



Published in final edited form as:

Accid Anal Prev. 2013 January ; 50: 992–1002. doi:10.1016/j.aap.2012.08.003.

Classifying performance impairment in response to sleep loss using pattern recognition algorithms on single session testing

Melissa A. St. Hilaire, Ph.D.^{a,b}, Jason P. Sullivan^b, Clare Anderson, Ph.D.^{b,c}, Daniel A. Cohen, M.D., M.Sc.^{b,c,d}, Laura K. Barger, Ph.D.^{b,c}, Steven W. Lockley, Ph.D.^{b,c}, and Elizabeth B. Klerman, M.D., Ph.D.^{a,b,c}

^aAnalytic and Modeling Unit, Division of Sleep Medicine, Brigham and Women's Hospital, 221 Longwood Avenue, Boston, Massachusetts 02115, USA

^bDivision of Sleep Medicine, Brigham and Women's Hospital, 221 Longwood Avenue, Boston, Massachusetts 02115, USA

^cDivision of Sleep Medicine, Harvard Medical School, 401 Park Drive, 2nd Floor East Boston, Massachusetts 02215, USA

^dDepartment of Neurology, Beth Israel Deaconess Medical Center, Harvard Medical School, 330 Brookline Avenue, Boston, Massachusetts 02215, USA

Abstract

There is currently no “gold standard” marker of cognitive performance impairment resulting from sleep loss. We utilized pattern recognition algorithms to determine which features of data collected under controlled laboratory conditions could most reliably identify cognitive performance impairment in response to sleep loss using data from only one testing session, such as would occur in the “real world” or field conditions. A training set for testing the pattern recognition algorithms was developed using objective Psychomotor Vigilance Task (PVT) and subjective Karolinska Sleepiness Scale (KSS) data collected from laboratory studies during which subjects were sleep deprived for 26 – 52 hours. The algorithm was then tested in data from both laboratory and field experiments. The pattern recognition algorithm was able to identify performance impairment with a single testing session in individuals studied under laboratory conditions using PVT, KSS, length of time awake and time of day information with sensitivity and specificity as high as 82%. When this algorithm was tested on data collected under real-world conditions from individuals whose data were not in the training set, accuracy of predictions for individuals categorized with low performance impairment were as high as 98%. Predictions for medium and severe performance impairment were less accurate. We conclude that pattern recognition algorithms may be a promising method for identifying performance impairment in individuals using only current information about the individual's behavior. Single testing features (e.g., number of PVT lapses) with high correlation with performance impairment in the laboratory setting may not be the best indicators of performance impairment under real-world conditions. Pattern recognition algorithms should be further tested for their ability to be used in conjunction

© 2012 Elsevier Ltd. All rights reserved.

Corresponding Author: Melissa A. St. Hilaire, Analytic and Modeling Unit, Division of Sleep Medicine, Brigham & Women's Hospital, 221 Longwood Avenue, Boston, Massachusetts 02215, USA, msthilaire@rics.bwh.harvard.edu, Telephone: 617-278-0620, Fax: 617-732-4015.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

with other assessments of sleepiness in real-world conditions to quantify performance impairment in response to sleep loss.

Keywords

performance impairment; sleep deprivation; pattern recognition

1.0 Introduction

Laboratory studies have quantified the effects of insufficient sleep from either acute sleep deprivation or chronic sleep restriction on cognitive performance, including impaired reaction time, accuracy, visual attention, working memory and decision making, and subjective alertness (Belenky *et al.* 2003, Van Dongen *et al.* 2003, Santhi *et al.* 2007). It is now well accepted that multiple aspects of performance and alertness are affected by (i) a circadian process, an ~24-hour rhythm regulated by the suprachiasmatic nucleus of the hypothalamus; (ii) homeostatic processes in which sleep pressure builds during wake and declines during sleep; and (iii) non-linear interaction of these processes (Dijk *et al.* 1992, Dijk and Czeisler 1994, Wyatt *et al.* 2004, Cohen *et al.* 2010). In these controlled laboratory conditions, environmental factors such as schedule, light levels, activity level and meal timing have been carefully controlled and other activities such as caffeine consumption and pharmaceuticals have been eliminated. Therefore, translation and application of these experimental findings into an operational/real-world setting in which prediction of performance in an individual has been a goal, has been limited. Several mathematical models have used the results of these laboratory experiments to attempt to provide a predictive tool of the effect of a given sleep/wake schedule on cognitive performance (for review, see (Van Dongen 2004)); some of these models have also incorporated data collected under operational settings (see (Mallis *et al.* 2004)). These models, however, require inputs of prior sleep/wake history and possibly ambient light levels to appropriately estimate sleep/wake homeostasis (Akerstedt *et al.* 2004, Hursh *et al.* 2004) and circadian phase (timing) (Jewett and Kronauer 1999). In operational/real-world settings, requiring these inputs decreases the practical utility of these models. Furthermore, the output of the models are based on population averages, which also limits utility since there are large inter-individual differences in performance measures, such that some people maintain levels of performance comparable to their well-rested baseline values even after many hours or several days of sleep deprivation whereas others become impaired more quickly (Van Dongen *et al.* 2004).

Recent model developments have enabled predictions based on an individual (Van Dongen *et al.* 2007, Rajaraman *et al.* 2008). These models, however, require that individuals are monitored over several hours or days in order to identify the magnitude of their performance impairment in response to sleep loss relative to their baseline performance levels. In a real-world setting, prolonged multiple data collection sessions may not be feasible. What is needed is a real-time measure that can provide information about an individual's level of performance impairment at that particular moment in time with a single measurement session and without prior knowledge of variables such as past performance on cognitive tests, prior sleep/wake history, circadian phase, ambient light levels or prior light exposure history.

We hypothesized that pattern recognition algorithms could be used to extract important features using already collected data as the basis for categorizing (or classifying) performance impairment in a new individual using data collected from a single testing session by matching the features of the new data (test set) to the existing dataset (training

set). If successful, pattern recognition algorithms could utilize the large volumes of already-collected data from laboratory studies to create training sets against which to classify new data collected in the field. In addition, this classification can be made using only the features of the training set that are deemed necessary for reliably identifying the class (e.g., low or high performance impairment) of a new individual. The pattern recognition algorithm may not require information such as prior sleep/wake history, lighting levels and/or baseline performance information since it is presumed that the effects of these variables are embodied in the behavioral response and do not require explicit inclusion as features in the training set. In this paper, we introduce the use of pattern recognition algorithms to identify level of performance impairment and validate these pattern recognition algorithms on data previously collected in a field study of hospital interns (Lockley *et al.* 2004, Anderson *et al.* in press) and in a field study of ground control crews working on a Mars sol (T=24.65 hr) schedule (Barger *et al.*, in press).

2.0 Methods

2.1 Datasets

All studies were approved by the Partners Healthcare Institutional Review Board. Informed consent was obtained from all subjects prior to study.

(a) Laboratory Data: The laboratory data used to train and validate our pattern recognition algorithms were collected under three separate protocols that included sleep deprivation. All subjects were healthy, not on medications and were not allowed caffeine or other stimulants. Details of subject selection and experimental protocols are included in the published reports (Klerman and Dijk 2008, Cohen *et al.* 2010) of each protocol. In total, the data set used to train and validate our pattern recognition algorithms includes 33 subjects and 506 testing sessions.

In the first study, younger (N = 17, 9 female, mean age 23.1 ± 3.9 years, range 18–32 years) and older subjects (N = 7, 3 female, mean age 65.6 ± 4.2 years, range 60–71 years) were scheduled to 28 (N=14, 5 older) or 52 (N=10, 2 older) hours of sleep deprivation. During the sleep deprivation component of the protocol, the 10-minute version of the Psychomotor Vigilance Task (PVT) was administered every 2 hours and the Karolinska Sleepiness Scale (KSS) was administered every 30 minutes. Only KSS scores administered immediately prior to a PVT were used for this analysis.

In the second study, 9 subjects (4 female, mean age 27.1 ± 4.5 years, range 21–34 years) were scheduled to a T=42.85-hr forced desynchrony protocol with 10 hours of bedrest opportunity per 42.85-hr “day” (Cohen *et al.* 2010). Subjects were awake 32.85 hours per “day”, thereby having an extended wake duration similar to the acute sleep deprivation condition in the first laboratory study. The 10-minute version of the PVT was administered every 4 hours during this protocol and the KSS was administered every 30 minutes; only data from the first day of the forced desynchrony were included in our data set (i.e., a single 32.85-hr wake episode) and only KSS data associated with PVT data were used.

