Accident Analysis and Prevention 126 (2019) 105–114

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Prediction of drowsiness events in night shift workers during morning driving

Yulan Liang[a,*], William J. Horrey[a,1], Mark E. Howard[b,e], Michael L. Lee[c,d,2], Clare Anderson[c,d,e], Michael S. Shreeve[a], Conor S. O'Brien[d], Charles A. Czeisler[c,d]

[a] Liberty Mutual Research Institute for Safety, 71 Frankland Rd., Hopkinton, MA 01748, USA
[b] Department of Respiratory & Sleep Medicine, Institute for Breathing & Sleep, Austin Health, Heidelberg, VIC 3084, Australia
[c] Sleep Health Institute and Division of Sleep and Medicine, Harvard Medical School, 164 Longwood Ave., Room 106, Boston, MA 02115, USA
[d] Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, 75 Francis St., Boston, MA 02115, USA
[e] Monash Institute of Cognitive and Clinical Neuroscience, School of Psychological Sciences, 18 Innovation Walk, Clayton Campus,Wellington Rd., Monash University, Victoria, 3800, Australia

ARTICLE INFO

Keywords:
Drowsy driving
Fatigue
Predictive models
Electroencephalogram (EEG)
Driving performance
Infrared oculograph

ABSTRACT

The morning commute home is an especially vulnerable time for workers engaged in night shift work due to the heightened risk of experiencing drowsy driving. One strategy to manage this risk is to monitor the driver's state in real time using an in vehicle monitoring system and to alert drivers when they are becoming sleepy. The primary objective of this study is to build and evaluate predictive models for drowsiness events occurring in morning drives using a variety of physiological and performance data gathered under a real driving scenario. We used data collected from 16 night shift workers who drove an instrumented vehicle for approximately two hours on a test track on two occasions: after a night shift and after a night of rest. Drowsiness was defined by two outcome events: performance degradation (Lane-Crossing models) and electroencephalogram (EEG) characterized sleep episodes (Microsleep Models). For each outcome, we assessed the accuracy of sets of predictors, including or not including a driver factor, eyelid measures, and driving performance measures. We also compared the predictions using different time intervals relative to the events (e.g., 1-min prior to the event through 10-min prior). By examining the Area Under the receiver operating characteristic Curve (AUC), accuracy, sensitivity, and specificity of the predictive models, the results showed that the inclusion of an individual driver factor improved AUC and prediction accuracy for both outcomes. Eyelid measures improved the prediction for the Lane-Crossing models, but not for Microsleep models. Prediction performance was not changed by adding driving performance predictors or by increasing the time to the event for either outcome. The best models for both measures of drowsiness were those considering driver individual differences and eyelid measures, suggesting that these indicators should be strongly considered when predicting drowsiness events. The results of this paper can benefit the development of real-time drowsiness detection and help to manage drowsiness to avoid related motor-vehicle crashes and loss.

1. Introduction

Drowsy driving is a significant, but preventable public health hazard. According to Harding and Feldman (2008), 20% of all serious car crashes in the general population are associated with driver drowsiness, independent of alcohol effects. Driver drowsiness is often regarded as one of the greatest identifiable and preventable causes of accidents in all modes of transportation, remaining one of the leading causes of fatal-to-driver heavy truck crashes (Ayas et al., 2006) and a primary risk factor for all motor vehicle crashes (Connor et al., 2002; Douglas, 2001; Fuchs et al., 2001).

Moreover, the effects of drowsiness are particularly detrimental for those who work non-standard hours compared to the average driving population (Barger et al., 2005). In the US, nearly 15% of workers, accounting for over 15 million people, work the night shift or rotating shift schedules (McMenamin, 2007). These workers are required to work during the biological night, when performance decrements are exacerbated, and to sleep during the biological day, when sleep is often

* Corresponding author. Present address: 157 Berkeley Street, T04C, Boston, MA 02116, USA
E-mail address: Liang.yulan@libertymutual.com (Y. Liang).
[1] The author's present address: 607 14th Street, NW, STE 201, Washington, DC 20005, USA.
[2] The author's present address: 1417 N.E. 42nd St Box 354875, Seattle, WA 98105-6246, USA.

http://dx.doi.org/10.1016/j.aap.2017.11.004
Received 15 May 2017; Received in revised form 2 November 2017; Accepted 2 November 2017
Available online 07 November 2017
0001-4575/

short and of poor quality (Gold et al., 1992; Kunert et al., 2007). The commute home following a night shift, which often starts between 6:00 to 7:00 am, is an especially vulnerable time for these people as the strong circadian drive for sleep interacts with the additional sleepiness resulting from extended wakefulness and physical/mental labor at work (Gold et al., 1992; Horne and Reyner, 1999; Lee et al., 2009; Silva et al., 2010; Stutts et al., 2003). A recent study demonstrated significant performance degradation in steering wheel control and lane maintenance during prolonged morning drives after night-shift compared to a post-sleep condition, with 37% of drivers having a near-crash event (Lee et al., 2016).

To manage the risk associated with drowsy driving, one promising strategy is to monitor driver state in real time and to intervene when, or even before, sleep is experienced or imminent. For example, onboard systems could alert drivers after identifying that the drivers are sleepy and recommend interventions, such as stopping driving and taking a nap or drinking a cup of coffee (Clark, 2008). To ensure the timeliness and efficacy of such a system, correctly recognizing driver drowsiness in real time is essential. A growing number of research studies have attempted to address this goal (Jackson et al., 2016; Khushaba et al., 2011; Sahayadhas et al., 2012).

In previous studies, the indicators of drowsy driving were often selected according to how drowsiness affects the human body and/or performance. Development of drowsiness is a gradual process and involves a series of physiological and behavioral changes, which can serve as predictive cues for drowsiness. Drowsiness occurs as people progressively transition from wakefulness into stage I sleep (Lal and Craig, 2001), when rhythmic activities of the brain transition from higher frequency beta and gamma waves (13–30 Hz and 30–70 Hz) to lower frequency alpha waves (8–13 Hz) and then to theta waves (4–7 Hz). Electroencephalogram (EEG) can capture this change and is often considered as the gold standard indicator of drowsiness (Lal and Craig, 2001). Slow rolling eye movements captured by electro-oculography (EOG)(Santamaria and Chiappa, 1987) and heart rate (ECG) and its variation (HRV) are also indicative of drowsiness (Harris and Mackie, 1972; Hartley et al., 1994). As drowsiness advances, it can be measured by a series of behavioral and performance measures, such as prolonged eye closure, head nodding, and performance errors. Eyelid movement measures—the percentage of eyelid closure (PERCLOS) over the pupil (Dingus and Grace, 1998) and blink frequency (Schleicher et al., 2008), facial expression and head movements (Belyavin and Wright, 1987), and performance degradation (McDonald et al., 2014; Yabuta et al., 1985) have been used as drowsiness indicators.

Nonetheless, in spite of these efforts, it is still difficult to choose satisfactory cutoff values for these indicators to separate drowsy driving from normal driving. Most of the detection models have been created using data-driven approaches; that is, model parameters are trained from data using machine learning techniques, ranging from simple linear models (Jackson et al., 2016) to complex, nonlinear methods, such as Fuzzy Analysis (Khushaba et al., 2011), Artificial Neural Networks (Daza et al., 2011), and Support Vector Machines (Li and Chung, 2013). The training data used in these approaches often contain a large number of samples, each of which includes the values of selected predictors as well as the known state of the driver (i.e., drowsy or not). The assumption in such cases is that training data can represent the majority of situations to which the models are generalized. The resultant models are validated with “new” data.

