Assessment of the Unified Model of Performance: Accuracy of Group-Average and Individualized Alertness Predictions

Nikolai V. Priezjev[1,2], Francisco G. Vital-Lopez[1,2], and Jaques Reifman[1*]

[1]Department of Defense Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, United States Army Medical Research and Development Command, Fort Detrick, MD, USA

[2]The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., Bethesda, MD, USA

*Correspondence: Jaques Reifman, Ph.D., Senior Research Scientist and Director
Department of Defense Biotechnology High Performance Computing Software Applications Institute
Telemedicine and Advanced Technology Research Center
U.S. Army Medical Research and Development Command
ATTN: FCMR-TT, 504 Scott Street, Fort Detrick, MD 21702-5012
Phone: (301) 619-7915; Fax: (301) 619-1983; E-mail: jaques.reifman.civ@mail.mil

Short title: Assessment of the Unified Model of Performance

Total number of words (abstract and main text): 6978
Total number of references: 34

DISCLOSURE STATEMENT: This was not an industry-supported study. FGVL and JR receive royalties for the licensing of the *2B-Alert* technology to Integrated Safety Support. The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army, the U.S. Department of Defense, or The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc. This paper has been approved for public release with unlimited distribution.

AUTHOR CONTRIBUTORSHIP: JR conceived the research. NVP and FGVL performed the computations. NVP and JR wrote the manuscript. All authors have reviewed the manuscript and approved the submitted version.

ABSTRACT

To be effective as a key component of fatigue-management systems, biomathematical models that predict alertness impairment as a function of time of day, sleep history, and caffeine consumption must demonstrate the ability to make accurate predictions across a range of sleep-loss and caffeine schedules. Here, we assessed the ability of the previously reported Unified Model of Performance (UMP) to predict alertness impairment at the group-average and individualized levels in a comprehensive set of 12 studies, including 22 sleep and caffeine conditions, for a total of 301 unique subjects. Given sleep and caffeine schedules, the UMP predicted alertness impairment based on the psychomotor vigilance test (PVT) for the duration of the schedule. To quantify prediction performance, we computed the root mean square error (RMSE) between model predictions and PVT data, and the fraction of measured PVTs that fell within the models' prediction intervals (PIs). For the group-average model predictions, the overall RMSE was 43 ms (range from 15 to 74 ms) and the fraction of PVTs within the PIs was 80% (range from 41 to 100%). At the individualized level, the UMP could predict alertness for 81% of the subjects, with an overall RMSE of 64 ms (average range from 32 to 147 ms) and fraction of PVTs within the PIs conservatively estimated as 71% (range from 41 to 100%). Altogether, these results suggest that, for the group-average model and 81% of the individualized models, in three out of four PVT measurements we cannot distinguish between study data and model predictions.

Keywords: alertness prediction model; fatigue; neurobehavioral performance; psychomotor vigilance test; sleep deprivation

INTRODUCTION

Public and private sectors can use biomathematical fatigue models to help design, compare, and contrast work schedules for teams of workers as well as to provide individualized guidance for optimizing the use of fatigue countermeasures (Powell et al., 2014; Reifman et al., 2019, 2022; Vital-Lopez et al., 2021; Integrated Safety Support, 2022). As may be expected, decision-makers and users of such fatigue-management tools generally assume that the underlying mathematical models driving these tools have been well validated and peer-reviewed before they come to market or become publicly available. However, the opposite more closely reflects common practice.

Over the last decade, only two studies have performed side-by-side comparisons among multiple biomathematical fatigue-prediction models (Hilaire et al., 2017; Flynn-Evans et al., 2020). While such analyses are invaluable, especially if performed by independent reviewers, these studies compared and contrasted the models against a single study condition, which is insufficient to gauge model performance across a broad range of sleep/rest schedules reflective of everyday life, for which the tools are expected to be used. One exception is the work of Powell et al. (2014), which attempted to validate the System for Aircrew Fatigue Evaluation model for 11 distinct commercial flight operations. Unfortunately, as detailed by the authors, a number of simplifying assumptions regarding the inputs to the model may have contributed to the low agreement between model predictions and recorded quantitative measures of fatigue. In addition, because this model does not account for fatigue countermeasures, their assessment did not consider the

beneficial effects of caffeine, the most widely used stimulant compound consumed daily by ~85% of the U.S. population (Mitchell et al., 2014).

Over the past 15 years, our group at the U.S. Army has been incrementally developing and enhancing the Unified Model of Performance (UMP), which predicts alertness impairment—as determined by the psychomotor vigilance test (PVT)—at the group and individual levels, as a function of sleep history, time of day, and caffeine consumption (Rajdev et al., 2013; Liu et al., 2017; Ramakrishnan et al., 2016a and 2016b). As we enhanced the UMP over time, we continually validated the model predictions using an array of total sleep deprivation (TSD) and chronic sleep restriction (CSR) conditions (Ramakrishnan et al., 2016a) as well as a diverse set of caffeine-consumption schedules (Ramakrishnan et al., 2016b). However, we described such model validations in different reports, sometimes with slight changes in the model or model parameter values, involving a limited number of conditions. As we have now completed model enhancements and frozen the model, culminating in the development of the Web- and smartphone-based *2B-Alert* tools (Reifman et al., 2019 and 2022), we sought to perform a thorough validation of the UMP. To this end, we assessed its ability to predict alertness impairment at the group-average level and at the individual-specific level across a broad range of sleep and caffeine schedules (22 from 12 different studies), involving a total of 301 unique subjects (244 at the individual level). Given two inputs (sleep and caffeine schedules), the UMP predicted alertness impairment, as measured by the mean response time (RT) in the PVT. To assess prediction performance at the group-average level, we used two metrics: the root mean square error (RMSE) between the model predictions and the group-average PVT data and the fraction of the data that fell within the prediction intervals (PIs) of the model. Similarly, to assess the prediction performance at the individual level, we computed these two metrics using the

corresponding individual-specific model predictions and the individuals' PVT data. We used the fraction of the PVT data that fell within the PIs of the models to estimate the extent to which the UMP predictions were indistinguishable from the PVT measurements.

METHODS

Datasets for assessment of model performance

To validate the group-average and individualized models, we used the mean RT in the PVT as a measure of alertness impairment collected from 12 studies (V1–V12). These involved a total of 301 unique subjects and 22 diverse conditions, including 14 distinct sleep schedules and seven caffeine conditions. Table 1 provides a brief description of the 22 study conditions, including the participant's sex, PVT duration and number collected in each condition, sleep schedule, including baseline, type of sleep challenge, and recovery, as well as caffeine-consumption schedule. The sleep schedules typically included several baseline days with either habitual sleep [7–8 h of time in bed (TIB)] or extended sleep (10 h of TIB), a sleep challenge period involving either CSR, TSD, or both, and a recovery phase of 8–24 h of TIB, for one to five consecutive nights. The studies reflect PVT data collected in both laboratory (V1–V5, V7, and V10–V12) and field (V6, V8, and V9) conditions, with three studies (V1, V3, and V8) using a cross-over design, in which the same subject performed PVTs under two different sleep or caffeine conditions. At different stages of model development throughout the years, we did use data from three of the 12 studies (V3, V7, and V12) to develop or optimize earlier versions of the models developed at those times (Rajdev et al., 2013; Ramakrishnan et al., 2013; Liu et al., 2017). As described below, the actual parameter values of the group-average model validated herein were derived using data from two different, earlier studies.

< Table 1 here >

Unified model of performance

Based on the two-process model postulated by Borbély and Achermann (1999), we previously developed the UMP to predict the temporal patterns of alertness for conditions ranging from CSR to TSD challenges (Rajdev et al., 2013), extending the two-process model in three ways. First, by taking into account prior sleep-wake history, the UMP modulates alertness impairment and recovery as a function of sleep debt, leading to a relatively slow decrease in impairment rates after extended sleep, i.e., sleep banking (Rupp et al., 2009), and slow recovery rates after CSR. Second, the UMP accounts for the alertness-enhancing effects of caffeine by assuming that it has a multiplicative effect on alertness throughout the sleep-wake cycle (Ramakrishnan et al., 2016b). That is, the UMP predicts alertness impairment $P(t)$ at time t after consumption of caffeine dose c, as follows:

$$P(t) = P_0(t) \times g_{PD}(t, c), \quad (1)$$

where $P_0(t)$ represents the alertness impairment predicted at time t in the absence of caffeine and $g_{PD}(t, c)$ denotes the pharmacodynamic (PD) effect of caffeine, which varies from 0 to 1, where the upper bound 1 corresponds to PD effects in the absence of caffeine and the theoretical lower bound 0 represents the maximal PD effect on alertness impairment. In this formulation, the effect of caffeine is greater when the alertness impairment is higher, in accordance with experimental studies (Rétey et al., 2006; Landolt et al., 2012). Table 2 shows the UMP equations governing the caffeine-free portion of the model [$P_0(t)$] in Eqs. (2)–(5) and the caffeine effect [$g_{PD}(t, c)$] in Eq. (6). Overall, the UMP has a total of 12 model parameters, eight for the caffeine-free portion

of the model, which we obtained by fitting the parameters to the group-average PVT data in the study by Belenky et al. (2003), and four to represent the effects of caffeine (Ramakrishnan et al., 2016b), which we obtained by fitting the caffeine parameters to the group-average PVT data in the study by Kamimori et al. (2005). Table S1 shows all parameter values for the group-average model (see Supporting Information).

< Table 2 here >

Third, in addition to predicting the average alertness of a collection of individuals in a "group-average" model, the UMP can be customized to predict alertness impairment of a specific individual in an "individualized" model (Ramakrishnan et al., 2015; Reifman et al., 2019). We develop an individualized model for a subject by customizing the model parameters of the caffeine-free portion of the UMP so that they reflect the subject's response to sleep deprivation measured by the PVT (Liu et al., 2017). Model customization starts by assuming that the subject has an average response to sleep deprivation and, initially, can be represented by the parameters of the group-average model. Then, after each PVT, we customize the model by recursively adjusting its parameters using a Bayesian-learning approach, where we balance the weight of the latest PVT measurement [second term in Eq. (7) in Table 3] against that of the group-average model [i.e., the prior, the first term in Eq. (7)]. As the number of PVT measurements increases, the weight of the latest PVT increases, leading to an individualized model that represents the individual's sleep-loss phenotype (Liu et al., 2017). The model parameters are recursively updated after each PVT by solving two algebraic equations [Eqs. (8) and (9) in Table 3], where only five UMP parameters (see Table 2) need to be customized (Ramakrishnan et al., 2015).