In the third study (St. Hilaire *et al.*, unpublished data), 12 subjects (6 female, mean age 23.3 ± 3.0 years, range 18–30 years) were scheduled to a 50-hr sleep deprivation preceded by 3 baseline days in which subjects were scheduled to 8 hours of sleep and 16 hours of wake at their habitual times. During the sleep deprivation component of the protocol, the PVT was administered every 2 hours. The subjects from this protocol were not used in the training set; instead the PVT data from these subjects were used as an independent data set to determine classification labels for each test session included in the training set (described below).

(b) Field data: Two data sets were used. In these two data sets, the health, medication, pharmaceutical and caffeine use, and sleep schedule (including napping) behavior of the subjects were not controlled. For the first study, as described in (Lockley *et al.* 2004; Anderson *et al.* in press), PGY-1 medical interns were enrolled in a study designed to quantify the effects of extended duration work hours (24–30 hours) on sleep, alertness and the rates of medical errors among interns working in critical care units. This study tested the hypothesis that eliminating extended work shifts with an intervention (IV) schedule would increase sleep duration and reduce attentional failures as compared to a traditional intern schedule in which shifts were scheduled for up to 30 continuous hours every 3rd night (Q3 schedule). Additional details about the Q3 and IV schedules can be found in (Lockley *et al.* 2004). For this analysis, data from 34 interns (11 female, mean age 28.0 ± 1.8 years, range 24–32 years) were available. PVT data were collected intermittently in each intern during both schedules. The PVT task used in this study was the same 10-minute version as the task used for the laboratory studies described above. The intern data were used as an independent test set to determine whether a training set consisting of data from laboratory studies is able to classify individual performance impairment using data collected under real-world (field) conditions. All 34 subjects were used for the final analysis of the Q3 schedule; 1 subject on the IV schedule had inadequate sleep/wake data and was excluded.

For the second study, (Barger *et al.*, in press), all subjects were scientists or engineers working on a 24.65-hr Mars sol at the Science Operations Center in Tucson AZ in support of the Phoenix Mars Lander, which landed on Mars on May 25, 2008. Nineteen subjects total (6 female, mean age 36.8 ± 9.7 years, range 25–63 years) participated in a study to assess performance and alertness and sleep/wake patterns from actigraphy and sleep diaries, while living and working on the Mars sol schedule. Participants were asked to complete the previously validated 5-minute version of the PVT using a portable handheld device (Loh *et al.* 2004, Lamond *et al.* 2005, Roach *et al.* 2006) at least twice per day. Additionally, we used this data set to test whether results from a 10-minute PVT (from laboratory studies) can be used to classify results from a 5-minute PVT data (from real-world conditions). Seventeen subjects from the Phoenix Mars Lander group were used for this final analysis; 2 subjects had inadequate sleep/wake or PVT data and were excluded.

2.2 Description of the Pattern Recognition Algorithms

2.2.1 Feature Space—Pattern recognition algorithms require a set of data called the *feature space* to represent each object (i.e., each testing session from each individual) as a point in n-dimensional space. For this analysis, each object in the feature space was derived from an individual testing session that included the Psychomotor Vigilance Task (PVT), the Karolinska Sleepiness Scale (KSS), and the length of time awake and time of day when the testing session was administered. These features were chosen because the PVT and KSS are relatively easy to administer in a field setting, for example on a hand-held device. Additionally, there are large amounts of PVT and KSS data available from laboratory sleep deprivation studies for different lengths of time awake and circadian phase.

The PVT is an objective performance test that measures sustained attention to a visual stimulus presented at a high signal rate with a randomized inter-stimulus interval distributed uniformly from 1–9 seconds. Subjects are instructed to respond as quickly as possible once the visual stimulus appears on the screen; the reaction time (RT) to this stimulus is recorded and provided as feedback to the subject (Dorrian *et al.* 2005). In typical analysis of PVT data, a summary statistic, such as mean or median RT or the number of lapses (RT > 500 ms), is computed from each testing session and changes are tracked across sessions. The *distribution* of RT percentiles (5th–95th) collected during a 10-minute session of the PVT has also been used to compare performance changes across sessions (Santhi *et al.* 2007).

The KSS is a subjective measure that asks individuals to rate their sleepiness in the past 5 minutes on a 1 – 9 scale, with 1 indicating “very alert” and 9 indicating “Sleepy – great effort to keep awake – fighting sleep” (Åkerstedt and Gillberg 1990). The version of the KSS administered in all laboratory and field studies used in this analysis included descriptors on the odd numbers only.

The full feature space to be tested included 8 dimensions: 1) the mean fastest 10% and 2) median response times from PVT, 3) number of PVT lapses, 4) KSS score, 5) length of time awake (LOTA) and 6) time of day (TOD) at the time the testing session was administered, 7) age and 8) sex of the individual. TOD was binned across 24 hours into six 4-hour bins: 2:00–5:59, 6:00–9:59, 10:00–13:59, 14:00–17:59, 18:00–21:59, and 22:00–1:59.

The laboratory training set (feature space) consisted of 506 testing sessions (objects). A total of 33 subjects contributed to these 506 testing sessions. Thus, each individual represented in the training set was associated with one or more objects in the feature space. For example, a subject studied under a 52-hour sleep deprivation that completed a test session every 2 hours contributes 28 testing sessions to the training set.

2.2.2 Classification Labels—In order to use pattern recognition algorithms to classify the relative performance impairment for each individual during each test battery session, it was necessary to classify each of the 506 sessions in the laboratory training set with a label reflecting relative performance impairment. There is currently no “gold” standard or biomarker for defining performance impairment in response to sleep loss. It has been shown that several measures extracted from the PVT change in response to sleep deprivation in laboratory studies (e.g., Van Dongen *et al.* 2003; Belenky *et al.*, 2003; Wyatt *et al.*, 2004; Cohen *et al.*, 2010), and the number of PVT lapses has been unofficially accepted as a potential marker of performance impairment. To our knowledge, however, there is no evidence that PVT lapses map onto real-world functioning; no studies have reported that X number of lapses indicates Y% increase in, for example, motor vehicle accidents, medical errors or aviation errors. In fact, at least one study suggests that PVT median RT may be a better indicator of performance impairment than PVT lapses in medical residents on light vs. heavy call schedules (Arnedt *et al.* 2005). PVT lapses, furthermore, represent an arbitrary cut-off (500 msec) that does not reflect inter-individual differences in response speed. The 90th percentile of reaction times, a measure based on the entire RT distribution rather than an absolute value, has been shown to be a robust measure of performance impairment in laboratory data (Santhi *et al.* 2007). The mean slowest 10% RT, a summary statistic generated from each PVT session and reported in multiple publications (e.g., Wyatt *et al.* 2004; Grady *et al.* 2010; Anderson *et al.* in press), is similar to the 90th percentile measure. We hypothesize that tracking the change in mean slowest 10% RT provides a more robust measure of an individual’s change in performance over sleep deprivation than the number of PVT lapses. Therefore, to classify each testing session for the training set, we first computed the relative mean slowest 10% RT for each individual across a sleep deprivation episode by calculating the percent change in mean slowest 10% RT for each session from the best (i.e. lowest) mean slowest 10% RT score; this assumes the testing session with the lowest mean slowest 10% RT represents the individual’s best possible performance. These relative mean slowest 10% RTs were then categorized into three groups, labeled “1”, “2” or “3”, representing low performance impairment, moderate performance impairment and severe performance impairment, respectively. Each testing session was categorized into one of these three groups based on the following cut-off values: a testing session was labeled as “1” if the percent increase in mean slowest 10% was less than 25%, “3” if the percent increase was greater than 100% and “2” if the percent increase was between 25% and 100%. A 25% increase in mean slowest 10% corresponds to an increase of ~90ms in this measure and a 100% increase in corresponds to an increase of ~360ms. The number of lapses occurring

during testing sessions in which there was a ~25% increase in mean slowest 10% was between 0 and 17, and in which there was a ~100% increase in mean slowest 10% RT, was between 6 and 17 lapses. The label for each testing session in the training set, therefore, reflected performance changes across multiple testing sessions within an individual; the label could not be reliably derived from an outcome variable from a single testing session, such as the number of PVT lapses from a single session.