Importantly, the majority of models developed in previous studies are based on data collected in simulator-based experiments (i.e., driving simulators); few studies use data collected from real vehicles. However, the utility of data-driven models relies highly on the source of the data because the prediction reflects the relationship between variables in the training dataset. Because simulated driving lacks the perception of the risk encountered in real driving, there could be bias in the detection models. That is, drivers’ behavior in a simulated environment may deviate from normal, real-world driving (Philip et al., 2005). In this study, we sought to close this gap by developing and evaluating detection models of drowsy driving using performance and physiological data gathered from an actual vehicle (Lee et al., 2016). Additionally, we sought to extend the application of predictive models of drowsy driving to a population of non-standard (night-shift) workers.

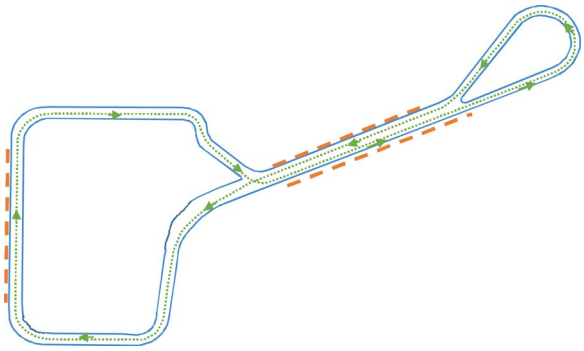


Fig. 1. The test track had three straight sections. The dotted line with arrows indicates the route of driving, and bolded dashed lines indicate the straight sections of the track.

2. Methods

2.1. Experiment and data source

We used data collected in a study that examined the impact of night shift work on driver performance in an instrumented vehicle on a closed test track with three straight sections (Lee et al., 2016) (ref. to Fig. 1). Sixteen night shift workers (mean age: 48.7; range 19–65; 7 male and 9 female) undertook two 2-h driving sessions: one following a night of rest (> 5 h sleep, no work, Post-Sleep condition), and another following a night of shift work (> 5 h work, no intervening sleep, Post-Shift condition). Under the Post-Sleep condition, participants reported that they slept 7.6 ± 2.41 h (range: 5–12 h) during the night before, did not work a night shift for 30 continuous hours prior to the session and had at least 2 sleep episodes during that interval. Under the Post-Shift condition, participants worked continuously for 8.3 ± 4.1 h (range: 5–20.5 h) within the interval of 10:00PM-8:00 AM prior to the session, with no intervening sleep between the end of the night shift and the start of the session. Sleep and wake times, medication use, and caffeine intake were recorded via diary and wrist activity monitor for the week prior to each session (Actiwatch-L; Mini Mitter, Bend, OR). Two out of 16 participants were classified as morning chronotype, 11 as intermediate chronotype, and 2 as evening chronotype. Five out of 16 were identified as high risk of sleep apnea. Participants did not consume caffeine for at least two hours before the drive. All driving sessions began between 9:30 AM and 2:30 p.m. Each participant’s post-sleep and post-shift drives were time-matched to control for potential time-of-day effects (1 ± 1.2 h; range: 6 mins–5 h). Although the lighting condition of environment was not particularly controlled, all driving sessions occurred under well-lit weather conditions for the ease of experiment conduction and consistent road surface condition.

During the driving sessions, participants drove on a two-lane 0.8 km closed-loop test track, consisting of a variety of curved and straight sections; a guard rail surrounded the majority of the track. Participants were asked to stop the vehicle approximately every 15 min in order to complete surveys related to sleepiness symptoms and their driving performance and completed up to seven consecutive 15-min driving blocks.

The vehicle was equipped with a passenger-side brake pedal so that the investigator could intervene when safety was concerned. Seven sessions were terminated early—most often due to safety considerations (e.g., investigator had to intervene multiple times; participants could no longer maintain adequate control of the vehicle or expressed safety concerns) and all occurred under the Post-Shift condition (Lee et al.,

Table 1
The objective measures of driver drowsiness including model predictors and prediction targets.

Measures		Description	Apparatus
Prediction Targets	Microsleep	Scored as 1 when at least one > 3-s microsleep episode (identified by theta wave intrusion in EEG data) occurred; otherwise, scored as 0	A portable Vitaport 4 system (Temec™) was used to measure EEG. Gold cup electrodes were fixed to the participant's scalp in the central, frontal, parietal, and occipital positions with adhesive electrode gel following scalp abrasion.
	Lane Crossing	Scored as 1 when the vehicle crossed the lane in any of the straight sections of the track (Fig. 1); otherwise, scored as 0	An in-vehicle investigator manually time-stamped whenever the vehicle exceeded the lane.
Predictors	SD of Lane Position	Average of standard deviation (SD) of vehicle lane position on the straight sections of the track	The instrumented vehicle equipped with several sensors and computers recorded vehicle lane position and driver steering input.
	SD of Steering Wheel Position	Average of SD of steering wheel position on the straight sections of the track	
	Mean Steering Wheel Error	Average of the difference between the actual steering wheel position and the steering wheel position predicted by a second-order Taylor expansion (Nakayama et al., 1999). It measured the smoothness of steering wheel movements on the straight sections of the track	
	Positive Amplitude/ Velocity Ratio (AVR)	Average and SD of the ratio of the maximum amplitude to maximum velocity of eyelid movements during closing phase	The Optalert™ system recorded the frequency and velocity of eye blinks. This device uses a small infra-red emitter (IR-A band: 760–1400 nm), and recorder mounted on an open lens glasses frame. Whenever necessary, the Optalert™ glasses were fitted with corrective lenses to accommodate a current prescription.
	Negative AVR	Average and SD AVR during reopening phase	
	% Eye Closure (PERLOC)	Percentage of time per minute when the eyes were closed	
	John's drowsiness score (JDS)	Composite score (1–10) of oculography measures. > 5: severe drowsiness	

Note: For the prediction targets, one-minute intervals were expressed as binomial values (either 0 or 1); the predictors were also aggregated across the one-minute intervals.

2016). The drivers whose sessions were terminated early had an average of 1.6 brake interventions by investigators while other drivers had none.

Objective physiological and performance measures were recorded continuously throughout the driving sessions, including ocular measures, polysomnography (e.g., EEG and EOG), and driving performance. An in-vehicle investigator, blind to the condition, monitored vehicle lane position and marked whenever the vehicle crossed the lane (i.e., if any wheel came into contact with the lane marking) or when the investigator had to apply an emergency brake to keep the vehicle on a safe course because a driver failed to respond to loss of vehicle control. There were two investigators who performed the recording across the sessions, and the investigators reached commonality on recording criteria for the events prior to the experiment. The same investigator recorded for both Post-Sleep and Post-Shift sessions of one driver. A subset of the objective measures, defined in Table 1, were used in the subsequent analyses and summarized across one-minute, non-overlapping intervals.

2.2. Predictive models

We built two types of predictive models that varied by prediction target: a Microsleep model and a Lane-Crossing model. For Microsleep models, drowsiness was defined as at least one > 3-s microsleep episode occurring in a one-minute interval (Table 1). Microsleep episodes were identified, through visual inspection, by theta wave intrusions lasting longer than 3 s. The inspection of data was carried out by a single expert scorer who was blinded to the experimental condition. A Microsleep indicates a severe level of drowsiness and/or the onset of sleep (Lal and Craig, 2001). For Lane-Crossing models, the drowsiness state was defined by performance degradation and operationalized as lane-crossing events that occurred on any of the longer straight sections of the track during a given one-minute interval (Table 1 and Fig. 1). Lane-crossing events occurring on the straight sections of the track were highly likely related to driver drowsiness because drivers were instructed to drive in the center of the lane and no in-vehicle or out-vehicle distraction might cause the lane-crossing events. We excluded curves and corners from the models, given drivers' tendency to regularly cut corners under both conditions, suggesting this was strategic. The lane-crossing events represented the performance breakdowns caused by drowsiness of the driver and could lead to severe safety outcomes. The one-minute intervals when either microsleep or lane-crossing events occurred were defined as drowsy epochs. In total, there were 60 drowsy epochs used to build Microsleep models (occurring on eight drivers, ranging from one to 12 Microsleep events for a driver) and 112 drowsy epochs used to build Lane-Crossing models (occurring on 10 drivers, ranging from one to 26 lane-crossing events for a driver).