< Table 3 here >

Learning an individual's sleep-loss trait

To determine whether the UMP recursively learned an individual's sleep-loss trait after *n* PVTs, we computed the difference between RMSEs of a recursively learned model and the best-fit model:

$$\Delta RMSE_n = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[y_i - P_0(t_i, \hat{\theta}_n)]^2} - \sqrt{\frac{1}{N}\sum_{i=1}^{N}[y_i - P_0(t_i, \theta^*)]^2}\,, \qquad (10)$$

where y_i, i = 1, 2,…, *n*,..., *N*, represents the complete set of *N* PVTs, $\hat{\theta}_n$ denotes the recursively estimated model parameters after *n* PVTs, and θ^* indicates the parameters of the best-fit model, defined as the solution of Eq. (7) obtained using all *N* PVTs of an individual, for the corresponding study condition. We assumed that the model "learned" the sleep traits of an individual after *n* PVTs when $\Delta RMSE_n$ reached and remained <5 ms with increasing numbers of PVTs. This arbitrary threshold guaranteed near-optimal predictions, while retaining sufficient remaining PVTs (10% of *N* or 5, whichever was smaller) to independently assess the goodness of the learned model to predict an individual's trait.

In addition, we assessed the ability of the individualized model to learn a subject's sleep-loss trait with a reduced set of PVT measurements. For this analysis, we focused on the five CSR study conditions of 3 h (V2a, V2b, and V3a), 5 h (V10a), and 6 h (V1b) of sleep per night in Table 1, because they more closely resembled everyday sleep-deprivation conditions and provided a sufficient number of PVTs to train the individualized models (at least five consecutive CSR nights). We did not consider CSR conditions with caffeine consumption to reduce potential confounding factors. For the selected study conditions, we obtained an

individualized model for each subject using a subset of the available PVT measurements and compared the performance of these models with those of the corresponding best-fit models.

Assessment of model performance

To assess the performance of the group-average model predictions for each study condition, we used two metrics. First, we calculated the RMSE between the model-predicted alertness impairment $P(t)$ in Eq. (1) and the mean PVT for each session. Second, we estimated PIs around $P(t)$ and computed the fraction of PVT data that fell within the PIs. For the group-average predictions, we defined PI_j for PVT session j in a study condition, as follows:

$$\mathrm{PI}_j = P_j(t) \pm z\sqrt{\sigma_{fit}^2 + \sigma_{sc,j}^2}\,, \tag{11}$$

where $P_j(t)$ denotes the group-average model prediction for PVT session j, z (≈1.96) represents the standard score for 95% confidence level, and the uncertainty term comprises two components: the standard deviation (σ_{fit} = 26 ms) in the PVT data of the study used to estimate the parameters of the group-average model (Belenky et al., 2003) and the standard deviation of the mean PVT ($\sigma_{sc,j}$) in session j, for the study condition we wish to predict. This definition of PI is more stringent than the alternative of separately computing PIs for the model prediction (based on σ_{fit}) and for the PVT data (based on $\sigma_{sc,j}$), and determining if they overlap [because $(\sigma_{fit}^2 + \sigma_{sc,j}^2)^{1/2} < (\sigma_{fit} + \sigma_{sc,j})$].

To assess the performance of the individualized model predictions, for each of the 301 unique subjects across the 22 study conditions, we first determined whether the subject's individualized model was capable of learning the subject's sleep-loss trait using a subset of the study-condition

data. We labeled a subject as "learnable" if the model predictions for the subject satisfied two arbitrary criteria: *1*) the percentage of PVT measurements that fell within the PIs of the individualized model predictions using the total number of measurements for the subject was >50% and *2*) after learning a subject using a subset of the study-condition data, there were sufficient remaining PVT measurements to assess the performance of the model predictions (at least 5 PVTs or 10% of the total number of PVT measurements in the study condition). Subjects who did not meet these criteria were labeled as "not-learnable." The first criterion speaks to the variance in the data. If the variance is too large (i.e., the PVT measurements have too much variability), given the limited number of adjustable parameters in the model, it does not have sufficient degrees of freedom to capture the sleep-loss trait of the subject. This criterion attempts to capture the fact that even if we were to use *all* PVT measurements in the study condition to *fit* the model to the data, it would still be an inadequate model because more than 50% of the data were not within the PIs. Notably, our analyses included *all* reported data from each study, even though we observed obvious outliers in a number of subjects across the studies. The second criterion is to guarantee that we have sufficient PVT measurements not used in model training to assess the performance of the model predictions. (For the cross-over design studies, we required these conditions be met for both arms of the study for a subject to be labeled as learnable.)

For the individualized models that met these criteria, we assessed their performance using equivalent metrics as in the group-average case. Using the PVT data for the sessions *after* the model had learned the subject's trait, we computed the RMSE between the model-predicted alertness impairment $P(t)$ in Eq. (1) and the subject's measured PVT, and calculated the fraction of the PVT data that fell within the PIs. For each subject, we defined PI_j for each PVT session *j* in each study condition, as follows:

$$\mathrm{PI}_j = P_j(t) \pm z\,\sigma_{ws}\,, \tag{12}$$

where $P_j(t)$ denotes the individualized prediction for the *j* PVT session, *z* = 1.96 as above, and σ_{ws} denotes the variance of the PVT data upon repeated measurements by the same subject under the same condition, i.e., a measure of within-subject variability. In the absence of repeated data, to err on the side of caution, we assumed σ_{ws} to be ~30 ms (Khitrov et al., 2014) for all subjects in all study conditions. This is a conservative estimate of σ_{ws} because it was obtained for subjects under well-rested conditions and σ_{ws} is known to be larger under sleep-loss conditions (Rupp et al., 2012). For both the group-average and the individualized model predictions, the larger the fraction of PVT data laying within the corresponding PI, the higher the accuracy of the UMP prediction.

Importantly, choosing a sufficiently wide PI can artificially inflate the number of predictions that fall within the interval, making it an inappropriate metric of performance. Precisely for this very same reason, by design, the PIs we used in our analyses were based *solely* on the underlying variability of the *PVT data*, rather than on the uncertainty of the model predictions. For example, for the group-average predictions, in addition to the variability in the data used to develop the model, the width of the PI depended on the variability of the study-condition data we wished to predict. Hence, if the measured PVT data in the study we wished to predict had large variability, then the PIs would be wider, as one would expect. In fact, had we designed a new study to reproduce the original experimental study, we would expect the results to fall within these very same PIs with a 95% confidence level.

We estimated the parameters for the group-average model using 10-min PVT data. Thus, for study conditions that used a 10-min PVT, we directly compared the group-average or individualized predictions with the data. For study conditions that used a 5-min PVT (marked with an asterisk in the first column in Table 1), we first obtained predictions for the 10-min PVT, converted the 10-min PVT predictions into an equivalent 5-min PVT prediction using an affine transformation (Hastie et al., 2001), and then compared the equivalent 5-min PVT predictions with the 5-min PVT data. The individualized models required an additional pre-processing step, where we first transformed the 5-min PVT data into 10-min PVT data, using the inverse affine transformation, before performing the steps above. We refer the reader to the Supporting Information for a detailed description of the affine transformation. In addition, we observed between-study differences in the PVT data across the study conditions. To normalize these differences, we added a constant value δ to the group-average predictions for each study condition, where δ was computed as the average difference between UMP predictions and PVT data within the first 16 h of wakefulness on the first day of the sleep-deprivation challenge (Ramakrishnan et al., 2016a).

RESULTS

We validated the group-average and individualized models by comparing their predictions against measured PVT data in the 12 studies (V1–V12). Table 4 summarizes the prediction performance for the 301 unique subjects under the 22 different sleep and caffeine-consumption conditions described in Table 1. Overall, the group-average model predictions demonstrated that the UMP captured the alertness impairment at the population level for a wide range of sleep and caffeine conditions, with an average RMSE between the UMP predictions and the PVT data of 43 ms, ranging from 15 ms in study condition V8a to 74 ms in V3b. Importantly, 80% of the

predictions fell within the PI in Eq. (11), suggesting that the majority of the group-average predictions were indistinguishable from the mean of the observed PVT measurements. This fraction ranged from 63% (V6a, V12a, and V6b) to 100% in four study conditions (V1b, V2b, V8b, and V10b), except for study V3a (41%).

< Table 4 here >

Table 4 also shows the prediction performance of the individualized model for 244 unique subjects (out of 301) who were learnable (i.e., had >50% of PVT data within the PIs) and had a sufficient number of PVTs (at least 5 or 10% of *N*) for independent assessment after the model had learned the subject's trait. Of the 57 (301 minus 244) not-learnable subjects, 42 did not meet the 50% criterion and 15 did not have a sufficient number of PVTs for model assessment (see Methods). Overall, the individualized models captured the alertness-impairment levels with varying accuracy, with RMSEs ranging from 32 ms in study condition V6b to 147 ms in V11a, for an average prediction error of 64 ms and 71% of the predictions falling within the PI in Eq. (12). Importantly, our analyses included *all* PVT data for each subject, regardless of how inaccurate they seemed, including obvious outliers.

We also computed the number of PVTs required by the individualized model to learn an individual's trait-like response to sleep deprivation and caffeine consumption for the same 22 conditions. Table 4 shows that the average number of PVTs needed to individualize the model parameters was 27, with 25% of the models requiring up to 33% of the data to learn a subject and 75% requiring up to 80% of the data. Overall, the number of PVTs required for the model to learn a subject depended on the study condition and varied between subjects.

Figures 1–4 show the observed PVT mean RT data along with the group-average and individualized model predictions for four study conditions that included CSR and TSD sleep challenges, with and without caffeine. Figure 1 shows the results for the CSR plus TSD challenge in study condition V1b (Table 1), consisting of two baseline nights of 8 h of TIB per night, seven nights of 6 h of TIB per night, followed by 41 h of TSD. Figure 1a shows the group-average data and corresponding model predictions, where the model yielded very accurate predictions with a RMSE of 28 ms and 100% of the experimental data falling within the PIs. Figures 1b–1d show the PVT data and individualized model predictions for three subjects in V1b with distinct sleep-loss phenotypes and model-prediction accuracies. The mean RT data for subject #20 showed negligible variation and consistently low alertness impairment throughout the multiple phases of the study, suggesting that this subject was resilient to sleep deprivation. The individualized model learned this subject by CSR day C3, after 19 PVTs (open circles, Fig. 1b), and accurately predicted alertness impairment throughout the remaining CSR days and TSD challenge (filled circles, RMSE = 37 ms and fraction = 93%). In contrast, subject #13 showed considerable PVT variability over the last day of the TSD challenge (T1, Fig. 1d). It took 50 PVTs for the individualized model to learn this subject (open circles), and the predictions that followed yielded a large RMSE (231 ms) with none of the data (filled circles) falling within the PIs. Note that the data showed clear outliers on day T1.