2.2.3 Algorithm selection—We tested two methods for pattern recognition: the k Nearest Neighbor (kNN) algorithm and a Naïve Bayes classifier. Both are supervised learning methods that compare a test object of unknown class to the training set, which consists of a set of objects with known classes. We chose to test these two methods for their relatively small computational requirements to highlight the feasibility of transitioning these methods to use in the field.

kNN classifies the unknown object by a majority vote of its k “nearest” neighbors, where “nearest” is defined by minimizing the distance between the test object and each object in the training set across the n-dimensions of the feature space. If $k = 1$, the object in the testing set is simply assigned to the class of its nearest neighbor. For the kNN implementation used in this analysis, Euclidean distance was used to determine the k nearest neighbors. This method was programmed and run in MatLab v7.11.0.

The Naïve Bayes classifier estimates the parameters of a defined probability distribution (e.g., Gaussian) during training. To test a new object, a posterior probability of that test object belonging to the class is computed. The Naïve Bayes classifier assumes that all features used in the feature space are conditionally independent; however, even when features are not independent, however, the Naïve Bayes classifier can still be used. For the Naïve Bayes implementation used in this analysis, both a normal (Gaussian) distribution and a kernel distribution were tested. When the Gaussian distribution is specified, the Naïve Bayes classifier assumes each feature is normally distributed for each class. This assumption is not made when the kernel distribution is used; instead, a separate kernel density estimate is computed for each class. The NaiveBayes function from the MatLab v7.11.0 Statistics Toolbox was used for this analysis.

2.2.4 Parameter validation and feature space selection—The appropriate use of pattern recognition algorithms requires a validation step before applying the chosen algorithms to the test set. The validation step includes choosing both the optimal parameters for the algorithm (e.g. parameter “k” for kNN, and the normal or kernel distribution for Naïve Bayes) and choosing the optimal feature space (out of the 8 available features) that best classifies the majority of the data. The test set cannot be used for the validation step, thus we used data from the training set to validate the parameters and choose the feature space. There are several methods for using the training set at the validation step. For example, one method is to set aside a proportion (e.g., one-third) of the training set as a validation set, and then use the validation set as a mock test set to optimize parameters and the feature space to the remainder of the training set (e.g., the other two-thirds of the data). For this analysis, however, we used the leave-one-out method to generate a unique validation set for each subject: each of the 33 subjects was tested on a subset of the full training set (506 test sessions) that omitted their own test sessions but included all of the test sessions from each of the remaining 32 subjects (e.g., Subject 1 contributing 28 test sessions to the full training set would be tested against a subset of training data containing only the 478 other test sessions).

To further improve on this validation approach, a method called bootstrap aggregation (or “bagging”) was employed. The bagging method improves classification accuracy and

reduces variance (Witten and Frank 2005). From each unique validation set created for each subject, we further generated 100 training sets that included a subset of test sessions from each validation set. Each of these 100 subset validation sets contained 150 test sessions randomly sampled with replacement from the validation set. For example, for Subject 1, their validation set contained 478 test sessions from each of the other 32 subjects, and each of their 100 subset validation sets contained 150 test sessions from this sample of 478 test sessions. Thus, for Subject 1, each test session (a total of 28 test sessions) belonging to Subject 1 was classified independently on 100 subset validation sets, which each generated a classification label for that test session, resulting in 100 classifications for each of the 28 test sessions for Subject 1. The final predicted classification label for each of these 28 test sessions was chosen by majority rule: for example, if the test session was classified as a '1' for 60 of the subset validation sets, '2' for 30 of the subset validation sets and '3' for 10 of the subset validation sets, the final predicted classification label for that test session for that subject would be chosen as '1'. See Supplementary Figure 1 for further details.

The validation steps just outlined above were used to determine the optimal value of k for the kNN algorithm and the optimal probability distribution – normal or kernel – to use for the Naïve Bayes classifier. For the kNN algorithm, general practice limits the value of k to less than the square root of the number of objects (in this case, $\sqrt{506} \approx 23$) in the full training set, and thus values of $k = 1$ to $k = 22$ were tested. Values of k in multiples of 2 and 3 were, however, omitted to avoid tie-breakers among the three classification groups. For kNN the value of k which resulted in the highest percentage of correctly classified test sessions across all subjects was chosen as the optimal k to be used for running the algorithm on the test set data. For the Naïve Bayes classifier, the probability distribution which resulted in the highest percentage of correctly classified test sessions across all subjects was chosen as the optimal distribution to be used for running the algorithm on the test set data

Once optimal parameters were chosen for each of the algorithms, it was necessary to determine the optimal feature space from the full feature space to use for classification. The same validation procedures outlined above were used to create the appropriate subset validation sets for each of the 33 subjects in the full training set. To choose the optimal feature space from the set of 8 features available, a method called forward feature selection was used. In the first step of forward feature selection, each of the 8 features (PVT mean fastest 10%, median and lapses, KSS score, LOTA, TOD, age, sex) were independently used to classify all of the data in the training set (using the leave-one-out method and bagging as described above). The one feature with the highest classification percentage (i.e., that maximized the percent of testing sessions in the training set that were correctly classified) was selected. In the next step, each of the remaining features was paired with the selected feature from the first step and all of the data in the training set were re-classified. For example, if LOTA was found to provide the highest classification percentage at the first step, then at the second step of forward feature selection, the following 2-dimensional feature spaces were tested: LOTA and PVT mean fastest 10% RT, LOTA and median RT, LOTA and lapses, LOTA and KSS, LOTA and TOD, LOTA and age, LOTA and sex. The feature pair with the highest classification percentage was chosen. At the third step of forward feature selection, this feature pair (e.g. LOTA and KSS) was paired with the remaining features (i.e., PVT mean fastest 10% RT, median RT, lapses, TOD, age, sex) Feature selection continued in this way, testing the addition of each feature one by one to the optimal feature space selected at the previous forward selection step, until the classification percentage no longer improved.

2.2.5 Testing on field data—The optimal parameters and feature spaces chosen for the kNN and Naïve Bayes algorithms were determined from validation against the laboratory training set. These optimized parameters and feature spaces were used to classify the data in

our two field-collected data sets. Each object in the test set, which represented a test session completed by an individual at a given LOTA and TOD, was compared to all 506 testing sessions (33 subjects) in the laboratory training set. In a true test set, the true class of a test object would be unknown, and the purpose of the analysis would be to predict the class to which the object belongs. For this analysis, however, all objects in the test set were classified *a priori* (see above) in order to report classification sensitivity and specificity results.

To compare the ability of the kNN and Naïve Bayes algorithms to classify correctly each object in our test set, we computed sensitivity and specificity scores. Sensitivity measures the proportion of true positives that are correctly identified as such. To compute the sensitivity value for testing sessions labeled as “1” (low performance impairment), for example, we computed the proportion of testing sessions categorized as “1” that were correctly predicted by the algorithm as belonging to “1”. For example, if 100 testing sessions were *a priori* labeled as “1” and the algorithm correctly identified 50 of these testing sessions as belonging to “1”, then the sensitivity of the algorithm for classifying “1” would be 50%. Specificity measures the proportion of true negatives that are correctly identified as such. To compute the specificity value for testing sessions labeled as “1”, for example, we computed the proportion of testing sessions categorized as “2” or “3” that were not identified by the algorithm as “1” by summing the number of testing sessions that were correctly predicted as “2” or “3” and dividing this sum by the total number of testing sessions originally labeled as “2” or “3” in the test set. For example, if 50 testing sessions were labeled in the test set as “2” and 25 were labeled in the test set as “3”, and the algorithm correctly predicted 60 of those testing sessions as not belonging to “1” (meaning 15 of those testing sessions were incorrectly predicted as belonging to “1”), then the specificity for “1” would be 80%. The false positive rate for “1” can be computed as the specificity value subtracted from 100%; for this example, the false positive rate would be 20%. Both sensitivity and specificity were computed in this way for each label “1”, “2” and “3”.

Positive predictive value (PPV) and negative predictive value (NPV) were also computed for classification results from each algorithm. PPV is often used in diagnosis of disease and reflects the probability that a positive test result reflects the underlying condition being tested. A PPV of 100% for a disease indicates that all patients that tested positive for the disease were found to have the disease (i.e., no false positives). In contrast, NPV reflects the probability that a negative result means the patient does not have the disease. An NPV of 100% for a disease indicates that all patients that tested negative for the disease were found to not have the disease (i.e., no false negatives). For our classification results, for example, PPV for “1” was computed as the number of test sessions labeled in the test set as “1” that were also predicted as “1” (true positives) divided by all the test sessions predicted as belonging to “1” (true positives and false positives). For example, if 100 test sessions were predicted as “1”, and 90 of these were labeled in the test set as “1” while 10 were labeled in the test set as “2” or “3”, then the PPV for “1” would equal 90%. NPV for “1” was computed as the number of test sessions labeled in the test set to “2” or “3” that were also predicted as “2” or “3”, divided by the number of all test sessions predicted by the algorithm as “2” or “3” (a number which may include test sessions that are labeled in the test set as “1”). For example, if 100 test sessions were predicted by the algorithm as “2” or “3” and 80 of these test sessions were labeled as “2” or “3” (i.e. not labeled as “1”), then the NPV for “1” would equal 80%. Both PPV and NPV were computed in this way for each label “1”, “2” and “3”.