Paired baseline (non-drowsy) epochs were sampled for each type of model separately. For each drowsy epoch, these one-minute baseline epochs were taken from the same driver under the post-sleep condition and occurred before the earliest observed drowsy epoch. For example, if Driver A had experienced a drowsy epoch at the 60-min mark of their post-shift drive, baseline epochs would be drawn only from the 59th minute or earlier from Driver A's post-sleep drive. The analysis included a total of 259 baseline epochs corresponding to 60 Microsleep drowsy epochs and a total of 702 baseline epochs corresponding to 112 Lane-Crossing drowsy epochs.

The predictors applied in both types of model were driving performance and/or ocular measures recorded using Optalert™ (Table 1). These two categories of predictors were chosen because they are relatively nonintrusive compared to both physiological measures (e.g., EEG, EOG, ECG) and subjective measures and, therefore, could be more suitable for a drowsy driving detection application used in real driving. The driving performance measures collected by the instrumented vehicle included standard deviation (SD) of lane position, SD of steering wheel position, and mean steering error when the vehicle was driven on the straight sections. Steering error measures the smoothness of steering wheel movements, calculated as the difference between the actual

steering wheel position and the steering wheel position predicted by a second order Taylor expansion (Nakayama et al., 1999). For Lane-crossing Model, these driving performance measures were calculated for only the sections of the track where the vehicle did not go across the lane (i.e., a target event).

The ocular measures were collected by monitoring eyelid movements via infrared reflectance oculography (Optalert™, Melbourne, Australia) that uses a small infra-red transducer mounted on an open-lens glasses frame that emits and detects infra-red light (IR-A band: 760–1400 nm). The measures included mean and SD of Amplitude/Velocity Ratio (AVR), PERCLOS, and Johns Drowsiness Score (Anderson et al., 2013; Ftouni et al., 2013; Wilkinson et al., 2013). Positive and negative AVRs represent how fast eyelids move as people close and reopen their eyes during blinks (Table 1). PERCLOS is sensitive to long blinks or eyelid closure caused by sleepiness and has been used as a benchmark drowsy indicator in many studies (Dingus and Grace, 1998). Johns Drowsiness Score (JDS) has been validated as a scale to measure fatigue (Johns et al., 2007) and is recorded by the Optalert™ device. All these measures were summarized across a one-minute interval.

To identify which ocular and performance measures could most accurately predict drivers' drowsiness state, we evaluated and compared different models that included the following variations: (a) whether or not the model considered the effect of an individual driver (individualized vs. general); (b) whether the model was improved by the inclusion of the set of driving performance measures (Driving vs. No-Driving), and c) how the inclusion of ocular measures impacted model performance (AVR + JDS, AVR + PERCLOS, PERCLOS, JDS, or No-Ocular). In total, we built 19 models with different sets of predictors, which are outlined in Table 2.

In a subsequent modeling exercise, we traced the performance of predictors backwards in time (e.g., N minutes before the event), in order to examine the transition from wakefulness to sleep. The time scale (t) used here ranged from zero to ten minutes before the event: zero (t_0) represented physiological and performance data originating from the same time interval as the event and ten (t_{-10}) representing physiological and performance data gathered ten minutes prior to the event (aggregated at a one minute interval).

2.3. Model evaluation

In model evaluation, we split drowsy epochs and baseline epochs randomly into 75% for training and 25% for testing. Logistic regression models were built with the training data and tested with the "unseen" epochs in the testing set. The models were evaluated using testing sensitivity, specificity, accuracy and the area under receiver operating characteristic (ROC) curve (AUC). All these measures range from zero to one, and the higher the value, the better the model predicts. Sensitivity is true positive rate (true positive/total positive), specificity is true negative rate (true negative/total negative), and accuracy is the ratio of the sum of true positive and true negative to total cases. ROC curve uses true positive rate (sensitivity) against false positive rate (1-specificity); the area under ROC illustrates the predictive ability of a binary-classification algorithm as its determining threshold varies.

To estimate variance of these measures for model evaluation, we used a bootstrapping resampling method. For each model, we re-sampled the data 1000 times and used 90% confidence interval (CI) of quartiles (from 5% to 95% quartiles) as significant criterion to compare the performance of the predictive models. The statistical analyses associated with resampling and building logistic regression models were conducted with SAS 9.4, and 90% CIs were calculated with R 3.3.2.

3. Results

First, the models at t_0 were compared to identify which sets of predictors were the most indicative to drowsy driving among the 19 models. We examined the models with and without the effect of drivers (variation a); and then among the winning models (individualized or general models), we identified the best ones among those with difference driving performance and ocular predictors (variations b and c). Then, we used the best model to evaluate different time scales (from t_0 to t_{-10}). These analyses were performed for both the Microsleep and Lane-Crossing models.

3.1. Driver effect

Across all Lane-Crossing models at t_0, AUC was 0.82 (SD: 0.08), accuracy was 88% (SD: 3%), sensitivity was 0.36 (SD: 0.22), and specificity was 0.98 (SD: 0.02). For the Microsleep models at t_0, AUC was 0.82 (SD: 0.10), accuracy was 89% (SD: 4%), sensitivity was 0.39 (SD: 0.24), and specificity was 0.98 (SD: 0.02).

We calculated the paired difference for each of the four evaluation measures between individualized models and general models. While an individualized model included driver ID as a predictor, its paired general model did not; however, the same set of performance and ocular predictors were trained and tested by the same set of resampled data. For Lane-Crossing models, the individualized models yielded improvements in AUC, accuracy, and sensitivity from the general models (shown in the left panel in Fig. 2 and Table 3), showing the advantages of considering individual characteristics of drivers.

Similar results were found for the Microsleep model (shown in the right panel in Fig. 2 and Table 2); the individualized models yielded higher AUC and sensitivity than the general models. However, unlike for the Lane-Crossing model, the individualized models for Microsleep did not show significant improvements for accuracy.

3.2. Driving performance and ocular measures

We compared different combinations of predictors using the individualized models. For Lane-Crossing models (Fig. 3 and Table 4), the results of AUC, accuracy, and sensitivity showed that there were no significant improvements in the models by the inclusion of driving performance measures as predictors compared to models without these variables (sdoj vs s-oj, sdop vs s-op, sdj vs s-j, sdp vs s-p). In contrast, the models with ocular measures performed better compared to models without (i.e., sdoj, sdop, sdj, sdp, s-oj, s-op, s-j, s-p versus sd-, s–). Comparing the models with different sets of ocular measures, there was a consistent, but not statistically significant, decline of model performance in the following order: the models with AVRs and JDS, the models with AVRs and PERCLOS, the model with JDS, the model with

Table 2
Model symbols by three types of variations.

		AVR_JDS	AVR_PERCLOS	JDS	PERCLOS	No-Ocular
Individualized models	Driving	sdoj	sdop	sdj	sdp	sd-
	No-Driving	s-oj	s-op	s-j	s-p	s–
General models	Driving	-doj	-dop	-dj	-dp	-d-
	No-Driving	–oj	–op	–j	–p	

s: driver effect, d: driving measures, o: AVRs measures, j: JDS, and p: PERCLOS.