< Fig. 1 here >

Figure 2 shows the mean RT data as well as the group-average and individualized model predictions for study condition V11b, consisting of one baseline night of 7 h of TIB followed by 54.5 h of TSD, with a single 600-mg dose of caffeine at 23:50 after 41.5 h into the challenge. In this case, the group-average model predictions in Fig. 2a yielded results slightly worse than the

average model performance over the 22 conditions (RMSE = 51 ms vs. 43 ms and fraction = 76% vs. 80%). As expected, alertness impairment temporarily improved immediately after caffeine consumption at the end of challenge day T1, and the model accurately captured this behavior. Figures 2b–2d show the results for three subjects where the individualized model learned each subject on challenge day T2 after caffeine consumption (open circles), yielding varying degrees of accuracy (RMSE from 68 ms to 103 ms and fraction from 50% to 20%). The individualized model learned subject #10 slightly faster and more accurately than the other two subjects, who were more vulnerable (#3) or whose data showed greater variability (#3 and #9).

< Fig. 2 here >

Figure 3 shows the mean RT data and model predictions for study condition V2b, which consisted of eight baseline nights with 8 h of TIB and a CSR phase with 3 h of TIB for seven nights, followed by a recovery phase of 8 h of TIB for five nights. In this case, the group-average model predictions agreed very well with the experimental data (RMSE = 26 ms and fraction = 100%, Fig. 3a). Figures 3b–3d illustrate the effect of PVT variability in the model's ability to learn a subject's trait and its prediction accuracy, for three subjects in this study. The PVT data for subject #11 showed minor changes from day to day and little variability, resulting in quick model learning after only 19 PVTs by CSR day C2 and excellent predictions thereafter (Fig. 3b). In contrast, the data for subjects #3 and #12 showed more variability, resulting in a larger number of PVTs to learn the subjects' traits and, for subject #12, yielding a relatively low prediction accuracy (RMSE = 118 ms and fraction = 65%, Fig. 3d), likely due to the large amount of scatter in the PVT data during the last three days of CSR and the first recovery day, including a few outliers.

< Fig. 3 here >

Figure 4 shows the results for study condition V10b, where, after five nights of extended sleep (10 h of TIB), subjects were challenged with five nights of 5 h of TIB per night, followed by three recovery nights of 8 h of TIB. Subjects consumed 200 mg of caffeine twice a day at 08:00 and 12:00 during the CSR phase. In this case, the group-average predictions yielded excellent agreement with the experimental data (RMSE = 19 ms and fraction = 100%, Fig. 4a). Nevertheless, we did observe variability in the individualized predictions and the number of PVTs required to learn each subject. For example, while the individualized model learned subject #6 by CSR day C3 and yielded excellent agreement with the experimental data (RMSE = 31 ms and fraction = 98%, Fig. 4b), the model only learned subject #8 after 3.5 days of CSR and yielded relatively low performance (RMSE = 98 ms and fraction = 43%, Fig. 4d). In this case, the individualized model underpredicted alertness impairment during the last day of CSR and the first two days of recovery, where the experimental data were scattered.

< Fig. 4 here >

We also assessed the ability of the individualized models to learn a subject's sleep-loss trait with a reduced set of PVT measurements. We focused on the CSR conditions (V1b, V2a, V2b, V3a, and V10a; see Table 1) because they more closely resembled everyday sleep-deprivation conditions. Interestingly, we found that by using only two PVTs per day (taken at 10:00 and at around 18:30) for five consecutive days, for a total of 10 PVTs, the individualized models were able to learn at least 70% of the learnable subjects in each study condition. For these subjects, the individualized models yielded RMSEs slightly larger (<11 ms) than those of the best-fit models.

In particular, for two study conditions (V2b and V3a), the fraction of subjects learned with 10 PVTs was at least 80%.

DISCUSSION

To be useful, mathematical models must be able to accurately predict an individual's neurobehavioral performance as a function of time of day, sleep history, and caffeine consumption across a wide range of sleep and caffeine-consumption conditions. They must also be able to capture an individual's trait-like response to sleep deprivation, so as to account for the large between-subject variability (Van Dongen et al., 2004; Rupp et al., 2012). In this work, we validated the UMP, demonstrating its ability to adequately represent alertness impairment of a population as well as of a specific individual across a comprehensive set of 22 distinct conditions spanning the continuum of sleep loss and caffeine consumption.

Overall, the group-average UMP accurately predicted the mean PVT response for each sleep and caffeine condition, including seven caffeine-dosing schedules and 14 distinct sleep-deprivation conditions (Table 1). Quantitatively, we showed that the mean RMSE between the group-average predictions and the study-average PVT data across the 22 conditions and 301 unique subjects was 43 ms (range = 15 to 74 ms), whereas the fraction of PVTs within the PIs was relatively high (80%; range = 41 to 100%), suggesting that in 80% of the cases the UMP predictions were indistinguishable from group-average PVT measurements. The fraction was lower than 60% for only one study condition (V3a), which involved an extended sleep period during the baseline phase, followed by seven nights of 3 h of TIB and a recovery phase (Table 4). In this case, the relatively low fraction (41%) is attributed to an abnormally small decrement in alertness impairment during CSR [in comparison to the decrement in similar sleep schedules in V2a

(fraction = 69%) and V10a (fraction = 97%)], which resulted in an overprediction by the group-average model. The fractions for all other study conditions were greater than 63%, and reached 100% for four conditions, indicating accurate predictions across a diverse set of sleep and caffeine schedules. To characterize any potential systematic error not captured by the RMSE or fraction metrics, we carried out an analysis of residuals (the difference between PVT measurements and predictions; see Supporting Information). The analysis showed that the residuals for the combined 22 conditions appeared normally distributed and without any systematic patterns other than an average over-prediction of the measurements by 14 ms, which corresponds to 20% of the average half-width of the PIs. Thus, there was a positive bias in the predictions, but the bias was relatively small, confirming the validity of the group-average model. Although we observed obvious PVT outliers in a number of subjects across the studies, our analysis included *all* reported data from each study.

We also assessed the ability of the UMP to learn the sleep-loss trait of each of the 301 unique individuals across the 22 study conditions. For the 81% of the subjects (244 out of 301) deemed to be learnable (see Methods), we assessed model performance by comparing the individualized model predictions against each subject's PVT data (Table 4). Overall, the average RMSE across all predictions was 64 ms (average range per study = 32 to 147 ms), where, on average, 71% of the PVTs of each subject in each session of each study condition fell within two standard deviations of the within-subject variability (=59 ms for well-rested conditions; Khitrov et al., 2014). This suggests that, for these 244 subjects, in nearly three out of four PVTs we cannot distinguish between a single PVT measurement and the individualized model prediction. This is a conservative estimate because the within-subject variability is known to increase with sleep loss (Rupp et al., 2012). Assuming a 25% increase in within-subject variance (from 59 to 74 ms)

during the sleep challenge phase of the studies would have increased the number of PVTs falling within this error bar to ~80%. To characterize any potential systematic error in the individualized model predictions, we also carried out an analysis of residuals (see Supporting Information). The analysis showed that the residuals for the combined 244 subjects in the 22 conditions appeared normally distributed and without systematic patterns other than an average over-prediction of the measurements by 6 ms, which corresponds to 10% of the half-width of the PIs. Thus, there was a positive bias in the individualized predictions, but the bias was relatively small, confirming the validity of the individualized models. As in the group-average predictions, our analysis used *all* reported data, including obvious outliers.

The UMP did not capture the sleep-loss traits of 57 (or 19%) of the 301 unique subjects. These subjects were not-learnable primarily because of the excessive variability in the PVT data (42 subjects) and because this variability slowed the learning process, not leaving sufficient data to assess the model predictions (15 subjects). To characterize the variability in the data used for validation between learnable and not-learnable subjects, we observed that the standard deviation of the data for the not-learnable subjects was 270% larger. Notably, the variability was concentrated in a few studies. For example, 55% (or 23 subjects) of the 42 not-learnable subjects discussed above came from only two studies [V1 (15) and V9a (8)], where the data showed excessive variability. In addition, the model does not have enough degrees of freedom to fit any one individual perfectly, even if we were to use all available data to fit the model to the data (i.e., to develop a best-fit model). In fact, developing individualized best-fit models for these 57 subjects and computing the performance metrics for the smallest of the last 5 PVTs, or 10% of the total number of PVTs available for each subject, yielded an average RMSE of 135 ms and an

average fraction of 32%. In contrast, for the learnable 81% of the subjects, we obtained an average RMSE of 64 ms and an average fraction of 71%.

Because one of the motivations to develop the UMP was to bridge the continuum from CSR to TSD with a single model, we investigated whether there were differences in the performance of the models between the TSD studies (11 conditions) and the CSR studies (5 conditions) in Table 4. For the group-average predictions, the average RMSE was 44 ms (standard deviation = 21 ms) and 36 ms (20 ms), and the average fraction of PVTs within the PIs was 80% (10%) and 81% (26%) for the TSD and CSR conditions, respectively. Similarly, for the individualized predictions, the average RMSE was 72 ms (31 ms) and 53 ms (4 ms) and the average fraction of PVTs within the PIs was 64% (20%) and 82% (6%) for the TSD and CSR conditions, respectively. Based on two-sample *t*-tests, there were no statistical differences at the 0.05 significance level in the performance metrics between the TSD and CSR conditions, for either the group-average or individualized model predictions, confirming one of the distinctive features of the UMP, the ability to bridge the continuum of sleep loss with a single model.

Although the overall results for the individualized model were similar to those of the group-average model, the latter cannot accurately predict each specific individual unless the alertness impairment is close to that of the "average" subject (Liu et al., 2017). To assess the benefit of model individualization, we computed the RMSE between each subject's PVT data and the group-average model predictions for the same subset of PVT measurements used for validating the individualized models. The average RMSE across the 22 study conditions was 88 ms (vs. 64 ms), a 38% increase in prediction error, demonstrating that model customization produced more accurate predictions at the individual level. In general, the individualized model failed to accurately capture a subject's trait-like response to sleep loss when the subject's PVT data

showed large variability, resulting in lower prediction accuracy, for example, as for subject #8 in study condition V10b (Fig. 4d).