2.3 Comparison of Pattern Recognition Algorithm Results to Classification Based on Lapses

The goals of using pattern recognition algorithms for this analysis were 1) to demonstrate the ability to classify performance impairment from a single observation of the state of an individual's neurobehavioral performance level, i.e. a single testing session and 2) to use a measure of performance impairment that relates the current level of performance to the individual's baseline or optimal performance, e.g. the relative mean slowest 10% RT, rather than an absolute value, such as the number of PVT lapses in a testing session. To show the benefit of using a relative vs. absolute measure of an individual's performance impairment, plus additional information derived from the individual's age, sex, LOTA, and subjective sleepiness assessment, we compared the ability of our pattern recognition algorithms to classify performance impairment from a single testing session in data collected from the field to predictions based not on our algorithm (the feature space of which includes PVT lapses as a potential predictor), but on absolute PVT lapses only. Although PVT lapses have been accepted as a potential marker of performance impairment under total sleep deprivation studies (e.g., Van Dongen et al., 2003), to our knowledge no studies have been conducted to correlate the number of PVT lapses with a real-world outcome (e.g., X PVT lapses equates to Y% increase in motor vehicle accidents). Thus, we chose to designate the three levels of performance impairment ("1" low, "2" medium and "3" severe) using cut-off values based on the number of hours awake. These cut-off values were determined using an independent laboratory data set (i.e., Laboratory study 3, which was not included in the laboratory training set described above) (Figure 1). The average number of PVT lapses was computed after 16 and 32 hours of wakefulness: 16 hours awake was chosen as the initial cut-off based on the 2:1 wake:sleep ratio on a normal 24-hr day and the fact that performance has been documented to deteriorate rapidly after 16 hours of continuous wakefulness (Jewett and Kronauer 1999). At 16 hours awake, the average number of PVT lapses in all 12 subjects was ~3; therefore all testing sessions in which the number of PVT lapses was 3 were classified to group "1", representing low performance impairment. After 32 hours awake, the average number of PVT lapses observed across these 12 subjects was ~21, and therefore all testing sessions with the number of PVT lapses 21 were predicted as belonging to group "3" representing severe performance impairment. All testing sessions with PVT lapses between 3 and 21 were predicted as belonging to group "2" representing medium performance impairment. Sensitivity, specificity, PPV and NPV were computed for these classification predictions and compared to those obtained from using the pattern recognition algorithms for the field-collected data test sets.

3.0 Results

3.1 Classification Results for Laboratory Data (Training Set)

Our first step in using the pattern recognition algorithms involved choosing optimal parameters for the kNN and Naïve Bayes methods. The optimal value of "k" for the kNN algorithm was determined using the full feature space (8 dimensions) to classify all the data in the training set (33 subjects) using a leave-one-out method and bagging for each value of k. An optimal value of $k = 1$ was found, which resulted in 67% correct classification of all the data in the training set. The optimal Naïve Bayes distribution was also determined using the full feature space (8 dimensions) to classify all the data in the training set (33 subjects) using a leave-one-out method for each distribution. The kernel distribution was found to be optimal and resulted in 72% correct classification of all data in the training set. Bagging was not used for the Naïve Bayes algorithm due to the increased computational time, which was ~6 seconds without bagging and over 1 hour with bagging. No considerable increase in classification accuracy was found when bagging was used.

The next step was to determine the best combination of the 8 available features to use for our final feature space using forward feature selection. For the kNN algorithm, feature selection resulted in an optimal feature space consisting of PVT lapses, LOTA, TOD, KSS and sex, with a 73% correct classification rate. For the Naïve Bayes classifier, feature selection resulted in the same optimal feature space with a 75% correct classification rate.

Sensitivity, specificity, PPV and NPV results were computed separately for each of the groups 1, 2, and 3 for the optimal feature space for each method. These results are presented in Table 1. Table 1 also shows the “confusion matrix” of the raw values of actual vs. predicted classifications for each group for each method. kNN and Naïve Bayes produce similar predictions across all three classification groups.

It is possible to estimate probabilities of future susceptibility given current status. Only one of the 33 subjects in the training set was labeled with low performance impairment (“1”) on all test sessions. For the remaining 32 subjects, once a session was labeled as a “2” or a “3” (starting anywhere from 2 hours to 26 hours after wake), 84% of subsequent testing sessions for that wake episode were also labeled as “2” or “3”.

3.2 Classification of Real World Data (Test Set)

3.2.1 Intern data set—For the intern data set with Q3 and IV schedules, age, sex, PVT mean fastest 10% RTs, medians, and lapses, LOTA and TOD data were available; KSS was not. As naps were allowed in this field study, LOTA was determined by computing the length of time since any sleep episode >30 minutes. The optimal feature spaces derived above from the laboratory training set were used in the kNN and Naïve Bayes algorithms to classify testing sessions from the Q3 and IV datasets; KSS was omitted from the feature space because it was not collected in these test sets.

In post-hoc analysis of data from the intern Q3 (standard) schedule, 329 out of the 940 sessions across the 34 subjects were labeled as ‘1’ using our labeling criteria described above (i.e. derived from relative mean slowest 10% RTs), whereas 409 were labeled as ‘2’ and 202 as ‘3’. In post-hoc analysis of data from the intern IV (intervention) schedule, 319 out of the 865 PVT trials across 34 subjects were labeled as ‘1’, while 431 were labeled as ‘2’ and 115 were labeled as ‘3’. The sensitivity, specificity, PPV and NPV results for both the kNN and Naïve Bayes methods for both the Q3 and IV schedules are presented in Table 2. The kNN algorithm correctly classified 49% of all test sessions on the Q3 schedule and 51% of all sessions on the IV schedule; the Naïve Bayes classifier correctly classified 51% of all sessions on the Q3 schedule and 51% of all sessions on the IV schedule. Using PVT lapses alone to classify the data, based on the cut-off values discussed above in section 2.3, 52% of sessions on Q3 and 56% of sessions on IV were classified correctly. Both methods, either employing one of the algorithms (kNN or Naïve Bayes) or the classification based on PVT lapse cut-offs, were most effective at correctly classifying testing sessions labeled as “1”, and considerably underestimated testing sessions labeled as “2” and “3”. The confusion matrices for the kNN and Naïve Bayes algorithms are presented in Table 3. Confusion matrices for the classification based on PVT lapses are presented in Table 4.

Since our optimal feature space used in the kNN and Naïve Bayes algorithms was based on classification of laboratory data, we considered the possibility that different features may be optimal for classifying field-collected data. By forward feature selection using laboratory data for training on the kNN algorithm, we found the optimal feature space for classifying the Q3 data included lapses and mean fastest 10% RTs only, and the optimal feature space for classifying the IV data included median, lapses, TOD and sex. Similarly, a different optimal feature space emerged for the Naïve Bayes algorithm, including only lapses and age for the Q3 data and lapses only for the IV data. Using these optimal feature spaces improved

the overall percentage correctly classified by only 1–5%, but considerably decreased the number of test sessions for which performance impairment was underestimated. Table 5 shows the results of these feature spaces found post-hoc from the intern data.

3.2.1.1 Chronic sleep deprivation effects on PVT performance on the Intern Q3

schedule: Anderson *et al.* (in press) compared acute vs. chronic effects of sleep deprivation on reaction time when subjects were on the Q3 schedule by comparing the change in RT distributions at the beginning of the first work shift (Extended Duration Work Shift or EDWS 1) on Q3 to the end of the first work shift (acute effect) and PVT performance at the beginning and end (acute effect) of the sixth work shift (EDWS 6) on Q3, which occurred ~3 weeks after the first work shift in EDWS 1 (chronic effect). They found a significant acute effect for both shifts (EDWS 1 and EDWS 6) and a chronic effect between EDWS 1 and EDWS 6. To determine whether a chronic effect could be observed in our classification results (Table 3), we grouped the classification results across the same test sessions included in their analysis. We used only subjects who had an equal number of test sessions in EDWS 1 and EDWS 6; if extra sessions were recorded in either EDWS, they were omitted from this analysis. We found that from EDWS 1 to EDWS 6, the number of test sessions classified as 1 was decreased and that the number of sessions classified as either 2 or 3 increased (Figure 2) for both the kNN and Naïve Bayes methods as well as the classification based on PVT lapses. Using a chi-square test for goodness-of-fit, assuming a null hypothesis that the proportion of 1, 2 and 3 does not change from EDWS 1 to EDWS 6, we observed a significant change from EDWS 1 to EDWS 6 (kNN $\chi^2 = 23.79$, $p < 0.001$; Naïve Bayes $\chi^2 = 83.73$, $p < 0.001$; PVT Lapses $\chi^2 = 43.49$, $p < 0.001$), in accordance with the results reported in Anderson *et al.* (in press).