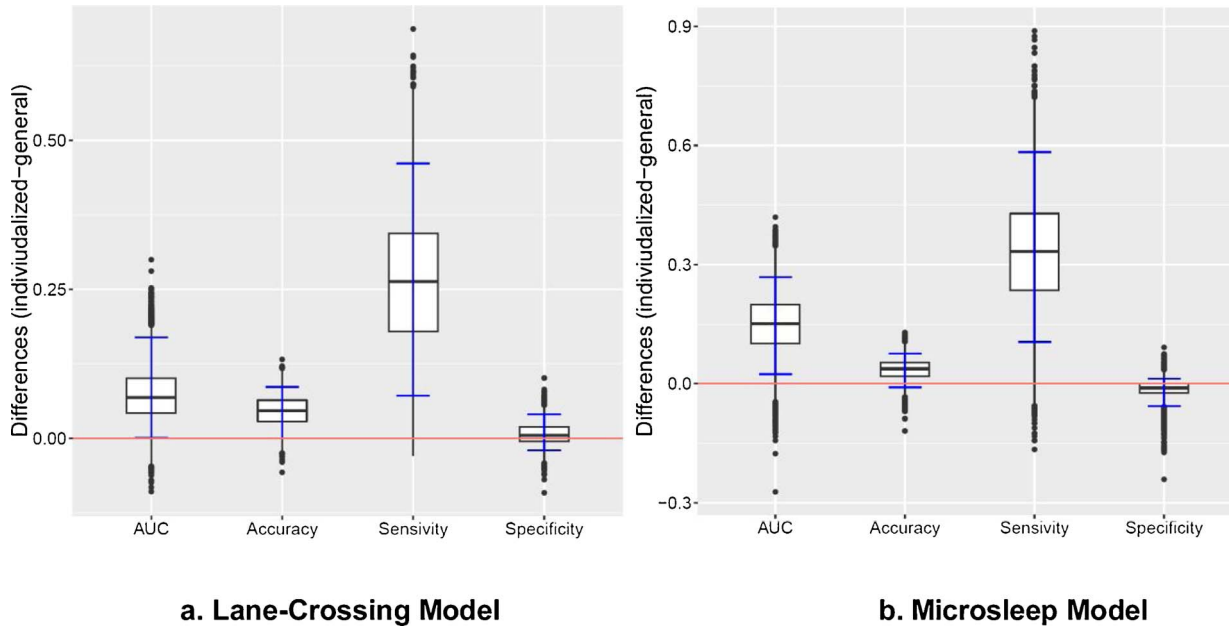


Fig. 2. Comparison between individualized models and general models. The error bars indicate 90% confident interval of the paired differences.

Table 3
Performance difference between individualized models and general models.

	Lane-Crossing Model	Microsleep Model
AUC	0.07 [0.00, 0.17]*	0.15 [0.02, 0.27]*
Accuracy	0.05 [0.00, 0.09]*	0.04 [−0.01, 0.08]
Sensitivity	0.26 [0.07, 0.46]*	0.33 [0.11, 0.58]*
Specificity	0.01 [−0.02, 0.04]	−0.01 [-0.06, 0.01]

Median [lower bound, upper bound] of 90% CI. AUC: area under curve.
* Significant difference between individualized models and general models using 90% CI as criterion.

PERCLOS, and finally the model without any ocular measures (i.e., sdoj- > sdop- > sdj- > sdp- > sd- and s-oj- > s-op- > s-j- > s-p- > s–). Specificity did not show any difference across the models. Overall, the highest performing models were those that included AVRs and JDS (sdoj and s-oj) as predictors.

The performance of all Microsleep models was similar with respect to AUC, accuracy, and sensitivity. However, there were differences with respect to sensitivity (see Table 4). Similar to the Lane-Crossing models, there were patterns of non-significant declines in AUC, accuracy, and sensitivity among the models with different combinations of ocular measures (i.e., sdoj- > sdop- > sdj- > sdp- > sd- and s-oj- > s-op- > s-j- > s-p- > s–). Again, there was no difference in specificity across these models. Based on the median of evaluation measures, the highest performing models were those that included AVRs and JDS or PERCLOS (sdoj sdop, s-oj and s-op) as predictors.

3.3. Time factor

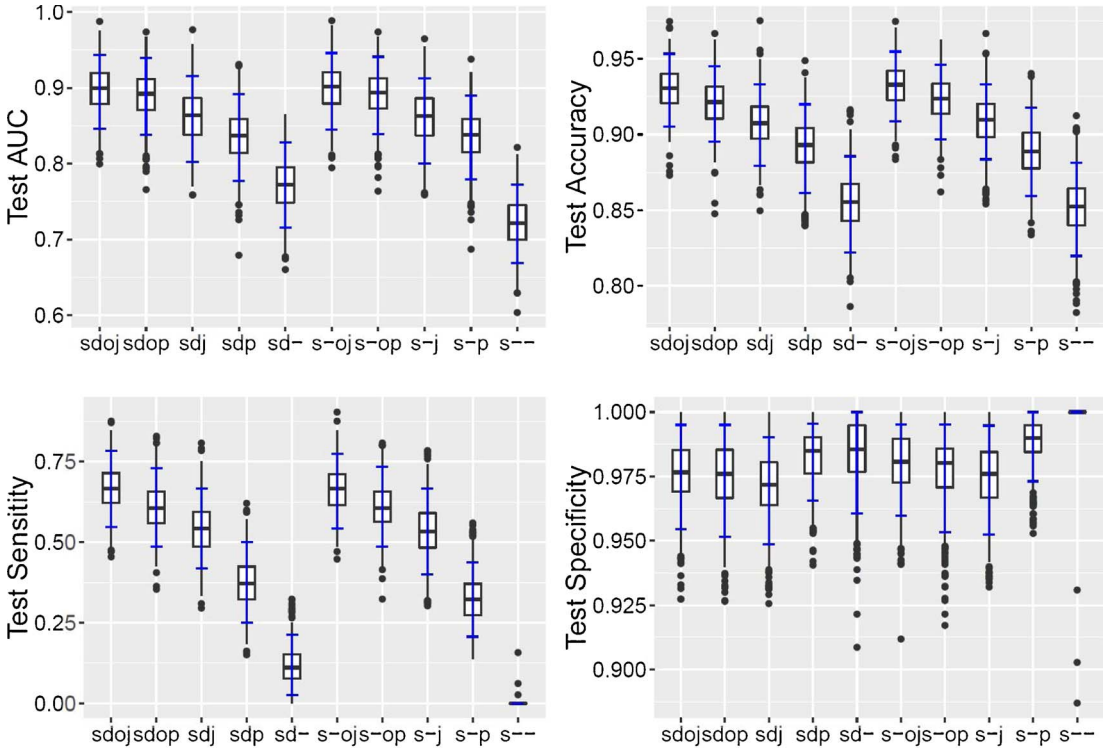
We used the model with driver ID, driving performance, AVRs and JDS (sdoj) as predictors to examine how the time factor (from t_0 to t_{-10}) affected the model performance. Although there were nominal differences and variations across the different time scales, for both Lane-Crossing and Microsleep models, there were no significant differences in model performance (Fig. 4).