We expected the number of PVTs required to individualize the model parameters to depend on the individual's sleep-loss phenotype, sleep schedule, caffeine consumption, as well as the frequency and timing of PVT sessions. For example, while it took on average 53 PVTs to learn the subjects in study condition V10a (placebo), it took only 36 PVTs in the caffeine arm of the study (V10b). To assess the effect of some of these factors, we plotted the cumulative distribution of the percentage of subjects in a given study condition learned by the model as a function of the number of PVTs needed to individualize the UMP parameters. However, we could only compare nine of the 12 studies in Fig. 5 where the study had two arms that differed by only one factor (Table 1): V2 (baseline sleep of 8 h vs. 10 h of TIB), V3 (CSR vs. TSD), and V6–V12 (caffeine vs. placebo). As expected, the model was able to learn subjects with shorter baseline TIB duration (Fig. 5b) and more acute sleep deprivation (Fig. 5c) faster, consistent with previous results (Liu et al., 2017). We could not reach a conclusion on the comparison of caffeine vs. placebo (Figs. 5f–5l), primarily because the model only learned some subjects after caffeine consumption. Nevertheless, we found variability in PVT data to be the chief factor driving the speed of model individualization. As a result, because PVT measurements of resilient subjects have consistently low variability (see, for example, Figs. 1b–4b), the individualized model learned the traits of resilient subjects considerably faster than those of more vulnerable subjects.

< Fig. 5 here >

Although the PVT is widely used to assess alertness in sleep studies, the task is rather tedious and time consuming. Therefore, to determine if it is possible to reduce this burden, we investigated whether the individualized models could learn the sleep-loss trait of an individual using only a subset of PVT measurements. Interestingly, for the five CSR conditions without caffeine consumption (V1b, V2a, V2b, V3a, and V10a; see Table 1), at least 70% of the learnable subjects in each study condition required only 10 PVTs (taken at around 10:00 and 18:30 during 5 days of CSR) to generate models with similar performance to those obtained with the best-fit models. This represents nearly an 80% reduction in the average number of PVTs (48) available and used by the individualized models to learn the subjects' sleep-loss traits. This analysis suggests that taking two PVTs per day, one in the morning and one in the evening, for five consecutive days of CSR conditions is sufficient to learn the trait-like response to sleep loss of most individuals.

While both caffeine intake and sleep opportunities were controlled in the three field studies in Table 1 (V6, V8, and V9), we investigated whether other factors not controlled in the studies could have influenced the PVT measurements and affected the models' performance. We did not observe significant differences in the performance metrics between models based on laboratory and field studies, for either the group-average or the individualized predictions (for the RMSEs, *p*-values from two-sample *t*-tests were 0.34 and 0.56, and for fractions they were 0.57 and 0.88, respectively). We did observe a large variability in the data of study condition V9a (a field study), which precluded us from learning 8 out of 15 subjects (Table 4). However, on the other arm of the study in condition V9b, a large variability was not an issue. We also observed a large variability in study V1, which was a laboratory study. These results support our modeling

assumption that sleep and caffeine intake are the major factors influencing the prediction of alertness impairment.

Our study has limitations. The UMP was developed for healthy young adults without a history of sleep or neurological disorders and, therefore, the conclusions may be different for a heterogeneous and older population. Another possible limitation is that the UMP does not consider chronic caffeine consumption or withdrawal effects. Alertness enhancement for habitually high caffeine users may require larger caffeine doses than for habitually low caffeine users (Einother et al., 2013). In addition, our results are based on PVT statistics, and it is unclear to what extent our findings can be applied to other neurocognitive performance measures. Finally, with the current approach, we were not able to individualize the model for 19% of the subjects. In theory, we could reduce this fraction by extending the model to account for additional individual characteristics or other factors that influence the PVT not included in the current model. However, adding new parameters to the model also increases the challenge of real-time parameter estimation.

Based on our analyses, we believe that real-world deployment of the UMP at the individualized level should follow two sequential phases of prospective, real-time validation: first in laboratory studies then in field studies. We started the first phase by integrating the UMP predictive engine into a smartphone to allow for prospective, real-time assessment. To this end, we created the *2B-Alert* app, which automatically learns the sleep-loss traits of individuals and predicts alertness impairment in real time as a function of sleep history, time of day, and caffeine consumption. To assess these capabilities, we recently performed a *prospective* study where 21 subjects used the *2B-Alert* app during a 62-h TSD laboratory challenge (study V5; Reifman et al., 2019). The results showed that the individualized models could capture the sleep-loss traits of the subjects in

real time by using the first 36 h (12 PVTs) to learn the individuals, and then predicting their alertness for the last 24 h of the study. The average RMSE between the *2B-Alert* app predictions and the data was only 8 ms larger than that obtained with the best-fit model using all the data (54 ms vs. 46 ms) (Reifman et al., 2019). For the same study, here we obtained a comparable average RMSE of 41 ms (Table 4), which is different because each individualized model assessed here was obtained with a different number of PVTs (see Methods). The next logical step is to assess individualized caffeine recommendations in a similar prospective, real-time laboratory study, paving the way for future field testing and the transition of individualized model predictions from the bench to the real world.

In summary, here we validated the group-average and individualized UMP models, demonstrating their ability to adequately predict alertness impairment at the population and individual-specific levels for 22 distinct conditions, spanning the continuum of sleep loss and caffeine consumption. Notably, we showed that the UMP was able to capture the sleep-loss trait of 81% of the subjects, and that the individualized predictions for these subjects and the group-average predictions were indistinguishable from PVT measurements in nearly 80% of the cases, highlighting the benefits of these models as an integral element of fatigue-management tools (Reifman et al., 2019, 2022).

ACKNOWLEDGEMENTS

This work was sponsored by the Military Operational Medicine Research Program of the U.S. Army Medical Research and Development Command (USAMRDC), Fort Detrick, MD, and was supported by USAMRDC Contract No. W81XWH20C0031.

DATA AVAILABILITY STATEMENT

All data will be made available following a written request to the corresponding author, along with a summary of the planned research.

REFERENCES

Belenky, G., et al. (2003). Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. *Journal of Sleep Research*, **12**, 1-12. doi:10.1046/j.1365-2869.2003.00337.x.

Borbély, A. A., and Achermann P. (1999). Sleep homeostasis and models of sleep regulation. *Journal of Biological Rhythms*, **14**, 557-568. doi:10.1177/074873099129000894.

Einother, S. J. L., et al. (2013). Caffeine as an attention enhancer: reviewing existing assumptions. *Psychopharmacology*, **225**, 251-274. doi:10.1007/s00213-012-2917-4.

Flynn-Evans, E. E., et al. (2020). Changes in performance and bio-mathematical model performance predictions during 45 days of sleep restriction in a simulated space mission. *Scientific Reports*, **10**, 15594. doi:10.1038/s41598-020-71929-4.

Hastie, T. J., Tibshirani, R. J., Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Hilaire, M. A. S., Ruger, M., Fratelli, F., Hull, J. T., Phillips, A. J. K., Lockley, S. W. (2017). Modeling neurocognitive decline and recovery during repeated cycles of extended sleep and chronic sleep deficiency. *Sleep*, **40**, 1. doi:10.1093/sleep/zsw009.

Integrated Safety Support. Smartphone Apps. https://integratedsafety.com.au/eclipse/smartphone-apps/#. Accessed January 3, 2022.

Kamimori, G. H., Johnson, D., Thorne, D., and Belenky, G. (2005). Multiple caffeine doses maintain vigilance during early morning operations. *Aviation, Space and Environmental Medicine*, **76**, 1046-1050.

Kamimori, G. H., McLellan, T. M., Tate, C. M., Voss, D. M., Niro, P., and Lieberman, H. R. (2015). Caffeine improves reaction time, vigilance and logical reasoning during extended periods with restricted opportunities for sleep. *Psychopharmacology (Berl)*, **232**, 2031-2042. doi:10.1007/s00213-014-3834-5.

Khitrov, M. Y., Laxminarayan, S., Thorsley, D., et al. (2014). PC-PVT: a platform for psychomotor vigilance task testing, analysis, and prediction. *Behavior Research Methods*, **46**, 140-147. doi:10.3758/s13428-013-0339-9.

Killgore, W. D., Rupp, T. L., Grugle, N. L., Reichardt, R. M., Lipizzi, E. L., and Balkin, T. J. (2008). Effects of dextroamphetamine, caffeine and modafinil on psychomotor vigilance test performance after 44 h of continuous wakefulness. *Journal of Sleep Research*, **17**, 309-321. doi:10.1111/j.1365-2869.2008.00654.x.

Landolt, H. P., Retey, J. V., and Adam, M. (2012). Reduced neurobehavioral impairment from sleep deprivation in older adults: contribution of adenosinergic mechanisms. *Frontiers in Neurology*, **3**, 62. doi:10.3389/fneur.2012.00062.

Liu, J., et al. (2017). Real-time individualization of the unified model of performance. *Journal of Sleep Research*, **26**, 820-831. doi:10.1111/jsr.12535.

Lo, J. C., Groeger, J. A., Santhi, N., et al. (2012). Effects of partial and acute total sleep deprivation on performance across cognitive domains, individuals and circadian phase. *PLoS ONE*, **7**, e45987. doi:10.1371/journal.pone.0045987.

McLellan, T. M., Bell, D. G., and Kamimori, G. H. (2004). Caffeine improves physical performance during 24 h of active wakefulness. *Aviation, Space and Environmental Medicine*, **75**, 666-672.

McLellan, T. M., Kamimori, G. H., Bell, D. G., Smith, I. F., Johnson, D., and Belenky, G. (2005). Caffeine maintains vigilance and marksmanship in simulated urban operations with sleep deprivation. *Aviation, Space and Environmental Medicine*, **76**, 39-45.

Mitchell, D. C., Knight, C. A., Hockenberry, J., Teplansky, R., and Hartman, T. J. (2014). Beverage caffeine intakes in the U.S. *Food and Chemical Toxicology,* **63**, 136-142. doi:10.1016/j.fct.2013.10.042.

Powell, D., Spencer, M. B., and Petrie, K. (2014). Comparison of in-flight measures with predictions of a bio-mathematical fatigue model. *Aviation, Space and Environmental Medicine*, **85**, 1177-1184. doi:10.3357/ASEM.3806.2014.

Rajdev, P., et al. (2013). A unified mathematical model to quantify performance impairment for both chronic sleep restriction and total sleep deprivation. *Journal of Theoretical Biology*, **331**, 66-77. doi:10.1016/j.jtbi.2013.04.013.

Ramakrishnan, S., et al. (2013). A biomathematical model of the restoring effects of caffeine on cognitive performance during sleep deprivation. *Journal of Theoretical Biology*, **319**, 23-33. doi: 10.1016/j.jtbi.2012.11.015.

Ramakrishnan, S., et al. (2015). Can a mathematical model predict an individual's trait-like response to both total and partial sleep loss? *Journal of Sleep Research*, **24**, 262-269. doi:10.1111/jsr.12272.

Ramakrishnan, S., et al. (2016a). A unified model of performance: validation of its predictions across different sleep/wake schedules. *Sleep*, **39**, 249-262. doi:10.5665/sleep.5358.