3.2.2 Non-24-hr Phoenix Mars work schedule—For this dataset, age, sex, PVT, LOTA, KSS and TOD data were available. As in the Q3/IV test data, the optimal feature spaces derived above from the laboratory training set were used in the kNN and Naïve Bayes algorithms to classify testing sessions and included lapses, LOTA, TOD, KSS and sex.

In post-hoc analysis of the individuals, 203 out of the 1239 sessions across 17 subjects were labeled as “1”, while 565 were labeled as “2” and 471 as “3”. The sensitivity, specificity, PPV and NPV results for both the kNN and Naïve Bayes methods are presented in Table 2. The kNN algorithm correctly classified 33% of all test sessions on the non-24-hour schedule and the Naïve Bayes classifier correctly classified 27% of all sessions. In contrast, classifying the data using the absolute number of PVT lapses alone, only 24% of sessions on the non-24-hour schedule were classified correctly. Both the kNN and Naïve Bayes algorithms were more effective at correctly classifying test sessions labeled as “3” than the classification based on PVT lapses. The confusion matrices for the kNN and Naïve Bayes algorithms are presented in Table 3. Table 4 shows the confusion matrix for classification based on PVT lapses of the non-24h-schedule test set.

For this field-collected data set, we again tested the possibility that a different “optimal” feature space existed than the one based on laboratory data. For the kNN algorithm, the optimal feature space for classifying the non-24-hour schedule data included median PVT RTs and KSS, and for the Naïve Bayes algorithm only age and sex were included in the optimal feature space. Using these feature spaces improved overall classification accuracy to 40% and 42% for the kNN and Naïve Bayes algorithms, respectively, and considerably decreased the number of test sessions for which performance impairment was underestimated. Table 5 shows the results of these feature spaces found post-hoc from the non-24-hour schedule data.

4.0 Discussion

Currently, performance impairment is defined by absolute performance of an individual at one point in time compared to group averages. Such methods do not take into account individual differences in baseline performance, and impairment is often detected only *after* performance has already declined to a dangerous level. As performance can deteriorate rapidly depending on the sleep-wake and circadian history (Cohen *et al.* 2010), it would be useful to detect the signature of an individual experiencing *relative* impairment to identify individuals that are on the verge of rapid deterioration in performance *before* they actually reach dangerous levels. In this paper, we tested the ability of pattern recognition algorithms to classify impairment in response to sleep loss, using features extracted from a single testing session, after being trained on a separate data set. In order to be able to use individual testing sessions to classify an individual's performance impairment, we first needed to construct an objective classification scheme. There is significant inter-individual variability in performance, both at baseline and across sleep deprivation. We chose a classification scheme that classified individuals on the percent change in mean slowest 10% RTs on the PVT during sleep deprivation. Each testing session for each individual was labeled as reflecting low, medium or severe performance impairment using 25% and 100% relative increases in mean slowest 10% RT as cut-offs to determine classification to each impairment level. We tested two pattern recognition algorithms, kNN and Naïve Bayes classifier. Although both algorithms performed similarly, overall the kNN algorithm is simpler to implement and computation time is less than for the Naïve Bayes classifier.

It is important to note that the classification of each testing session takes into account only current information about the individual's behavior. For example, if the pattern recognition algorithm classifies a test object as "2", this indicates that the person's current performance level is equivalent to someone with moderate performance impairment relative to their best or optimal performance, but does not indicate whether the individual will continue to respond poorly to subsequent sleep loss. We know from analysis of our training set data taken from laboratory experiments, however, that once an individual has a testing session labeled as "2" or "3", 84% of their subsequent testing sessions for that wake episode will also be labeled as "2" or "3", indicating that they will probably continue to experience medium to severe performance impairment without an intervention such as sleep.

One shortcoming of our approach is that we are using PVT measures both in our feature space and in our classification labels. Using forward feature selection methods, it was found that PVT measures combined with additional information (e.g., TOD) were better than PVT measures alone at predicting performance impairment. Additionally, we used relative mean slowest 10% RTs rather than an absolute measure such as lapses based on a 500 ms threshold, because some individuals at baseline who are presumably well-rested may have a similar mean slowest 10% RT score to individuals who are sleep-deprived; while their baselines are slower, they are not necessarily at risk of rapid deterioration in performance compared to someone with a faster baseline that is starting show relative decline from fatigue. Although the absolute number of PVT lapses for one testing session are a useful measure to classify performance impairment, as evidenced by its inclusion in our optimal feature spaces, our analysis here has shown that the absolute number of PVT lapses over a single testing session cannot alone determine an individual's relative performance impairment, particularly in our field-collected data when subjects show severe performance impairment in the mean slowest 10% RT but not in the number of PVT lapses. This suggests that lapses may not map onto real-world functioning, since the field-collected data sets demonstrate that individuals can have a low number of PVT lapses but still show large decrements in performance based on other measures. A previous study of medical residents found that the number of PVT lapses had no significant change between a light vs. heavy

call schedule, but that PVT median RT significantly increased (Arnedt et al. 2005). Additionally, the cut-off of 500 ms to define a lapse may not be appropriate for all individuals. Distribution of reaction times shows large inter-individual differences, including age and sex effects (Wyatt et al. 2004, Blatter et al. 2006, Duffy et al. 2009). Furthermore, time to respond on the PVT depends on whether eyes are open, closed or looking away at the time of stimulus presentation and a lapse just above 500 ms may represent a different mechanism of attentional failure than a lapse of several seconds (Anderson et al. 2010).

We tested our validated feature space and training set from laboratory data on data collected in two field studies: a population of hospital interns who frequently work extended hours or at night and a population of ground crew working the Phoenix Mars mission on a 24.65-hr Mars sol schedule. Our pattern recognition method was able to correctly classify subjects labeled as “1” low performance impairment 80–98% of the time, depending on which algorithm was used. Our pattern recognition methods were less successful at predicting “2” medium and “3” severe performance impairment in the field-collected data sets, and often underpredicted the level of performance impairment. The PVT administered during the non-24-hr Mars sol schedule, however, was the 5-minute version of the task. The PVT shows a time-on-task effect, with longer RTs occurring in the later minutes of the task (Paus et al., 1997, Tucker et al. 2009). Although the 5-minute version of the PVT has been previously validated (Loh et al. 2004, Lamond et al. 2005, Roach et al. 2006), a difference in the distribution of RTs may explain why our training set and classification labels, which were based on the 10-minute version of the PVT, were less successful predicting medium and severe performance impairment in the subjects on the non-24-hr Mars sol schedule.

A possible limitation of the training set used here to optimize parameters and feature spaces of our pattern recognition algorithms is that it comes from laboratory data in which subjects were not allowed any substances including caffeine, nicotine or alcohol, whereas in both the intern and non-24-hr schedules subjects had free access to stimulants and many reported using them. Using laboratory data, therefore, to quantify performance impairment in populations which have access to performance-altering substances may not always be appropriate. These differences may also explain why different feature spaces were found to be better predictors when applied to the field-collected data compared to the feature spaces derived from the laboratory data. Our analysis suggests that features that are important indicators of performance impairment in a laboratory setting (e.g. LOTA, TOD) may not be important indicators of impaired performance under real-world conditions. For example, the TOD as a proxy for circadian phase as defined in this analysis (i.e., six 4-hr bins over a 24-hr cycle) is an inappropriate measure for the non-24-hr Phoenix Mars schedule because the majority of subjects were in phase with their work schedule, which moved forward 0.65-hr per day (Barger et al. in press). Unfortunately, “optimal” feature spaces must be determined *a priori* from existing data, and therefore more work should be done to identify features that are robust predictors of performance impairment both in the laboratory and the field. Future work should also develop training sets that include laboratory and non-laboratory data, indicating data in which caffeine and other stimulants were used.