4. Discussion

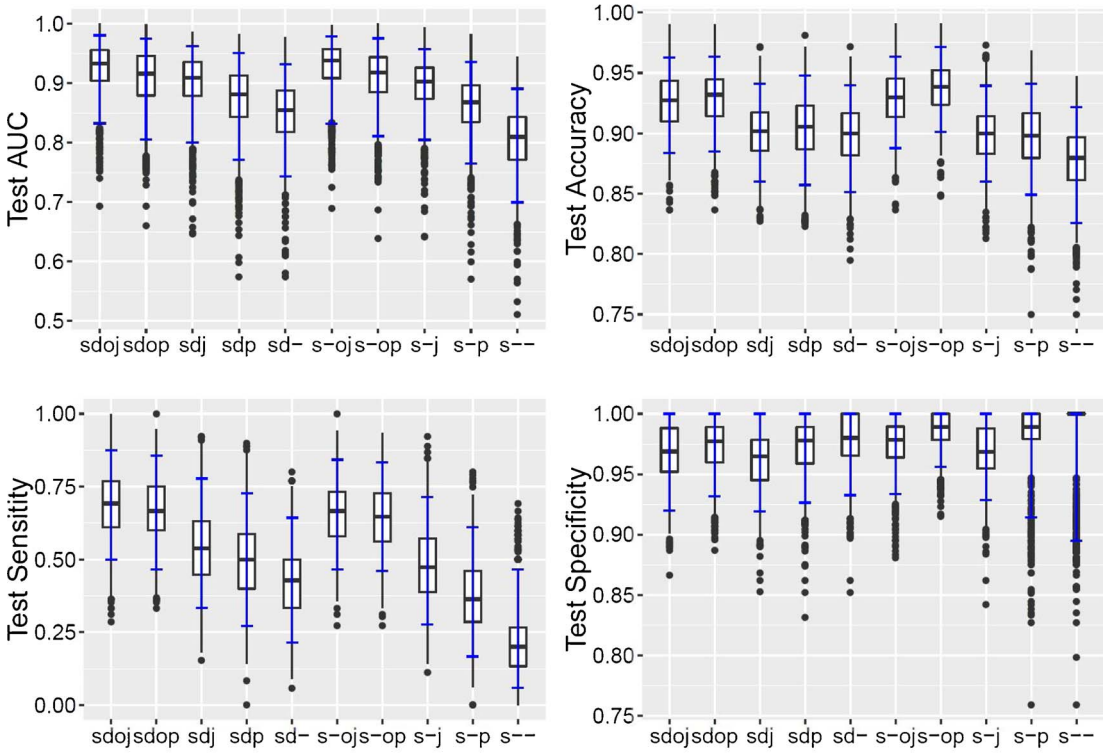
This study found that drivers' physiological data was predictive of driver drowsiness (Lane-Crossing and Microsleep events) during real vehicle driving in a population of at-risk drivers (i.e., night shift workers). Using data drawn from Lee et al. (2016), we carried out a series of modeling exercises to explore the role and influence of various predictors including driver individual differences, ocular measures, and driving performance. The results showed that including driver ID as a predictor improved AUC and prediction accuracy for both types of models. The Lane-Crossing models with different sets of ocular measures as predictors improved prediction accuracy while the same predictors in the Microsleep models did not produce any statistically significant improvements. Prediction performance was not changed by adding driving performance predictors for either model; an examination of different time to event epochs (i.e., t_0–t_{-10}) did not reveal any significant differences. Overall, the best performing models for both Lane Crossing models and Microsleep models were those considering driver individual differences and ocular measures.

This study applied a data-driven approach to build predictive models for drowsy driving; the data used to train predictive models were collected in an experiment where drivers drove an instrumented vehicle on a test track (Lee et al., 2016). Although driving on the testing tack, and knowing they were participating an experiment could affect driver behavior and drowsiness, when compared with many studies that used data collected in a driving simulator (Jackson et al., 2016; Li and Chung, 2015), behavioral patterns captured from the data collected in real driving situations could be closer to those posited in naturalistic driving on the road, implying that these models could be more likely generalized to the real world.

The overall performance of these models was promising (Lane-Crossing model: AUC 0.82 and accuracy 88%; Microsleep model: AUC 0.82 and accuracy 90%). Nonetheless, the models had high specificity (true negative rate) and moderate sensitivity (true positive rate), which means that the prediction made by the models had low false alarms, but were likely to miss drowsy cases. This could be driven by the unbalanced samples between drowsy and non-drowsy epochs. That is, samples were more likely to be predicted as a non-drowsy state because logistic regression models optimize the overall accuracy, which leads the models to favor the class with the majority of training samples. This represents a limitation of the training data because drowsy events (Microsleep and Lane-crossing events) were relatively rare, although representing the true state of real driving. One potential remedy for this is to use other classification methods which are more tolerant to unbalanced data (e.g., Support Vector Machines)—a possibility that can



a. Lane-Crossing model



b. Microsleep model

Fig. 3. Comparison of Lane-Crossing (a) and Microsleep (b) Models with different performance predictors. Model symbols refer to Table 2 (s: driver, d: driving measures, o: AVRs measures, j: JDS, and p: PERCLOS). AUC: area under curve. In the boxplots, the middle lines indicate the median and the error bars indicate 90% of confidence intervals.

Table 4
Performance and comparison of the models with different combinations of performance predictors.

		AUC		Accuracy		Sensitivity		Specificity
Lane-Crossing Model	sdoj	0.90 [0.85, 0.94]	A	0.93 [0.91, 0.95]	A	0.67 [0.55, 0.78]	A	0.98 [0.95, 1.00]
	sdop	0.89 [0.84, 0.94]	A	0.92 [0.90, 0.95]	A	0.61 [0.49, 0.73]	AB	0.98 [0.95, 1.00]
	sdj	0.86 [0.80, 0.92]	AB	0.91 [0.88, 0.93]	ABC	0.54 [0.42, 0.67]	ABC	0.97 [0.95, 0.99]
	sdp	0.84 [0.78, 0.89]	AB	0.89 [0.86, 0.92]	ABC	0.37 [0.25, 0.50]	ABC	0.98 [0.97, 1.00]
	sd-	0.77 [0.72, 0.83]	AB	0.86 [0.82, 0.89]	ABC	0.11 [0.03, 0.21]	ABCD	0.99 [0.96, 1.00]
	s-oj	0.90 [0.84, 0.95]	A	0.93 [0.91, 0.95]	A	0.67 [0.54, 0.77]	A	0.98 [0.96, 1.00]
	s-op	0.89 [0.84, 0.94]	A	0.92 [0.90, 0.95]	A	0.61 [0.49, 0.73]	AB	0.98 [0.95, 1.00]
	s-j	0.86 [0.80, 0.91]	AB	0.91 [0.88, 0.93]	ABC	0.53 [0.40, 0.67]	ABC	0.98 [0.95, 0.99]
	s-p	0.84 [0.78, 0.89]	AB	0.89 [0.86, 0.92]	ABC	0.32 [0.21, 0.44]	ABCD	0.99 [0.97, 1.00]
	s–	0.72 [0.67, 0.77]	AB	0.85 [0.82, 0.88]	ABC	0.00 [0.00, 0.00]	ABCDE	1.00 [1.00, 1.00]
Microsleep Model	sdoj	0.93 [0.83, 0.98]		0.93 [0.88, 0.96]		0.69 [0.50, 0.88]	A	0.97 [0.92, 1.00]
	sdop	0.92 [0.81, 0.97]		0.93 [0.88, 0.96]		0.67 [0.47, 0.86]	AB	0.98 [0.93, 1.00]
	sdj	0.91 [0.80, 0.96]		0.90 [0.86, 0.94]		0.54 [0.33, 0.78]	AB	0.96 [0.92, 1.00]
	sdp	0.88 [0.77, 0.95]		0.91 [0.86, 0.95]		0.50 [0.27, 0.73]	AB	0.98 [0.93, 1.00]
	sd-	0.85 [0.74, 0.93]		0.90 [0.85, 0.94]		0.43 [0.21, 0.64]	AB	0.98 [0.93, 1.00]
	s-oj	0.94 [0.83, 0.98]		0.93 [0.89, 0.96]		0.67 [0.47, 0.84]	AB	0.98 [0.93, 1.00]
	s-op	0.92 [0.81, 0.98]		0.94 [0.90, 0.97]		0.65 [0.46, 0.83]	AB	0.99 [0.96, 1.00]
	s-j	0.90 [0.80, 0.96]		0.90 [0.86, 0.94]		0.47 [0.28, 0.71]	AB	0.97 [0.93, 1.00]
	s-p	0.87 [0.77, 0.94]		0.90 [0.85, 0.94]		0.36 [0.17, 0.61]	AB	0.99 [0.91, 1.00]
	s–	0.81 [0.70, 0.89]		0.88 [0.83, 0.92]		0.20 [0.06, 0.47]	ABC	1.00 [0.89, 1.00]

Median [lower bound, upper bound] of 90% CI; AUC: area under curve.
s: driver, d: driving measures, o: AVRs measures, j: JDS, and p: PERCLOS.
ABCDE: post-hoc comparisons using 90% CI as criterion.

be explored in future work.