Ramakrishnan, S., et al. (2016b). A unified model of performance for predicting the effects of sleep and caffeine. *Sleep*, **39**, 1827-1841. doi:10.5665/sleep.6164.

Reifman, J., et al. (2019). *2B-Alert* App: a mobile application for realtime individualized prediction of alertness. *Journal of Sleep Research*, **28**, e12725. doi:10.1111/jsr.12725.

Reifman, J., et al. (2022). *2B-Alert* Web 2.0, an open-access tool for predicting alertness and optimizing the benefits of caffeine: utility study. *Journal of Medical Internet Research*, **24**, e29595. doi:10.2196/29595.

Rétey, J. V., Adam, M., Gottselig, J. M., et al. (2006). Adenosinergic mechanisms contribute to individual differences in sleep deprivation-induced changes in neurobehavioral function and brain rhythmic activity. *Journal of Neurosci*ence, **26**, 10472-10479. doi:10.1523/JNEUROSCI.1538-06.2006.

Rupp, T. L., Wesensten, N. J., and Balkin, T. J. (2012). Trait-like vulnerability to total and partial sleep loss. *Sleep*, **35**, 1163-1172. doi:10.5665/sleep.2010.

Rupp, T. L., Wesensten, N. J., Bliese, P. D., and Balkin, T. J. (2009). Banking sleep: realization of benefits during subsequent sleep restriction and recovery. *Sleep*, **32**, 311-321. doi:10.1093/sleep/32.3.311.

So, C. J., Quartana, P. J., and Ratcliffe, R. H. (2016). Caffeine efficacy across a simulated 5-day work week with sleep restriction. *Sleep*, **39**, A92.

Van Dongen, H. P., Baynard, M. D., Maislin, G., and Dinges, D. F. (2004). Systematic interindividual differences in neurobehavioral impairment from sleep loss: evidence of trait-like differential vulnerability. *Sleep*, **27**, 423-433. doi:10.1093/SLEEP/27.3.423.

Vital-Lopez, F. G., Doty, T. J., and Reifman, J. (2021). Optimal sleep and work schedules to maximize alertness. *Sleep*, **44**, 144. doi:10.1093/sleep/zsab144.

Wesensten, N. J., Belenky, G., Thorne, D. R., Kautz, M. A., and Balkin, T. J. (2004). Modafinil vs. caffeine: effects on fatigue during sleep deprivation. *Aviation, Space and Environmental Medicine*, **75**, 520-525.

Wesensten, N. J., Killgore, W. D., and Balkin, T. J. (2005). Performance and alertness effects of caffeine, dextroamphetamine, and modafinil during sleep deprivation. *Journal of Sleep Research*, **14**, 255-266. doi:10.1111/j.1365-2869.2005.00468.x.

Wesensten, N. J., Reichardt, R. M., and Balkin, T. J. (2007). Ampakine (CX717) effects on performance and alertness during simulated night shift work. *Aviation, Space and Environmental Medicine*, **78**, 937-943.

FIGURE CAPTIONS AND TABLES

Figure 1. Psychomotor vigilance test (PVT) mean response time (RT) data along with the group-average and individualized model predictions for study condition V1b. The study consisted of two baseline nights of 8 h of time in bed (TIB; for B1 and B2), followed by 7 nights of chronic sleep restriction (CSR; 6 h of TIB for C1–C7) and 41 h of total sleep deprivation (TSD; T1). (a) Group-average mean RT and group-average model predictions, representative of above-average model performance. The error bars denote two standard errors of the mean. (b-d) Individualized predictions for three subjects: subject #20 [strong individualized model predictions, (b)], subject #14 [average individualized model predictions, (c)], and subject #13 [weak individualized model predictions, (d)]. Open and filled circles correspond to data used by the individualized model to "learn" the subject and assess model predictions, respectively. The shaded regions in all panels represent the width of the prediction intervals. The gray vertical stripes represent sleep episodes. RMSE, root mean square error.

Figure 2. Psychomotor vigilance test (PVT) mean response time (RT) data along with the group-average and individualized model predictions for study condition V11b. The study consisted of one baseline night (B1) of 7 h of time in bed (TIB), followed by 54.5 h of total sleep deprivation (TSD; T1 and T2), where subjects consumed 600 mg of caffeine (vertical dashed green line) at 23:50 (i.e., after 41.5 h of wakefulness). (a) Group-average mean RT and group-average model predictions, representative of below-average model performance. The error bars denote two standard errors of the mean. (b-d) Individualized predictions for three subjects: subject #10 [average individualized model predictions, (b)], subject #3 [weak individualized model predictions, (c)], and subject #9 [weak individualized model predictions, (d)]. Open and filled circles correspond to data used by the individualized model to "learn" the subject and assess

model predictions, respectively. The shaded regions in all panels represent the width of the prediction intervals. The gray vertical stripes represent sleep episodes. RMSE, root mean square error.

Figure 3. Psychomotor vigilance test (PVT) mean response time (RT) data along with the group-average and individualized model predictions for study condition V2b. The study consisted of eight baseline nights of 8 h of time in bed (TIB; for B1–B8), followed by seven nights of chronic sleep restriction (CSR; 3 h of TIB for C1–C7) and five recovery nights of 8 h of TIB (R1–R5). (a) Group-average mean RT and group-average model predictions, representative of above-average model performance. The error bars denote two standard errors of the mean. (b-d) Individualized predictions for three subjects: subject #11 [strong individualized model predictions, (b)], subject #3 [strong individualized model predictions, (c)], and subject #12 [weak individualized model predictions, (d)]. Open and filled circles correspond to data used by the individualized model to "learn" the subject and assess model predictions, respectively. The shaded regions in all panels represent the width of the prediction intervals. The gray vertical stripes represent sleep episodes. RMSE, root mean square error.

Figure 4. Psychomotor vigilance test (PVT) mean response time (RT) data along with the group-average and individualized model predictions for study condition V10b. The study consisted of five baseline nights (B1–B5) of 10 h of time in bed (TIB), followed by five nights of chronic sleep restriction (CSR; 5 h of TIB for C1–C5) and three recovery nights of 8 h of TIB (R1–R3). During the CSR phase, subjects consumed 200 mg of caffeine at 08:00 and 12:00 daily (vertical dashed green lines). (a) Group-average mean RT and group-average model predictions, representative of above-average model performance. The error bars denote two standard errors of the mean. (b-d) Individualized predictions for three subjects: subject #6 [strong individualized

model predictions, (b)], subject #21 [strong individualized model predictions, (c)], and subject #8 [weak individualized model predictions, (d)]. Open and filled circles correspond to data used by the individualized model to "learn" the subject and assess model predictions, respectively. The shaded regions in all panels represent the width of the prediction intervals. The gray vertical stripes represent sleep episodes. RMSE, root mean square error.

Figure 5. Cumulative distribution of the percentage of subjects in a given study condition learned by the model as a function of the number of psychomotor vigilance test (PVT) measurements needed to individualize the model parameters. Nine panels represent studies with two conditions that differed by only one factor: (b) study V2 (baseline sleep of 8 h vs. 10 h of time in bed), (c) study V3 (chronic sleep restriction vs. total sleep deprivation), and (f–l) studies V6–V12 (caffeine vs. placebo). The dashed green curves and solid blue curves denote the results for study conditions "a" and "b," respectively (see Table 1).

Table 1. Summary of studies used to assess the Unified Model of Performance

Study condition	# Subjects (men)	# PVTs	Sleep schedule			Caffeine-consumption schedule	
			Baseline (TIB, h)	TSD (wakefulness, h) or CSR (TIB, h)	Recovery (TIB, h)	Dose, mg	Time of day
Studies with no caffeine							
V1a	36 (18)	61	2 nights (8)+ 7 nights (10)	TSD (39)	-		
V1b	36 (18)	62	2 nights (8)	7 CSR nights (6) + TSD (41)	-		
V2a*	12 (7)	143	8 nights (10)	7 CSR nights (3)	5 nights (8)		
V2b*	12 (4)	143	8 nights (8)	7 CSR nights (3)	5 nights (8)		
V3a	19 (11)	109	7 nights (10)	7 CSR nights (3)	3 nights (8)		
V3b	19 (11)	64	7 nights (10)	TSD (63)	3 nights (8)		
V4	12 (12)	45	-	TSD (25) + 4 CSR day (4)	-		
V5*	21 (14)	23	1 night (8)	TSD (62)	1 night (12)		
Studies with caffeine and placebo							
V6a*	11 (11)	35	1 night (8)	TSD (31) + 2 CSR days (4)	-		
V6b*	10 (10)	35	1 night (8)	TSD (31) + 2 CSR days (4)	-	4×200	21:45, 01:00, 03:45, 07:00 (daily)
V7a*	14**	34	1 night (8)	TSD (61)	1 night (12)		
V7b*	11**	34	1 night (8)	TSD (61)	1 night (12)	1×600	After 44 h of wakefulness at 03:00
V8a*	21**	11	1 night (8)	TSD (28)	-		
V8b*	21**	11	1 night (8)	TSD (28)	-	1×400, 2×200	21:30, 03:00, 05:00 during TSD
V9a*	15 (15)	31	-	1 CSR night (3) + TSD (33)	-		
V9b*	15 (15)	31	-	1 CSR night (3) + TSD (33)	-	2×(100, 200)	21:45, 23:45, 01:45, 03:45 during TSD
V10a	24 (10)	145	5 nights (10)	5 CSR nights (5)	3 nights (8)		
V10b	24 (9)	145	5 nights (10)	5 CSR nights (5)	3 nights (8)	2×200	08:00, 12:00 (daily)
V11a	10 (6)	37	1 night (7)	TSD (54.5)	1 night (24)		
V11b	10 (6)	37	1 night (7)	TSD (54.5)	1 night (24)	1×600	After 41.5 h of wakefulness at 23:50
V12a	12 (11)	48	1 night (8)	TSD (85)	1 night (12)		
V12b	12 (11)	48	1 night (8)	TSD (85)	1 night (12)	1×600	After 65 h of wakefulness at 00:00

*5-min psychomotor vigilance test (PVT); otherwise, 10-min PVT. **Sex information was not available. CSR: chronic sleep restriction; TIB: time in bed; TSD: total sleep deprivation. References: V1 (Lo et al., 2012), V2 (Rupp et al., 2009), V3 (Rupp et al., 2012), V4 (Wesensten et al., 2007), V5 (Reifman et al., 2019), V6 (Kamimori et al., 2015), V7 (Killgore et al., 2008), V8 (McLellan et al., 2004), V9 (McLellan et al., 2005), V10 (So et al., 2016), V11 (Wesensten et al., 2004), and V12 (Wesensten et al., 2005).