Another limitation of the current training set is that it includes only data collected from one type of objective performance test. Although there are other neurobehavioral tasks administered in the laboratory, many of these cannot be administered easily in the field. A benefit of using pattern recognition algorithms is that there is no limit on the amount or type of data that can be included in the feature space. It may therefore be possible to include in the feature space other types of measurements that can be collected in a field setting to improve further the classification results that we have reported here. The percentage of eyelid closure time (PERCLOS), for example, has been shown as a potentially effective predictor of low vigilance (Abe et al. 2011).

In conclusion, we have presented a pattern recognition algorithm that makes use of data from laboratory studies to classify performance impairment in response to sleep loss in other individuals working at their jobs using data from a single collection period. We have shown that this method can be used in conjunction with administration of the PVT and subjective assessments of sleepiness in non-laboratory conditions to predict when individuals have low performance impairment; however, more analysis is necessary to improve the algorithms to predict more accurately when individuals reach medium and severe performance impairment levels, particularly under real-world conditions. The ability to accurately classify low impairment is useful, however, as individuals that are not in this classification can be flagged as a potential safety concern, and limited resources may be devoted to more direct supervision of their actual job responsibilities to judge safety of their performance or to choose another individual to perform the task. Furthermore, while our definitions of medium and severe performance impairment reflect increases in RT of ~90 msec and ~360 msec, respectively, it is not clear how such an increase in RT translates to the risk of accidents and errors in operational settings; more work in this area is needed to correlate objective measures of performance impairment such as the PVT with such outcomes.

Once the algorithms have been improved, it would be advantageous to implement these methods in a software device that could be used in a real-world setting to decide which individuals may not be safe to continue working. An important open question is what steps to take once an individual has been classified into a medium or severe performance impairment category. An ideal scenario would be to remove that individual from further work (and replace them with another individual if available) until adequate sleep could be obtained, although in operational settings this option may not be safe or feasible. Future work, therefore, should focus on interpreting the results of these algorithms in conjunction with existing fatigue management scheduling tools and mathematical models which predict levels of neurobehavioral performance and alertness under different sleep/wake and circadian phase combinations. These tools can be used to predict the relative effectiveness of various countermeasures, such as naps, caffeine or light, in an individual given their current level of performance impairment. Once a countermeasure has been given, the algorithm can be implemented again to determine whether the countermeasure has improved performance and alertness to an acceptable level.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The laboratory studies used in this analysis were supported by AFOSR FA9550-06-0080/O5NL132 and NIH P01-AG009975 and NIH RC2-HL101340-0. The intern study was supported by grants from the Agency for Healthcare Research and Quality (RO1 HS12032), affording data confidentiality protection by federal statute (Public Health Service Act; 42 U.S.C.) and the National Institute of Occupational Safety and Health within the U.S. Centers for Disease Control (R01 OH07567), which provided a Certificate of Confidentiality for data protection, and by the Department of Medicine, Brigham and Women's Hospital, the Division of Sleep Medicine, Harvard Medical School. Some studies were performed in the Brigham and Women's Hospital General Clinical Research Center, Harvard Clinical and Translational Science Center, supported by the National Center for Research Resources (M01 RR02635 and 1 UL1 RR025758). The Phoenix Mars study was supported by grant NNX08AD66A from the National Aeronautics and Space Administration. MSH is supported by T32-HL07901. EBK was supported by K02-HD045459, RC2-HL101340 and K24-HL105663. EBK, LKB and SWL were supported in part by the National Space Biomedical Research Institute, through the National Aeronautics and Space Administration NCC 9-58 HPF HFP01603 and HFP01701 (EBK).

References

- Abe T, Nonomura T, Komada Y, Asaoka S, Sasai T, Ueno A, Inoue Y. Detecting deteriorated vigilance using percentage of eyelid closure time during behavioral maintenance of wakefulness tests. *Int. J. Psychophysiol.* 2011; 82(3):269–274. [PubMed: 21978525]
- Akerstedt T, Folkard S, Portin C. Predictions from the three-process model of alertness. *Aviat.Space Environ.Med.* 2004; 75(Suppl)(3):A75–A83. [PubMed: 15018267]
- Åkerstedt T, Gillberg M. Subjective and objective sleepiness in the active individual. *Int J Neurosci.* 1990; 52(1–2):29–37. [PubMed: 2265922]
- Anderson C, Wales AWJ, Horne JA. Pvt lapses differ according to eyes open, closed, or looking away. *Sleep.* 2010; 33(2):197–204. [PubMed: 20175403]
- Arnedt JT, Owens J, Crouch M, Stahl J, Carskadon MA. Neurobehavioral performance of residents after heavy night call vs after alcohol ingestion. *Journal of the American Medical Association.* 2005; 294(9):1025–1033. [PubMed: 16145022]
- Belenky G, Wesensten NJ, Thorne DR, Thomas ML, Sing HC, Redmond DP, Russo MB, Balkin TJ. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study. *Journal of Sleep Research.* 2003; 12:1–12. [PubMed: 12603781]
- Blatter K, Graw P, Munch M, Knoblauch V, Wirz-Justice A, Cajochen C. Gender and age differences in psychomotor vigilance performance under differential sleep pressure conditions. *Behavioural Brain Research.* 2006; 168(2):312–317. [PubMed: 16386807]
- Cohen DA, Wang W, Wyatt JK, Kronauer RE, Dijk D, Czeisler C, Klerman EB. Uncovering residual effects of chronic sleep loss on human performance. *Science Translational Medicine.* 2010; 2(14):14ra3.
- Dijk DJ, Czeisler CA. Paradoxical timing of the circadian rhythm of sleep propensity serves to consolidate sleep and wakefulness in humans. *Neurosci Lett.* 1994; 166(1):63–68. [PubMed: 8190360]
- Dijk DJ, Duffy JF, Czeisler CA. Circadian and sleep/wake dependent aspects of subjective alertness and cognitive performance. *Journal of Sleep Research.* 1992; 1:112–117. [PubMed: 10607036]
- Dorrian, J.; Rogers, NL.; Dinges, DF. Psychomotor vigilance performance: Neurocognitive assay sensitive to sleep loss. In: Kushida, CA., editor. *Sleep deprivation. Clinical issues, pharmacology, and sleep loss effects.* New York: Marcel Dekker; 2005. p. 39-70.
- Duffy JF, Willson HJ, Wang W, Czeisler CA. Healthy older adults better tolerate sleep deprivation than young adults. *J Am Geriatr Soc.* 2009; 57(7):1245–1251. [PubMed: 19460089]
- Grady S, Aeschbach D, Wright KP Jr, Czeisler CA. Effect of modafinil on impairments in neurobehavioral performance and learning associated with extended wakefulness and circadian misalignment. *Neuropsychopharmacology.* 2010; 35(9):1910–1920. [PubMed: 20505660]
- Hursh SR, Redmond DP, Johnson ML, Thorne DR, Belenky G, Balkin TJ, Storm W, Miller JC, Eddy DR. Fatigue models for applied research in warfighting. *Aviation, Space, and Environmental Medicine.* 2004; 75:A44–A53.
- Jewett ME, Kronauer RE. Interactive mathematical models of subjective alertness and cognitive throughput in humans. *Journal of Biological Rhythms.* 1999; 14(6):588–597. [PubMed: 10643756]
- Klerman EB, Dijk DJ. Age-related reduction in the maximal capacity for sleep - implications for insomnia. *Current Biology.* 2008; 18(15):1118–1123. [PubMed: 18656358]
- Lamond N, Dawson D, Roach GD. Fatigue assessment in the field: Validation of a hand-held electronic psychomotor vigilance task. *Aviat.Space Environ.Med.* 2005; 76(5):486–489. [PubMed: 15892548]
- Lockley SW, Cronin JW, Evans EE, Cade BE, Lee CJ, Landrigan CP, Rothschild JM, Katz JT, Lilly CM, Stone PH, Aeschbach D, Czeisler CA. Effect of reducing interns' weekly work hours on sleep and attentional failures. *N Engl J Med.* 2004; 351(18):1829–1837. [PubMed: 15509816]
- Loh S, Lamond N, Dorrian J, Roach G, Dawson D. The validity of psychomotor vigilance tasks of less than 10-minute duration. *Behav Res Methods Instrum.Comput.* 2004; 36(2):339–346. [PubMed: 15354700]