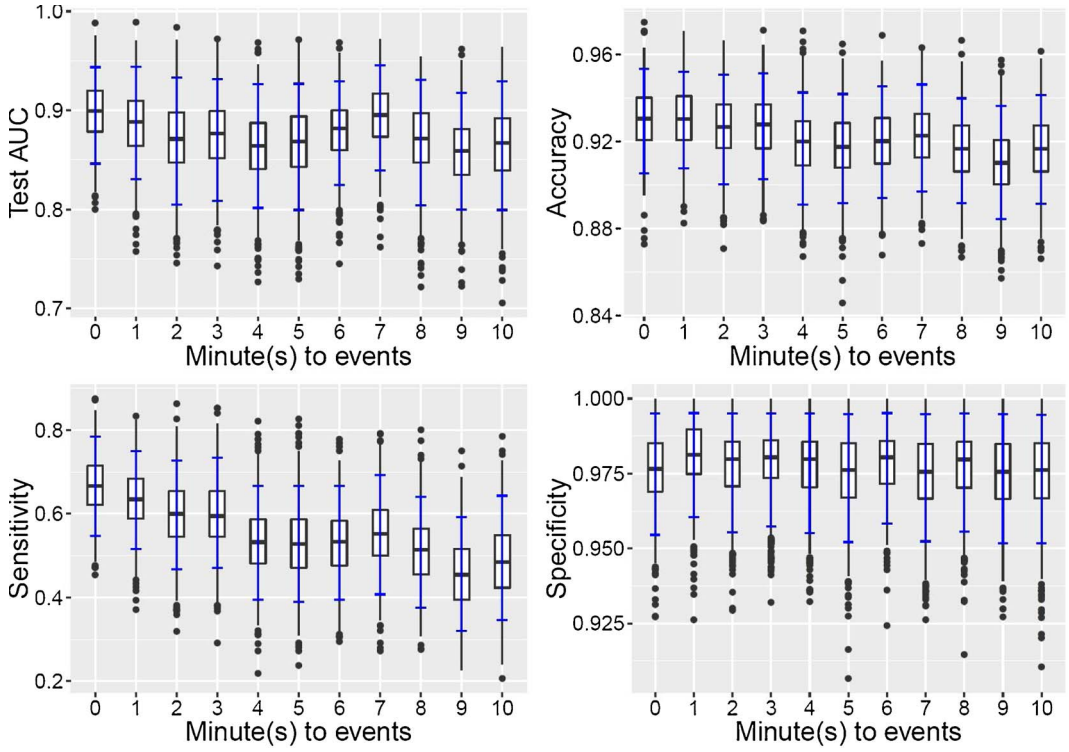
The results show that individual differences between drivers are an important consideration when building predictive models for drowsiness. This is not surprising, given the wide range of inter-individual differences respecting drowsiness etiology, chronic and acute sleep effects, chronotype, and many other factors (Roenneberg et al., 2003; Williamson et al., 2011). This result is also consistent with the observation found when building predictive models for driver cognitive distraction, where an individual model for each driver was a significant and necessary model component (Liang and Lee, 2014). In this study, the effects of driver individual differences were manifest as a series of parallel lines in the logistic regression models, meaning drivers shared common coefficients for other predictors (ocular and/or performance). It is expected that customized coefficients for other predictors by each driver (building the model for each driver separately) could further improve the overall predictions; however, this hypothesis would need to be tested with a much larger dataset than the current study can offer. Based on the predictive models, we are unable to locate the etiological causes of these differences, the predictive models could only quantify the magnitude of individual difference on drowsiness detection. To explore the etiological causes, researchers need to rely on a well-controlled laboratory experiment, which is out of the scope of the current study.

Although tailored models that can accommodate individual differences may be necessary, it might be possible to find a midway approach between a general model for all drivers and individualized models, one for each driver. That said, a group of drivers may share a common model based on their similar responses to drowsiness. Moreover, as drowsiness further advances to a high level, the cues of drowsiness may further intensify and become unified across different groups. In this way, the number of models may decrease as the different groups may merge. It merits more detailed research into how the effects of individual difference changes with drowsiness progress from wakefulness to sleep. On the other side, individual models also carry practical implications. The model can be stored on a specialized, personal device that the owner can carry from vehicle to vehicle or stored on the vehicle that the driver usually uses for morning/night/long commute/trips. This outlook is especially possible with the development of artificial intelligent (AI) technology.

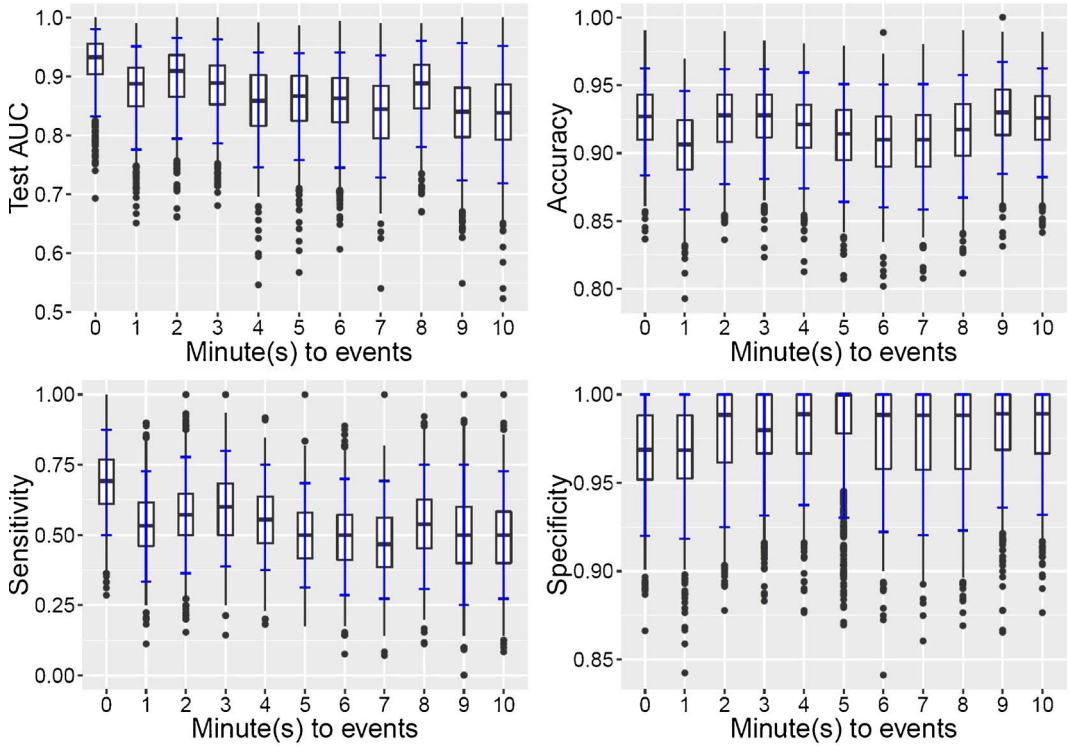
It is somewhat surprising that driving performance (SD of lane position, SD of steering angle, and mean of steering error) did not improve the prediction of drowsiness for either Microsleep models or Lane-Crossing models (McDonald et al., 2014; Yabuta et al., 1985). One possible explanation of this result is that we constrained the evaluation of performance to the straight sections of the track. Once the heading of the vehicle was aligned with the direction of the road, little effort was needed to adjust the steering wheel to maintain a straight heading. Thus, driving performance, as measured here, may be less sensitive to the effects of drowsiness than on more-demanding sections. This result may also indicate that the effects of severe drowsy episodes that led to microsleep and lane-crossing events may be transient in this dataset because the measures describing the performance before and/or after microsleep or lane-crossing events (the section where the events occurred was excluded when calculating these measures) were not predictive to the events; that is, drivers might experience these short severe drowsy episodes and quickly transition back to less drowsy levels.