Table 2. Equations of the Unified Model of Performance

UMP governing equations
Performance impairment without caffeine (P_o): $$P_o(t,\theta) = S(t) + \kappa C(t), \quad (2)$$ where θ represents the eight model parameters of the UMP, namely, $\theta = [U, \tau_w, \tau_s, \tau_{LA}, \kappa, \phi, S_0, L_0]^T$ as defined below. The time-dependent functions $S(t)$ and $C(t)$ denote the homeostatic and circadian processes, respectively, and κ denotes the circadian amplitude. Because the UMP predictions are not particularly sensitive to time constants τ_w, τ_s, and τ_{LA} (Ramakrishnan et al., 2015), we fixed them to 18.2 h, 4.2 h, and 7.0 days, respectively.
Circadian process (C): $$C(t) = \sum_{j=1}^{5} a_j \, sin\left[j\frac{2\pi}{\tau}(t+\phi)\right], \quad (3)$$ where a_j, j = 1, …, 5, denotes the amplitude of the five harmonics (a_1 = 0.97, a_2 = 0.22, a_3 = 0.07, a_4 = 0.03, and a_5 = 0.001), τ indicates the fundamental period of the circadian clock (~24 h), and ϕ represents the circadian phase.
Homeostatic process (S): $$\dot{S}(t) = \begin{cases} 1/\tau_w\,[U - S(t)] & \text{during wakefulness} \\ -1/\tau_s\,[S(t) - L(t)] & \text{during sleep,} \end{cases} \quad (4)$$ where U and L denote the upper and lower asymptotes, respectively, and τ_w and τ_s indicate the wake and sleep time constants of the increasing and decreasing sleep pressure, respectively. $S(0) = S_0$ and $L(0) = L_0$ correspond to the initial-state values for S and L.
Lower asymptote (L) of process S is defined as follows: $$L(t) = \begin{cases} max\{U - (U - L_0)\,exp(-t/\tau_{LA}),\ -0.11U\} & \text{during wakefulness} \\ max\{-2U + (2U + L_0)\,exp(-t/\tau_{LA}),\ -0.11U\} & \text{during sleep,} \end{cases} \quad (5)$$ where τ_{LA} denotes the time constant of the exponential decay of the effect of sleep history on performance.
The effect of caffeine (g_{PD}): $$g_{PD}(t,c) = \left[1 + M_c \frac{k_a}{k_a - k_c}\{\exp[-k_c(t-t_o)] - \exp[-k_a(t-t_o)]\}\right]^{-1} \text{ for } t \geq t_o$$ $$M_c = M_o \cdot c \text{ and } k_c = k_0 \exp(-z \cdot c), \quad (6)$$ where M_c and k_c indicate the amplitude factor and the elimination rate for a caffeine dose c scheduled at time t_o, respectively. Here, M_o, k_0, z, and k_a denote the amplitude slope, basal elimination rate, decay constant, and absorption rate, respectively. We fixed the caffeine parameters as described in Table S1.

Table 3. Individualization of the Unified Model of Performance

Bayesian learning:

$$\arg\min_{\theta}\left\{(\theta-\theta_0)^T\Sigma_0^{-1}(\theta-\theta_0)+\frac{1}{\sigma^2}\sum_{i=1}^{n}[y_i-P_o(t_i,\theta)]^2\right\}, \quad (7)$$

where θ_0 represents the parameters of an "average" individual, Σ_0 denotes the prior variance-covariance matrix of the model parameters θ_0, and σ^2 indicates the noise variance in PVT measurements y_i. The solution of Eq. (7) leads to the individualized model based on a set of n PVT measurements y_i, with $i = 1, 2,\ldots, n$, up to the current time t_n (where $n \leq N$, the total number of measurements).

Recursive learning based on the extended Kalman filter:

We recursively estimated the model parameter $\hat{\theta}_n$, at the current time t_n, with $n = 1, 2,\ldots, N$, as a function of the previous estimate $\hat{\theta}_{n-1}$ at time t_{n-1} and the current PVT measurement y_n, by solving the following algebraic equations:

$$\hat{\theta}_n=\hat{\theta}_{n-1}+\frac{\hat{\Sigma}_{n-1}J_n}{\sigma^2+J_n^T\hat{\Sigma}_{n-1}J_n}[y_n-P_o(t_n,\hat{\theta}_{n-1})] \quad (8)$$

$$\hat{\Sigma}_n=\left(I-\frac{\hat{\Sigma}_{n-1}J_nJ_n^T}{\sigma^2+J_n^T\hat{\Sigma}_{n-1}J_n}\right)\hat{\Sigma}_{n-1} \quad (9)$$

where $\hat{\Sigma}_n$ and $\hat{\Sigma}_{n-1}$ represent the estimated variance-covariance matrices of the model parameters at times t_n and t_{n-1}, respectively, $J_n=\partial P_o(t_n,\theta)/\partial\theta|_{\theta=\hat{\theta}_{n-1}}$ denotes the Jacobian of the model output with respect to the model parameters at time t_n, and I represents the identity matrix. To start the recursion, we assume that $\hat{\theta}_0=\theta_0$ and $\hat{\Sigma}_0=\Sigma_0$, where θ_0 and Σ_0 denote priors as in Eq. (7). To customize the model to an individual, we only needed to estimate five UMP parameters, i.e., the upper asymptote U, the circadian amplitude and phase, κ and ϕ, respectively, and the initial state values for process S and for the lower asymptote L_0 (Table S1).

Table 4. Performance of the group-average model and the individualized model. The values for the individualized models denote their mean [range] performance in each study condition for the PVTs after the model had "learned" the subject's trait-like response to sleep deprivation

Study condition	# Subjects (learnable subjects)	Total # PVTs[1]	Group-average model[3]		Individualized model[4]			
			RMSE, ms	Fraction,[2] %	# PVTs to learn a subject	Time to learn a subject, h[5]	RMSE, ms	Fraction,[2] %
Caffeine free								
V1a TSD	36 (12)	61 (43,18,0)	46	79	45 [1–59]	196 [0–245]	83 [41–229]	47 [0–88]
V1b both[6]	36 (12)	62 (43,19,0)	28	100	41 [4–60]	183 [18–245]	101 [25–231]	56 [0–93]
V2a CSR	12 (12)	143 (0,88,55)	38	69	48 [8–86]	99 [7–176]	56 [26–109]	86 [57–100]
V2b CSR	12 (10)	143 (0,88,55)	26	100	41 [15–96]	85 [27–199]	54 [20–118]	87 [65–100]
V3a CSR	19 (17)	109 (13,76,20)	70	41	45 [23–76]	221 [52–305]	57 [21–90]	72 [45–100]
V3b TSD	19 (17)	64 (13,31,20)	74	70	33 [25–47]	195 [177–227]	58 [26–79]	72 [47–100]
V4 both	12 (11)	45 (0,45,0)	60	80	24 [12–36]	48 [22–75]	39 [18–61]	87 [65–100]
V5 TSD	21 (20)	23 (0,20,3)	25	83	12 [7–19]	33 [18–54]	41 [17–95]	87 [50–100]
Placebo								
V6a both	11 (11)	35 (4,31,0)	55	63	17 [11–32]	56 [41–89]	46 [21–103]	81 [17–100]
V7a TSD	14 (13)	34 (0,30,4)	30	82	14 [1–26]	26 [0–54]	60 [26–179]	82 [42–100]
V8a TSD	21 (17)	11 (0,11,0)	15	82	9 [3–11]	23 [5–28]	66 [15–179]	57 [0–100]
V9a both	15 (6)	31 (0,31,0)	54	87	13 [8–17]	38 [34–42]	85 [22–309]	64 [0–100]
V10a CSR	24 (24)	145 (0,105,40)	29	97	53 [12–121]	68 [12–168]	48 [23–109]	83 [27–100]
V11a TSD	10 (9)	37 (3,31,1)	54	89	26 [18–32]	53 [40–60]	147 [53–267]	32 [17–68]
V12a TSD	12 (10)	48 (2,42,4)	73	63	29 [14–39]	63 [34–84]	78 [37–139]	60 [38–85]
Caffeine								
V6b both	10 (10)	35 (4,31,0)	47	63	13 [8–16]	45 [38–53]	32 [18–62]	91 [64–100]
V7b TSD	11 (11)	34 (0,30,4)	31	79	19 [11–28]	37 [20–54]	47 [27–73]	80 [50–100]
V8b TSD	21 (17)	11 (0,11,0)	22	100	9 [3–11]	23 [5–28]	41 [16–83]	90 [0–100]

V9b both	15 (13)	31 (0,31,0)	44	74	9 [1–16]	27 [0–41]	53 [17–115]	75 [0–100]
V10b CSR	24 (23)	145 (0,105,40)	19	100	36 [1–105]	44 [0–135]	50 [30–98]	80 [43–98]
V11b TSD	10 (6)	37 (3,31,1)	51	76	33 [29–35]	61 [57–63]	105 [68–147]	33 [0–67]
V12b TSD	12 (9)	48 (2,42,4)	64	75	21 [13–30]	48 [32–66]	69 [49–87]	61 [50–82]
Average			**43**	**80**	**27**	**76**	**64**	**71**

[1]The total number of PVTs and the number of PVTs during baseline, sleep challenge, and recovery phases, respectively (see Table 1). [2]Fraction is defined as the number of PVTs that fall within the prediction intervals of the model, divided by the total number of PVTs in the study condition (see Methods). [3]Based on 301 unique subjects. [4]Based on 244 unique subjects. [5]Time between the first and last PVT sessions required to learn the sleep-loss trait of a subject. [6]Both represents study conditions that included chronic sleep restriction (CSR) and total sleep deprivation (TSD) challenges. PVT: psychomotor vigilance test; RMSE: root mean square error between the model prediction and the measured PVT data. References: V1 (Lo et al., 2012), V2 (Rupp et al., 2009), V3 (Rupp et al., 2012), V4 (Wesensten et al., 2007), V5 (Reifman et al., 2019), V6 (Kamimori et al., 2015), V7 (Killgore et al., 2008), V8 (McLellan et al., 2004), V9 (McLellan et al., 2005), V10 (So et al., 2016), V11 (Wesensten et al., 2004), and V12 (Wesensten et al., 2005).

