiming to identify predictors of inter-individual differences in the effects of sleep loss is urgently needed to support this task. In the meantime, development of a new generation of biomathematical models should begin to accommodate inter-individual variability.

ACKNOWLEDGMENTS

We are grateful for helpful suggestions from two anonymous reviewers.

This research was supported by NIH grants NR04281, HL70154, and RR00040, NASA grant NAG9-1161, and NASA cooperative agreements NCC 9–58 with the National Space Biomedical Research Institute, and NCC 2-1077 and 2-1394 with the Institute for Experimental Psychiatry Research Foundation.

REFERENCES

1. Åkerstedt T. Consensus statement: Fatigue and accidents in transport operations. J Sleep Res 2000; 9:395.
2. Belenky G, Wesensten NJ, Thorne DR, et al. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study. J Sleep Res 2003; 12:1–12.
3. Box GEP. Robustness in the strategy of scientific model building. In: Launer RL, Wilkinson GN, eds. Robustness in statistics. New York: Academic Press; 1979:201–36.
4. Brown EN, Luithardt H. Statistical model building and model criticism for human circadian data. J Biol Rhythms 1999; 14: 609–16.
5. Burton P, Gurrin L, Sly P. Extending the simple linear regression model to account for correlated responses: An introduction to generalized estimating equations and multi-level mixed modeling. Stat Med 1998; 17:1261–91.
6. Davidian M, Gallant RA. The nonlinear mixed effects model with a smooth random effects density. Biometrika 1993; 80:475–88.
7. Diggle PJ, Liang K-Y, Zeger SL. Analysis of longitudinal data. Oxford, England: Clarendon Press; 1996:64–8.
8. Dijkman M, Sachs N, Levine E, et al. Effects of reduced stimulation on neurobehavioral alertness depend on circadian phase during human sleep deprivation. Sleep Res 1997; 26:265.
9. Dinges DF. An overview of sleepiness and accidents. J Sleep Res 1995; 4:4–14.
10. Dinges DF, Achermann P. Future considerations for models of human neurobehavioral function. J Biol Rhythms 1999; 14:598–601.
11. Dinges DF, Douglas SD, Zaugg L, et al. Leukocytosis and natural killer cell function parallel neurobehavioral fatigue induced by 64 hours of sleep deprivation. J Clin Invest 1994; 93:1930–9.
12. Dinges DF, Kribbs NB. Performing while sleepy: Effects of experimentally-induced sleepiness. In: Monk TH, ed. Sleep, sleepiness and performance. Chichester, England: John Wiley & Sons; 1991:97–128.
13. Dinges DF, Powell JW. Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. Behav Res Methods Instr Comp 1985; 17:652–5.
14. Doran SM, Van Dongen HPA, Dinges DF. Sustained attention performance during sleep deprivation: Evidence of state instability. Arch Ital Biol 2001; 139:253–67.
15. Dudewicz EJ, Mishra SN. Modern mathematical statistics. New York: John Wiley & Sons; 1988.
16. Dunn G. Design and analysis of reliability studies. Stat Methods Med Res 1992; 1:123–57.
17. Feldman HA. Families of lines: Random effects in linear regression analysis. J Appl Physiol 1998; 64:1721–32.
18. Fleiss JK. The design and analysis of clinical experiments. New York: John Wiley & Sons; 1986.
19. Folkard S, Åkerstedt T. Towards the prediction of alertness on abnormal sleep/wake schedules. In: Coblentz A, ed. Vigilance and performance in automated systems. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1989:287–96.
20. Gibbons RD, Hedeker D, Waternaux C, Davis JM. Random regression models: A comprehensive approach to the analysis of longitudinal psychiatric data. Psychopharmacol Bull 1988; 24: 438–43.