The models with different ocular measures provided reasonable prediction for driver drowsiness. For the Lane-Crossing models, there was a significant difference between the models with ocular measures as predictors and those without. The prediction of these models worsened as the number of ocular predictors was reduced although the results were not significant. This result posits that eyelid activities are important indicators of drowsy driving, which is consistent with many previous studies on detection of drowsiness (Dingus and Grace, 1998; Johns et al., 2007). Also, the results did not show any effect of time factor (from t_0 to t_{-10}) on model performance. This implies that the predictive utility of the model holds up even moving backwards in time as it opens up new possibilities for different countermeasures (e.g., when one has a bigger time window to react). Thus, when building detection models for drowsy driving, model builders should consider the effects of individual differences and include ocular measures as predictors.

Overall, Microsleep models and Lane-Crossing models obtained similar performance in terms of testing AUC and accuracy. Lane-Crossing models were more sensitive to the changes in the number of ocular predictors than the Microsleep models. Microsleep events and lane-crossing events indicated driver drowsiness from physiological perspectives and performance degradation perspectives, respectively. A more refined and nuanced examination of the order of the two types of events may shed light on how drowsiness develops and manifests itself



a. Lane-Crossing Model



b. Microsleep Model

Fig. 4. Comparison of models that used predictors tracked backward along time from zero to ten minutes. In the boxplots, the middle lines indicate the median and the error bars indicate 90% of confidence intervals.

in human physiology and performance across time. Nonetheless, we did not find any consistent pattern in the order of microsleep and lane-crossing events across drivers. One possible explanation is that there is not an etiological sequence in human physiological responses and performance degradation caused by drowsiness. Another possible explanation is that, although the drive to sleep is strong, drivers try to keep themselves alert and maintain acceptable performance, which interrupts natural development of sleep and any possible inherent order of the two types of events. Also, we only evaluated two types of outcomes of drowsiness (i.e., lane crossing and microsleep); there may be a pattern preset if more outcomes (e.g., rapid eye rolling, nodding) were also examined. However, such efforts to fight drowsiness could lead to prolonged driving under driver impairment and cause safety mishaps. Seven out of 16 drivers in this study were terminated early because of safety reasons (e.g., near-crashes) due to drowsiness (Lee et al., 2016). This demonstrates that continuing to drive when drowsiness manifests is very dangerous, and a safe and effective intervention is to stop driving and take a nap (Rosekind et al., 1995).

There are several limitations of this study. First, the size of data used to train and test the detection models was relatively small and unbalanced between drowsiness and non-drowsiness classes. This limitation arises from the infrequency of microsleep and lane-crossing events observed in the experiment. However, using microsleep and lane-crossing events to define drowsiness better controls the severity of drowsiness that the models detected and carries more realistic practical implications compared to a definition that uses the coarse classification based on experimental conditions (Vicente et al., 2016). Furthermore, the unbalanced data may reflect natural occurrences of drowsy driving: for average drivers, drowsy driving is likely to be rare compared with normal driving. We acknowledge that a small training data set can miss some situations and reduce the generalization of the models. There are plans to collect data from more drivers for longer periods of time in future studies.

Second, we only applied a relatively simple classification model, linear logistic regression, to build detection models. While logistic regression, which minimizes training error, could be biased toward the non-drowsiness state, there are other classification methods with a higher tolerance for unbalanced data. Nonetheless, this study aimed to identify important indicators for drivers' drowsiness in terms of the physiological gold standard (microsleep) as well as performance degradation (lane crossing). The classification method served as a tool to implement this comparison although it may not yet be an ideal method for identifying/predicting drowsiness. In future studies, we will compare different classification methods to identify driver drowsiness.

Moreover, the predictors were summarized across one minute—a constraint imposed by the output rate of the Optalert™ device—which is a relatively long time period when considering that microsleep and lane-crossing events can arise within the span of a few seconds. In contrast, it is possible with the current dataset for evaluating driving performance measures at a finer time scale. An alternative approach for the ocular measures is to use a high-resolution eye tracking system (e.g., 120 Hz) to monitor eyelid movements. During the driving sessions, the participant was asked to stop the vehicle approximately every 15-min in order to complete the various scales related to sleepiness symptoms and their driving performance. This may have inadvertently alerted the driver during these breaks and may have decreased the impact of drowsiness on these scales and on the subsequent driving session. In future studies, we will adopt longer time intervals between scale rating (e.g., 30 min) and/or also consider administrating the rating while drivers continue driving. In fact, the presence of investigators in the vehicle alongside the driver on the test track might also affect driver behavior, potentially rousing drivers from drowsiness or diminishing the motivation to stay alert. But this study design lay in safety consideration of participants and investigators and, thus, the generalization of the predictive models built based on the data collected in such a circumstance needs to be further evaluated in a naturalistic driving environment.

In spite of the study's limitations, the current study developed and evaluated predictive models for driver drowsiness using performance and physiological data gathered from an actual vehicle and from a population of at-risk drivers (i.e., night shift workers). The best models for both measures of drowsiness were those considering driver individual differences and eyelid measures, suggesting that these indicators should be strongly considered when predicting drowsiness events. These results can benefit the development of real-time drowsiness detection and help to manage drowsiness to avoid related motor-vehicle crashes and loss.

Acknowledgements

This study was supported by a grant from the Institute of Breathing and Sleep Research (to M.E.H.); by Liberty Mutual Insurance; National Institutes of Health Award 5T32HL7901-14 (to M.L.L.); National Space Biomedical Research Institute Award PF03002 (to M.L.L.); Department of Homeland Security Federal Emergency Management Agency Assistance to Firefighter Grant EMW-2010-FP-00521 (to C.A.C.); National Heart, Lung and Blood Institute Cooperative Agreement U01-HL111478 (to C.A.C.); National Institute of Occupational Safety and Health Grant R01-OH0103001 (to C.A.C.); National Institute on Aging Grant R01-AG044416 (to C.A.C.); and an endowed professorship provided to Harvard Medical School by Cephalon, Inc. (to C.A.C.). The authors acknowledge the contributions of Joseph Ronda for his technical expertise, and Brandon Lockyer for scoring of the EEG and EOG data.

References

Anderson, C., Chang, A.-M., Sullivan, J.P., Ronda, J.M., Czeisler, C.A., 2013. Assessment of drowsiness based on ocular parameters detected by infrared reflectance oculography. J. Clin. Sleep Med. 9 (9), 907.

Ayas, N.T., Barger, L.K., Cade, B.E., Hashimoto, D.M., Rosner, B., Cronin, J.W., et al., 2006. Extended work duration and the risk of self-reported percutaneous injuries in interns. JAMA 296 (9), 1055–1062.

Barger, L.K., Cade, B.E., Ayas, N.T., Cronin, J.W., Rosner, B., Speizer, F.E., Czeisler, C.A., 2005. Extended work shifts and the risk of motor vehicle crashes among interns. N. Engl. J. Med. 352 (2), 125–134. http://dx.doi.org/10.1056/NEJMoa041401.

Belyavin, A., Wright, N.A., 1987. Changes in electrical activity of the brain with vigilance. Electroencephalogr. Clin. Neurophysiol. 66 (2), 137–144.

Clark, J., 2008. Will Your Next Car Wake You up when You Fall Asleep at the Wheel? Retrieved June 15th, 2016, from http://auto.howstuffworks.com/car-driving-safety/safety-regulatory-devices/car-wake-you-up.htm.