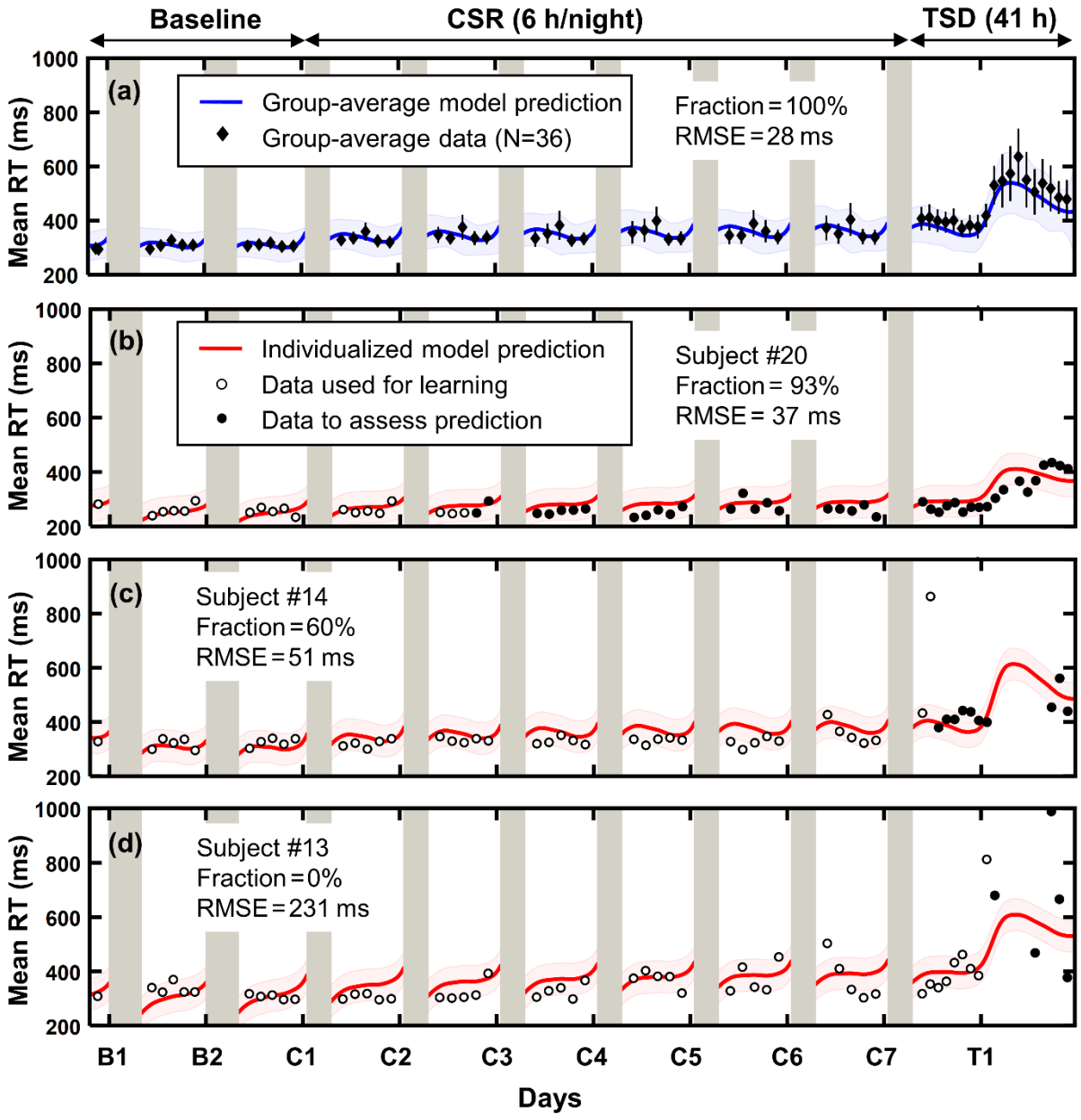
Baseline
CSR (6 h/night)
TSD (41 h)
(a)
Group-average model prediction
Group-average data (N=36)
Fraction = 100%
RMSE = 28 ms
(b)
Individualized model prediction
Data used for learning
Data to assess prediction
Subject #20
Fraction = 93%
RMSE = 37 ms
(c)
Subject #14
Fraction = 60%
RMSE = 51 ms
(d)
Subject #13
Fraction = 0%
RMSE = 231 ms
Mean RT (ms)
1000
800
600
400
200
B1
B2
C1
C2
C3
C4
C5
C6
C7
T1
Days

Figure 1

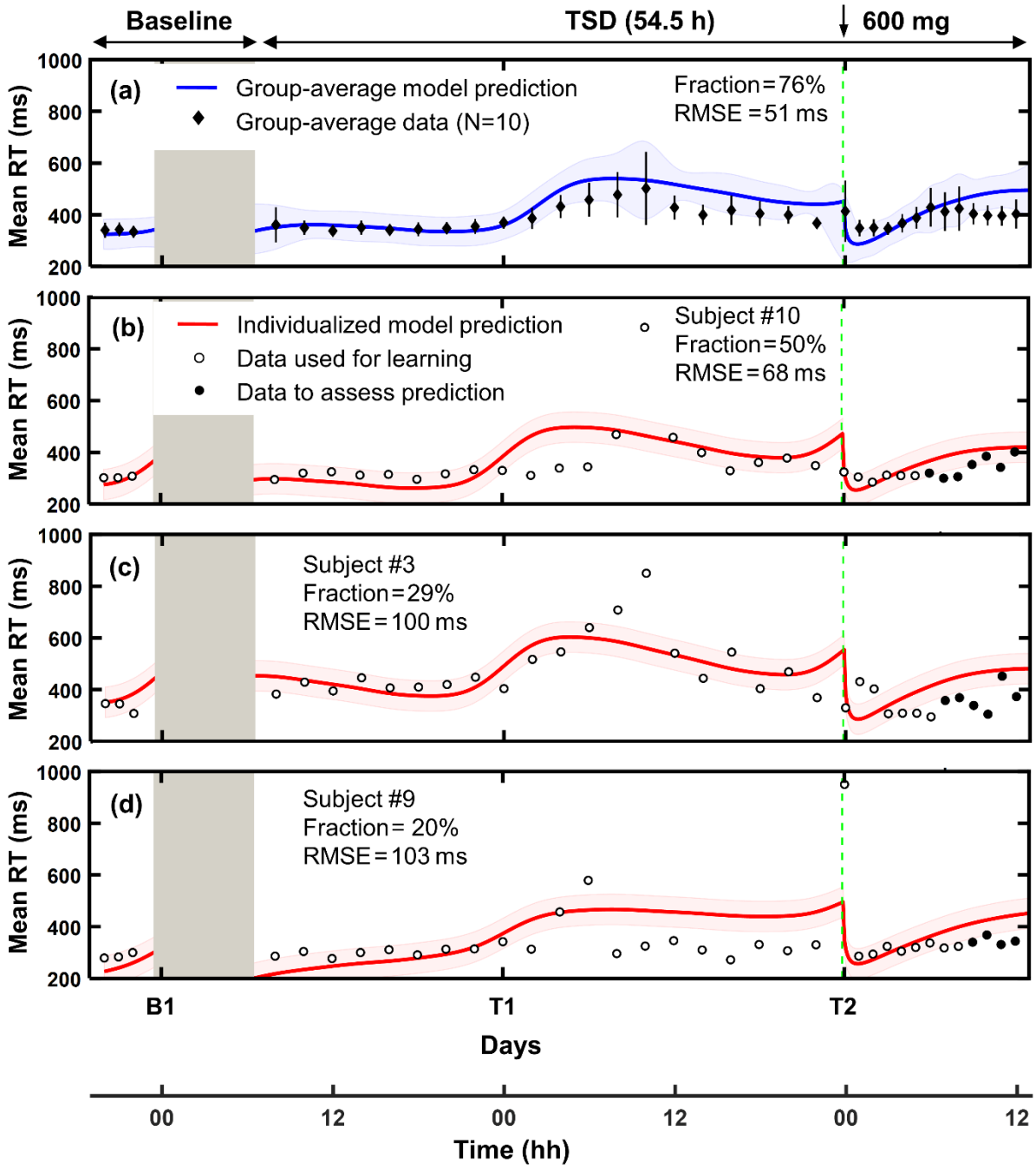
Baseline
TSD (54.5 h)
600 mg
(a)
Group-average model prediction
Group-average data (N=10)
Fraction=76%
RMSE =51 ms
(b)
Individualized model prediction
Data used for learning
Data to assess prediction
Subject #10
Fraction=50%
RMSE=68 ms
(c)
Subject #3
Fraction=29%
RMSE=100 ms
(d)
Subject #9
Fraction= 20%
RMSE=103 ms
Mean RT (ms)
1000
800
600
400
200
B1
T1
T2
Days
00
12
00
12
00
12
Time (hh)

Figure 2

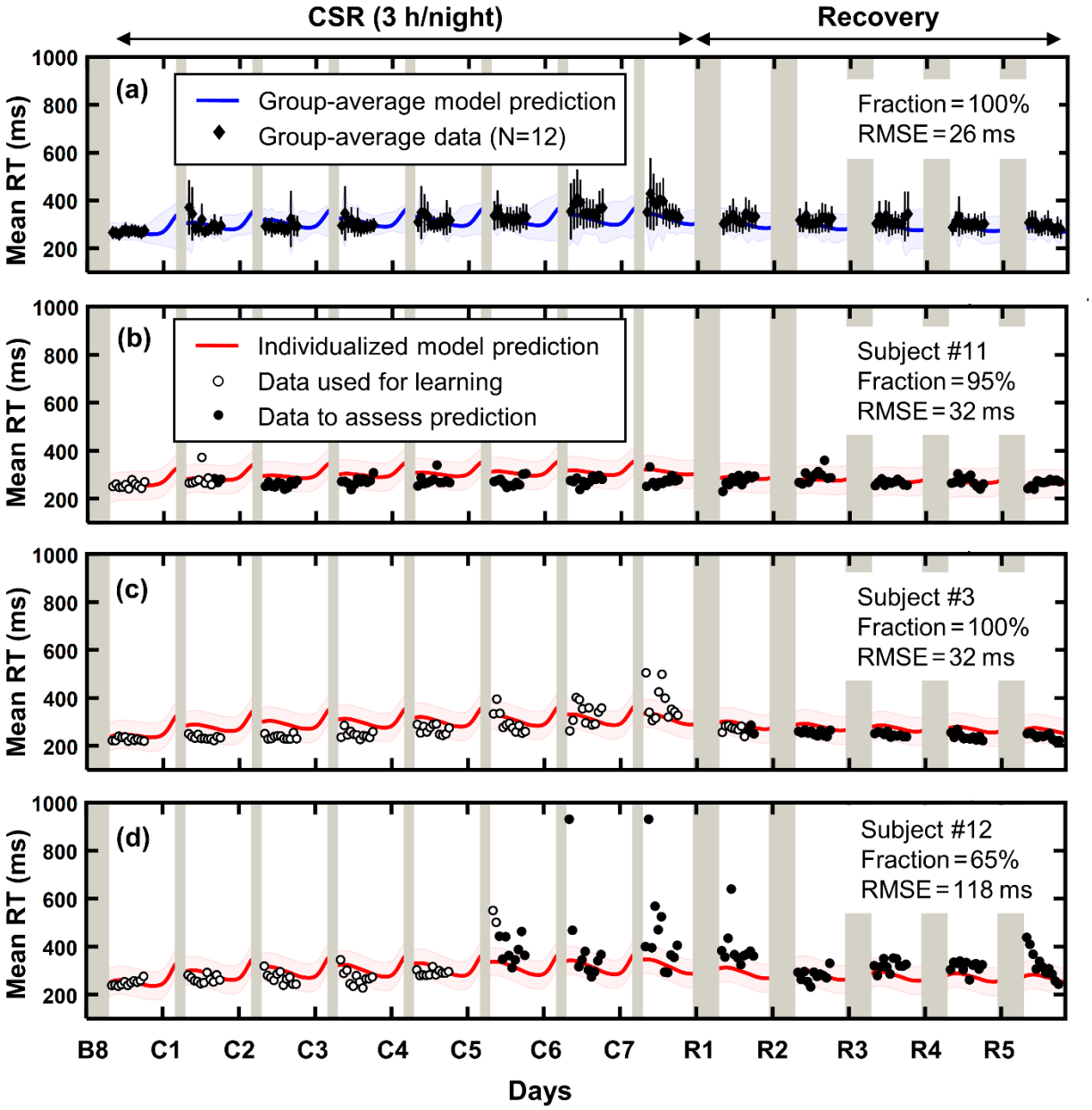
CSR (3 h/night)
Recovery
(a)
Group-average model prediction
Group-average data (N=12)
Fraction = 100%
RMSE = 26 ms
(b)
Individualized model prediction
Data used for learning
Data to assess prediction
Subject #11
Fraction = 95%
RMSE = 32 ms
(c)
Subject #3
Fraction = 100%
RMSE = 32 ms
(d)
Subject #12
Fraction = 65%
RMSE = 118 ms
Mean RT (ms)
1000
800
600
400
200
B8
C1
C2
C3
C4
C5
C6
C7
R1
R2
R3
R4
R5
Days