21. Harrison Y, Horne JA. The impact of sleep deprivation on decision making: A review. J Exp Psychol Appl 2000; 6:236–49.
22. Hoddes E, Zarcone V, Smythe H, et al. Quantification of sleepiness: A new approach. Psychophysiol 1973; 10:431–6.
23. Huynh H, Feldt LS. Estimation of the Box correction for degrees of freedom from sample data in the randomized block and split plot designs. J Educat Stat 1976; 1:69–82.
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159–74.
25. Leproult R, Colecchia EF, Berardi AM, et al. Individual differences in subjective and objective alertness during sleep deprivation are stable and unrelated. Am J Physiol Regulat Integr Comp Physiol 2003; 284:R280–90.
26. Lin LIK. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989; 45:255–68.
27. Littell RC, Pendergast J, Natarajan R. Tutorial in biostatistics: Modeling covariance structure in the analysis of repeated measures data. Stat Med 2000; 19:1793–1819.
28. Mallis MM, Mejdal S, Nguyen TT, Dinges DF. Summary of the key features of seven biomathematical models of human fatigue and performance. Aviat Space Environ Med 2004; 75(3, Suppl.): A4–14.
29. Mitler MM, Miller JC, Lipsitz JJ, et al. The sleep of long-haul truck drivers. New Engl J Med 1997; 337:755–61.
30. Morrison DF. Multivariate statistical methods. 2nd ed. New York: McGraw Hill; 1976.
31. NONMEM Project Group. NONMEM user's guide. San Francisco, CA: University of California at San Francisco; 1999.
32. Olofsen E, Dinges DF, Van Dongen HPA. Nonlinear mixed-effects modeling: Individualization and prediction. Aviat Space Environ Med 2004; 75(3, Suppl.):A134–40.
33. Patterson HD, Thompson R. Recovery of inter-block information when block sizes are unequal. Biometrika 1971; 58:545–54.
34. Pinheiro JC, Bates DM. Mixed-effect models in S and S-PLUS. New York: Springer-Verlag; 2000.
35. SAS Institute Inc. SAS OnlineDoc®, version 8. Cary, NC: SAS Institute Inc.; 1999.
36. Scheffe H. The analysis of variance. Chichester, England: John Wiley & Sons; 1959.
37. Van Dongen HPA. Comparison of mathematical model predictions to experimental data of fatigue and performance. Aviat Space Environ Med 2004; 75(3, Suppl.):A15–36.
38. Van Dongen HPA, Baynard MD, Nosker G, Dinges DF. Repeated exposure to total sleep deprivation: Substantial trait differences in performance impairment among subjects. Sleep 2002; 25:A89–90.
39. Van Dongen HPA, Dijkman MV, Maislin G, Dinges DF. Phenotypic aspect of vigilance decrement during sleep deprivation. Physiologist 1999; 42:A-5.
40. Van Dongen HPA, Dinges DF. Circadian rhythms in fatigue, alertness and performance. In: Kryger MH, Roth T, Dement WC, eds. Principles and practice of sleep medicine. 3rd ed. Philadelphia, PA: W.B. Saunders; 2000:391–9.
41. Van Dongen HPA, Maislin G, Mullington JM, Dinges DF. The cumulative cost of additional wakefulness: Dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. Sleep 2003; 26:117–26.
42. Van Dongen HPA, Rogers NL, Dinges DF. Sleep debt: Theoretical and empirical issues. Sleep Biol Rhythms 2003; 1:5–13.
43. Webb WB, Levy CM. Effects of spaced and repeated total sleep deprivation. Ergonomics 1984; 27:45–58.
44. Wilkinson RT. Interaction of lack of sleep with knowledge of results, repeated testing, and individual differences. J Exp Psychol 1961; 62:263–71.
45. Winer BJ. Statistical principles in experimental design. 2nd ed. New York: McGraw Hill; 1971.
46. Wolfinger RD. Fitting nonlinear mixed models with the new NLMIXED procedure. Cary, NC: SAS Institute Inc.; 2000.
47. Zeger SL, Liang K-Y, Albert PS. Models for longitudinal data: A generalized estimating equation approach. Biometrics 1988; 44:1049–60.