Connor, J., Norton, R., Ameratunga, S., Robinson, E., Civil, I., Dunn, R., Bailey, J., Jackson, R., 2002. Driver sleepiness and risk of serious injury to car occupants: population based case control study. BMJ 324 (7346), 1125.

Daza, I., Hernandez, N., Bergasa, L., Parra, I., Yebes, J., Gavilan, M., et al., 2011. Drowsiness monitoring based on driver and driving data fusion. Paper Presented at the Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference On.

Dingus, D.F., Grace, R., 1998. PERCLOS: A Valid Psychophysiological Measure of Alertness as Assessed by Psychomotor Vigilance. US Department of Transportation, Federal Highway Administration Publication Number FHWA-MCRT-98-006.

Douglas, N. (2001).Excessive daytime sleepiness and the sleep apnoea/hypopnoea syndrome: a major public health problem. Monaldi archives for chest disease = Archivio Monaldi per le malattie del torace/Fondazione clinica del lavoro, IRCCS [and] Istituto di clinica tisiologica e malattie apparato respiratorio, Universita di Napoli, Secondo ateneo, 56(1), 64–66.

Ftouni, S., Sletten, T.L., Howard, M., Anderson, C., Lenné, M.G., Lockley, S.W., Rajaratnam, S.M., 2013. Objective and subjective measures of sleepiness, and their associations with on-road driving events in shift workers. J. Sleep Res. 22 (1), 58–69.

Fuchs, B.D., McMaster, J., Smull, G., Getsy, J., Chang, B., Kozar, R.A., 2001. Underappreciation of sleep disorders as a cause of motor vehicle crashes. Am. J. Emerg. Med. 19 (7), 575–578.

Gold, D.R., Rogacz, S., Bock, N., Tosteson, T.D., Baum, T.M., Speizer, F.E., Czeisler, C.A., 1992. Rotating shift work, sleep, and accidents related to sleepiness in hospital nurses. Am. J. Public Health 82 (7), 1011–1014.

Harding, K., Feldman, M., 2008. Sleep disorders and sleep deprivation: an unmet public health problem. J. Am. Acad. Child Adolesc. Psychiatry 47 (4), 473–474.

Harris, W., Mackie, R.R., 1972. A Study of the Relationships Among Fatigue, Hours of Service, and Safety of Operations of Truck and Bus Drivers.

Hartley, L., Arnold, P.K., Smythe, G., Hansen, J., 1994. Indicators of fatigue in truck drivers. Appl. Ergon. 25 (3), 143–156.

Horne, J., Reyner, L., 1999. Vehicle accidents related to sleep: a review. Occup. Environ.

Med. 56 (5), 289–294.
Jackson, M.L., Kennedy, G.A., Clarke, C., Gullo, M., Swann, P., Downey, L.A., et al., 2016. The utility of automated measures of ocular metrics for detecting driver drowsiness during extended wakefulness. Accid. Anal. Prev. 87, 127–133. http://dx.doi.org/10.1016/j.aap.2015.11.033.
Johns, M.W., Tucker, A., Chapman, R., Crowley, K., Michael, N., 2007. Monitoring eye and eyelid movements by infrared reflectance oculography to measure drowsiness in drivers. Somnologie-Schlafforschung und Schlafmedizin 11 (4), 234–242.
Khushaba, R.N., Kodagoda, S., Lal, S., Dissanayake, G., 2011. Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm. IEEE Trans. Biomed. Eng. 58 (1), 121–131.
Kunert, K., King, M.L., Kolkhorst, F.W., 2007. Fatigue and sleep quality in nurses. J. Psychosoc. Nurs. Ment. Health Serv. 45 (8), 30–37.
Lal, S.K.L., Craig, A., 2001. A critical review of the psychophysiology of driver fatigue. Biol. Psychol. 55, 173–194.
Lee, M.L., Swanson, B.E., Horacio, O., 2009. Circadian timing of REM sleep is coupled to an oscillator within the dorsomedial suprachiasmatic nucleus. Curr. Biol. 19 (10), 848–852.
Lee, M.L., Howard, M.E., Horrey, W.J., Liang, Y., Anderson, C., Shreeve, M.S., et al., 2016. High risk of near-crash driving events following night-shift work. Proc. Natl. Acad. Sci. 113 (1), 176–181.
Li, G., Chung, W.-Y., 2013. Detection of driver drowsiness using wavelet analysis of heart rate variability and a support vector machine classifier. Sensors 13 (12), 16494–16511.
Li, G., Chung, W.Y., 2015. A context-aware EEG headset system for early detection of driver drowsiness. Sensors 15 (8), 20873–20893. http://dx.doi.org/10.3390/s150820873.
Liang, Y., Lee, J.D., 2014. A hybrid Bayesian Network approach to detect driver cognitive distraction. Transp. Res. Part C: Emerg. Technol. 38, 146–155.
McDonald, A.D., Lee, J.D., Schwarz, C., Brown, T.L., 2014. Steering in a random forest ensemble learning for detecting drowsiness-Related lane departures. Hum. Factors: J. Hum. Factors Ergon. Soc. 56 (5), 986–998.
McMenamin, T.M., 2007. A time to work: recent trends in shift work and flexible schedules. Monthly Lab. Rev. 130, 3.
Nakayama, O., Futami, T., Nakamura, T., Boer, E.R., 1999. Development of a steering entropy method for evaluating driver workload. In: SAE Technical Paper Series: #1999-01-0892: Presented at the International Congress and Exposition. Detroit, Michigan, March 1–4.
Philip, P., Sagaspe, P., Taillard, J., Valtat, C., Moore, N., Åkerstedt, T., et al., 2005. Fatigue, sleepiness, and performance in simulated versus real driving conditions. SLEEP 28 (12), 1511–1516.
Roenneberg, T., Wirz-Justice, A., Merrow, M., 2003. Life between clocks: daily temporal patterns of human chronotypes. J. Biol. Rhythms 18 (1), 80–90.
Rosekind, M.R., Smith, R.M., Miller, D.L., Co, E.L., Gregory, K.B., Webbon, L.L., et al., 1995. Alertness management: strategic naps in operational settings. J. Sleep Res. 4 (s2), 62–66.
Sahayadhas, A., Sundaraj, K., Murugappan, M., 2012. Detecting driver drowsiness based on sensor: a review. Sensor 12, 16937–16953. http://dx.doi.org/10.3390/s121216937.
Santamaria, J., Chiappa, K.H., 1987. The EEG of Drowsiness. Demos Publications.
Schleicher, R., Galley, N., Briest, S., Galley, L., 2008. Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? Ergonomics 51 (7), 982–1010.
Silva, E.J., Wang, W., Ronda, J.M., Wyatt, J.K., Duffy, J.F., 2010. Circadian and wake-dependent influences on subjective sleepiness, cognitive throughput, and reaction time performance in older and young adults. Sleep 33 (4), 481–490.
Stutts, J.C., Wilkins, J.W., Osberg, J.S., Vaughn, B.V., 2003. Driver risk factors for sleep-related crashes. Acc. Anal. Prev. 35 (3), 321–331.
Vicente, J., Laguna, P., Bartra, A., Bailón, R., 2016. Drowsiness detection using heart rate variability. Med. Biol. Eng. Comput. 54 (6), 927–937.
Wilkinson, V.E., Jackson, M.L., Westlake, J., Stevens, B., Barnes, M., Swann, P., et al., 2013. The accuracy of eyelid movement parameters for drowsiness detection. J. Clin. Sleep Med. 9 (12), 1315–1324.
Williamson, A., Lombardi, D.A., Folkard, S., Stutts, J., Courtney, T.K., Connor, J.L., 2011. The link between fatigue and safety. Acc. Anal. Prev. 43 (2), 498–515.
Yabuta, K., Iizuka, H., Yanagishima, T., Kataoka, Y., Seno, T., 1985. The Development of Drowsiness Warning Devices. SAE Technical Paper.