Figure 3

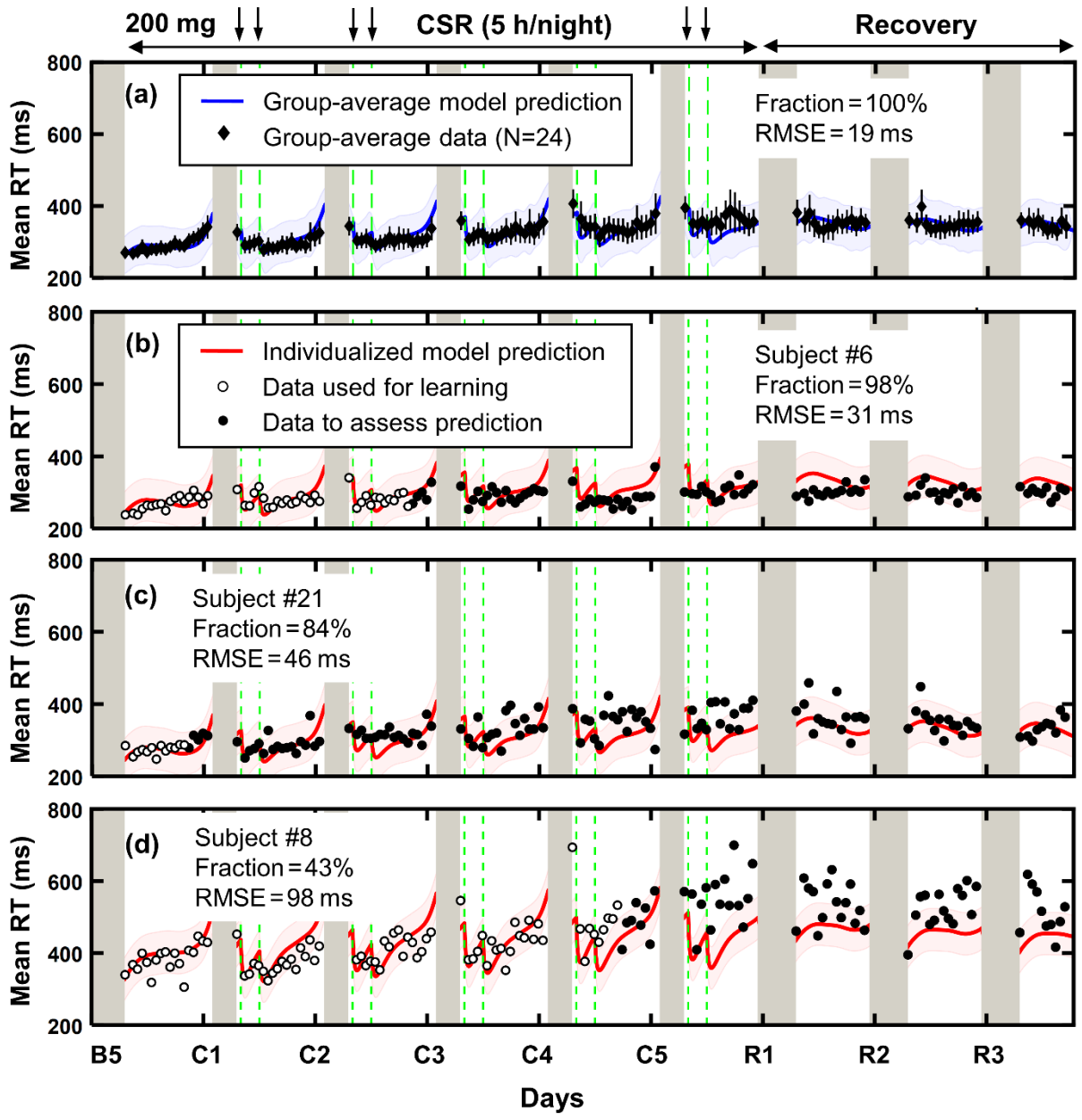
200 mg
CSR (5 h/night)
Recovery
(a)
Group-average model prediction
Group-average data (N=24)
Fraction = 100%
RMSE = 19 ms
(b)
Individualized model prediction
Data used for learning
Data to assess prediction
Subject #6
Fraction = 98%
RMSE = 31 ms
(c)
Subject #21
Fraction = 84%
RMSE = 46 ms
(d)
Subject #8
Fraction = 43%
RMSE = 98 ms
Mean RT (ms)
800
600
400
200
B5
C1
C2
C3
C4
C5
R1
R2
R3
Days

Figure 4

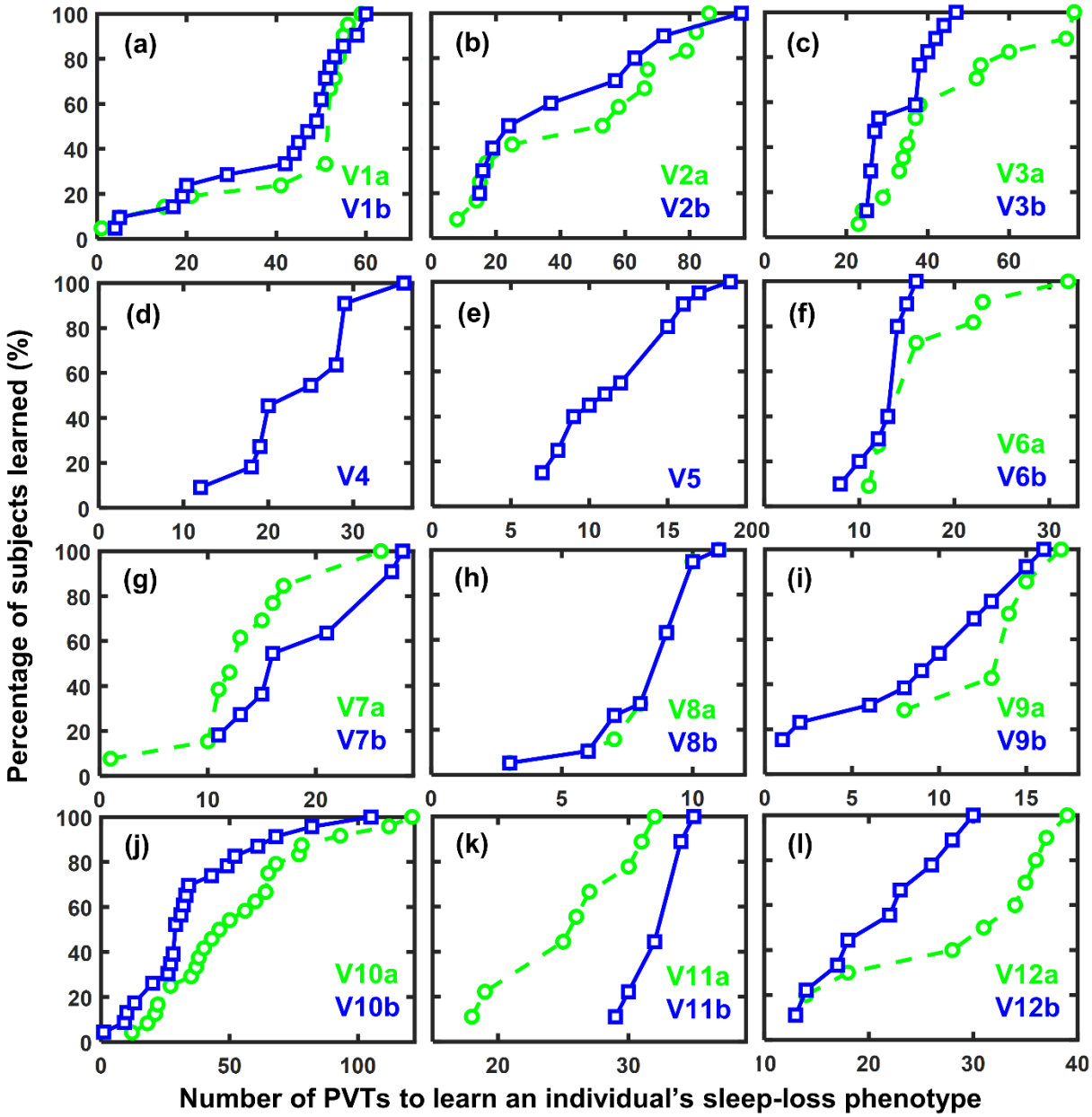
(a)
V1a
V1b
(b)
V2a
V2b
(c)
V3a
V3b
(d)
V4
(e)
V5
(f)
V6a
V6b
(g)
V7a
V7b
(h)
V8a
V8b
(i)
V9a
V9b
(j)
V10a
V10b
(k)
V11a
V11b
(l)
V12a
V12b
Percentage of subjects learned (%)
Number of PVTs to learn an individual's sleep-loss phenotype

Figure 5