- Mallis MM, Mejdal S, Nguyen TT, Dinges DF. Summary of the key features of seven biomathematical models of human fatigue and performance. *Aviation, Space, and Environmental Medicine*. 2004; 75(Suppl)(3):A4–A14.
- Paus T, Zatorre RJ, Hofle N, Caramanos Z, Gotman J, Petrides M, Evans AC. Time-related changes in neural systems underlying arousal and attention during the performance of an auditory vigilance task. *J Cogn Neurosci*. 1997; 9(3):392–408.
- Rajaraman S, Gribok AV, Wesensten NJ, Balkin TJ, Reifman J. Individualized performance prediction of sleep-deprived individuals with the two-process model. *J Appl Physiol*. 2008; 104(2):459–468. [PubMed: 18079260]
- Roach GD, Dawson D, Lamond N. Can a shorter psychomotor vigilance task be used as a reasonable substitute for the ten-minute psychomotor vigilance task? *Chronobiology International*. 2006; 23(6):1379–1387. [PubMed: 17190720]
- Santhi N, Horowitz TS, Duffy JF, Czeisler CA. Acute sleep deprivation and circadian misalignment associated with transition onto the first night of work impairs visual selective attention. *PLoS ONE*. 2007; 2(11):e1233. [PubMed: 18043740]
- Tucker AM, Basner RC, Stern Y, Rakitin BC. The variable response-stimulus interval effect and sleep deprivation: an unexplored aspect of psychomotor vigilance task performance. *Sleep*. 2009; 32(10):1393–1395. [PubMed: 19848367]
- Van Dongen HPA. Comparison of mathematical model predictions to experimental data of fatigue and performance. *Aviation, Space, and Environmental Medicine*. 2004; 75(Suppl)(3):A15–A36.
- Van Dongen HPA, Baynard MD, Maislin G, Dinges DF. Systematic interindividual differences in neurobehavioral impairment from sleep loss: Evidence of trait-like differential vulnerability. *Sleep*. 2004; 27:423–433. [PubMed: 15164894]
- 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(2):117–126. [PubMed: 12683469]
- Van Dongen HPA, Mott CG, Huang JK, Mollicone DJ, Mckenzie FD, Dinges DF. Optimization of biomathematical model predictions for cognitive performance impairment in individuals: Accounting for unknown traits and uncertain states in homeostatic and circadian processes. *Sleep*. 2007; 30(9):1129–1143. [PubMed: 17910385]
- Witten, IH.; Frank, E. *Data Mining: Practical Machine Learning Tool and Techniques*, second edition. San Francisco: Morgan Kaufmann; 2005. Transformations: Engineering the input and output; p. 285-344.
- Wyatt JK, Cajochen C, Ritz-De Cecco A, Czeisler CA, Dijk DJ. Low-dose repeated caffeine administration for circadian-phase-dependent performance degradation during extended wakefulness. *Sleep*. 2004; 27(3):374–381. [PubMed: 15164887]

Highlights

- Insufficient sleep results in cognitive performance impairment
- There is widespread inter-individual variability in response to sleep loss
- Current methods to assess impairment rely on tracking an individual over time
- Our proposed methods can classify level of impairment from one field assessment
- Such methods may identify individuals at risk before dangerous levels are reached

\$watermark-text

\$watermark-text

\$watermark-text

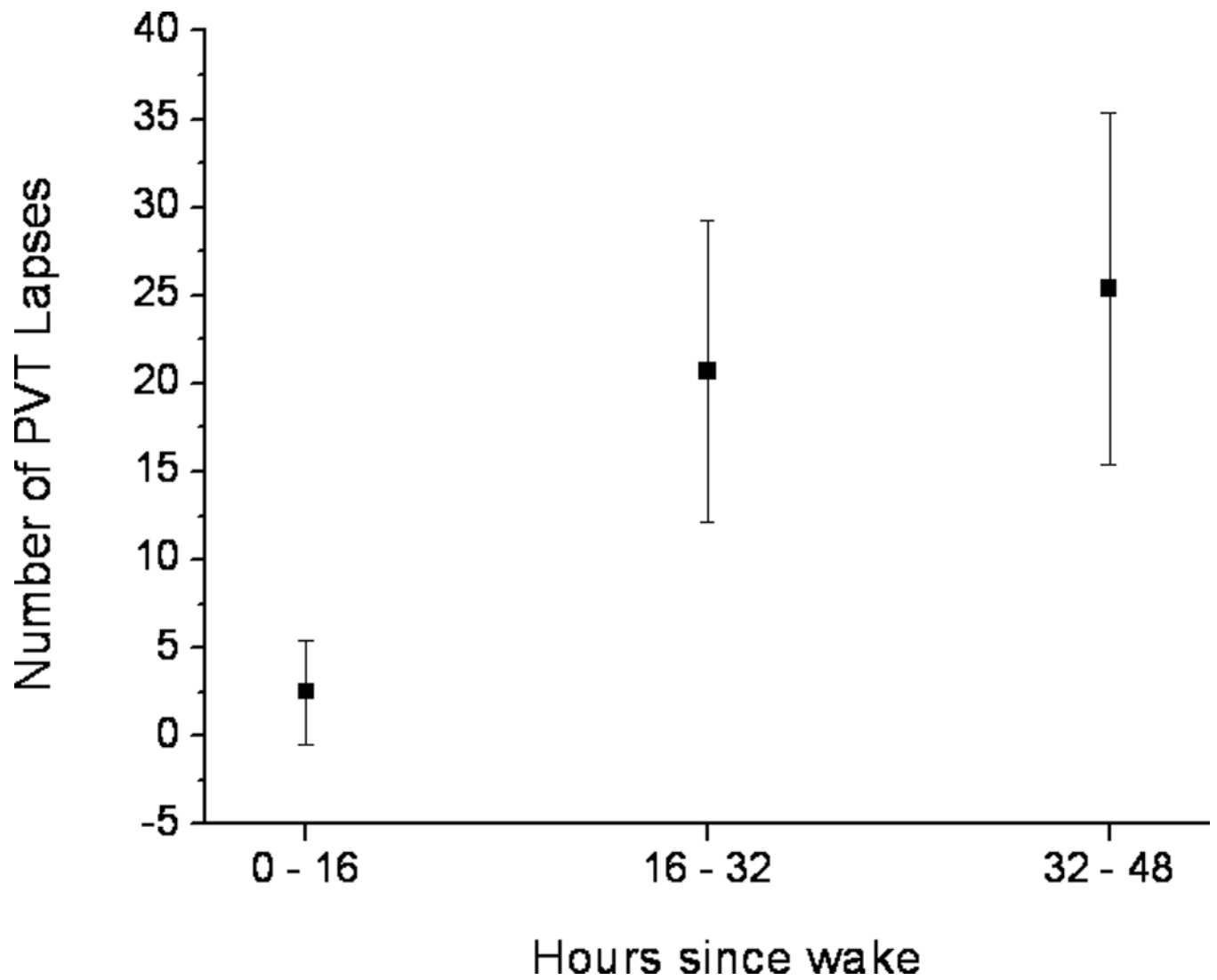


Figure 1. Number of PVT lapses as a function of length of time awake

Twelve subjects underwent a 50-hr sleep deprivation where the PVT was administered every 2 hours. The average number of PVT lapses (RT > 500 msec) were computed over the first, second and third 16 hours of wakefulness for each subject and then across subjects. The number of PVT lapses for each time awake bin is plotted as mean \pm standard deviation.

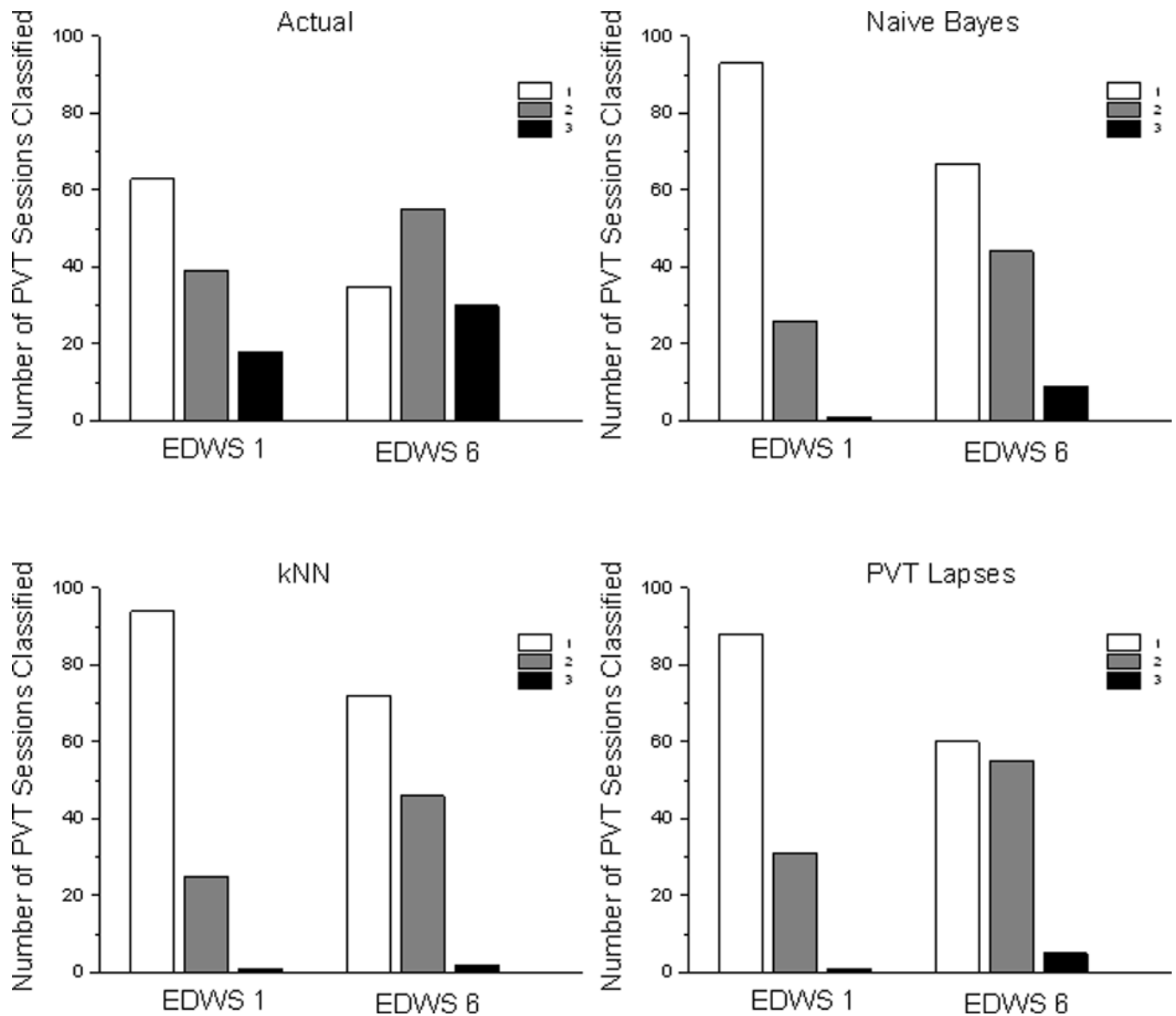


Figure 2. Classification of response by time within study for the two different methods

The number of test sessions classified as 1, 2 or 3 were compared between EDWS 1, the first extended duration work shift of interns on a Q3 schedule, and EDWS 6, the sixth extended duration work shift ~18 days later. Test sessions classified as 1 decreased from EDWS 1 to EDWS 6 and those classified as 2 or 3 increased across EDWS for both the kNN (lower left panel) and Naïve Bayes (upper right panel) methods as well as the classification based on PVT lapses (lower right panel). Actual classification, determined post-hoc based on relative mean slowest 10% RTs, is presented for comparison (upper left panel).

Table 1

Sensitivity, specificity, PPV and NPV results for the optimal kNN and Naïve Bayes algorithms found using the laboratory data. The confusion matrix shows the raw values of the actual vs. predicted classifications from the kNN and Naïve Bayes algorithms.

		kNN			Naïve Bayes		
		Actual			Actual		
		1	2	3	1	2	3
Predicted	1	139	30	2	147	34	2
	2	36	94	32	29	98	32
	3	4	32	137	3	24	137

Impairment	Sensitivity	Specificity	PPV	NPV	Sensitivity	Specificity	PPV	NPV
1	0.78	0.71	0.81	0.69	0.82	0.72	0.80	0.73
2	0.60	0.79	0.58	0.80	0.63	0.81	0.62	0.82
3	0.80	0.70	0.79	0.70	0.80	0.73	0.84	0.72

Table 2

Sensitivity, specificity, PPV and NPV results for the test sessions collected from the Q3 and IV medical intern schedules and the Phoenix Mars non-24-hr schedule.

Intern Q3 schedule										
<i>k</i> NN					<i>Naïve Bayes</i>					
Impairment	Sensitivity	Specificity	PPV	NPV	Sensitivity	Specificity	PPV	NPV	Sensitivity	Specificity
1	0.94	0.24	0.50	0.46	0.96	0.27	0.52	0.49		
2	0.33	0.61	0.45	0.50	0.33	0.65	0.45	0.54		
3	0.06	0.60	0.59	0.48	0.14	0.61	0.76	0.50		

Intern IV schedule										
<i>k</i> NN					<i>Naïve Bayes</i>					
Impairment	Sensitivity	Specificity	PPV	NPV	Sensitivity	Specificity	PPV	NPV	Sensitivity	Specificity
1	0.98	0.23	0.48	0.60	0.97	0.24	0.49	0.58		
2	0.24	0.77	0.58	0.49	0.27	0.75	0.56	0.50		
3	0.19	0.56	0.71	0.50	0.13	0.57	0.79	0.51		

Phoenix Mars non-24h-schedule										
<i>k</i> NN					<i>Naïve Bayes</i>					
Impairment	Sensitivity	Specificity	PPV	NPV	Sensitivity	Specificity	PPV	NPV	Sensitivity	Specificity
1	0.80	0.23	0.23	0.44	0.87	0.15	0.23	0.34		
2	0.34	0.31	0.41	0.28	0.14	0.38	0.28	0.27		
3	0.10	0.46	0.68	0.30	0.17	0.33	0.43	0.24		

Table 3

Confusion matrix results for the intern Q3 and IV schedules and Phoenix Mars non-24h-schedule from the kNN and Naïve Bayes algorithms using optimal feature spaces derived from laboratory data.

Intern Q3 schedule																
					<i>kNN</i>					<i>Naïve Bayes</i>						
					Actual					Actual						
					1	2	3	1	2	3	1	2	3	1	2	3
Predicted	1	309	267	46	1	315	263	23								
	2	19	134	143	2	14	137	151								
	3	1	8	13	3	0	9	28								

Intern IV schedule																
					<i>kNN</i>					<i>Naïve Bayes</i>						
					Actual					Actual						
					1	2	3	1	2	3	1	2	3	1	2	3
Predicted	1	312	318	22	1	311	309	16								
	2	6	105	71	2	8	118	84								
	3	1	8	22	3	0	4	15								

Phoenix Mars non-24h-schedule																
					<i>kNN</i>					<i>Naïve Bayes</i>						
					Actual					Actual						
					1	2	3	1	2	3	1	2	3	1	2	3
Predicted	1	163	349	183	1	177	400	191								
	2	40	193	240	2	7	80	202								
	3	0	23	48	3	19	85	78								

Table 4

Confusion matrix results for the intern Q3 and IV schedules and Phoenix Mars non-24h-schedule for classification based on number of PVT lapses.

		Intern Q3			
		Actual			
		1	2	3	
Predicted	1	309	232	7	
	2	20	167	180	
	3	0	10	15	

		Intern IV			
		Actual			
		1	2	3	
Predicted	1	308	261	3	
	2	11	161	97	
	3	0	9	15	

		Phoenix Mars non-24h-schedule			
		Actual			
		1	2	3	
Predicted	1	201	469	116	
	2	2	96	353	
	3	0	0	2	

Table 5

Confusion matrix results for the intern Q3 and IV schedules and Phoenix Mars non-24h-schedule from the kNN and Naïve Bayes algorithms using optimal feature spaces derived from the field-collected data. The optimal feature space used to derive each confusion matrix for each method is included for reference.

Intern Q3 schedule										
		kNN: lapses, mean fastest 10% RT			Naïve Bayes: lapses, age					
		Actual			Actual					
		1	2	3	1	2	3	1	2	3
Predicted	1	287	235	11	1	310	233	7		
	2	41	146	135	2	19	158	153		
	3	1	28	56	3	0	18	42		

Intern IV schedule										
		kNN: median RT, lapses, TOD gender			Naïve Bayes: lapses					
		Actual			Actual					
		1	2	3	1	2	3	1	2	3
Predicted	1	273	234	18	1	308	261	3		
	2	46	161	61	2	10	146	74		
	3	0	36	36	3	1	24	38		

Phoenix Mars non-24h-schedule										
		kNN: median RT, KSS			Naïve Bayes: age, gender					
		Actual			Actual					
		1	2	3	1	2	3	1	2	3
Predicted	1	144	295	183	1	118	215	102		
	2	56	203	142	2	57	213	174		
	3	3	67	146	3	28	137	195		