THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY
2011, 64 (9), 1714–1725

Psychology Press
Taylor & Francis Group

Do complex models increase prediction of complex behaviours? Predicting driving ability in people with brain disorders

Carrie R. H. Innes[1,2], Dominic Lee[3], Chen Chen[3], Agate M. Ponder-Sutton[3], Tracy R. Melzer[1,4], and Richard D. Jones[1,2,4,5,6]

[1]Van der Veer Institute for Parkinson's and Brain Research, Christchurch, New Zealand
[2]Department of Medical Physics and Bioengineering, Christchurch Hospital, Christchurch, New Zealand
[3]Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand
[4]Department of Medicine, University of Otago, Christchurch, New Zealand
[5]Department of Electrical & Computer Engineering, University of Canterbury, Christchurch, New Zealand
[6]Department of Psychology, University of Canterbury, Christchurch, New Zealand

Prediction of complex behavioural tasks via relatively simple modelling techniques, such as logistic regression and discriminant analysis, often has limited success. We hypothesized that to more accurately model complex behaviour, more complex models, such as kernel-based methods, would be needed. To test this hypothesis, we assessed the value of six modelling approaches for predicting driving ability based on performance on computerized sensory–motor and cognitive tests (*SMCTests*^{TM}) in 501 people with brain disorders. The models included three models previously used to predict driving ability (discriminant analysis, DA; binary logistic regression, BLR; and non-linear causal resource analysis, NCRA) and three kernel methods (support vector machine, SVM; product kernel density, PK; and kernel product density, KP). At the classification level, two kernel methods were substantially more accurate at classifying on-road pass or fail (SVM 99.6%, PK 99.8%) than the other models (DA 76%, BLR 78%, NCRA 74%, KP 81%). However, accuracy decreased substantially for all of the kernel models when cross-validation techniques were used to estimate prediction of on-road pass or fail in an independent referral group (SVM 73–76%, PK 72–73%, KP 71–72%) but decreased only slightly for DA (74–75%) and BLR (75–76%). Cross-validation of NCRA was not possible. In conclusion, while kernel-based models are successful at modelling complex data at a classification level, this is likely to be due to overfitting of the data, which does not lead to an improvement in accuracy in independent data over and above the accuracy of other less complex modelling techniques.

Keywords: Human performance modelling; Driving; Brain disorders; Prediction; Cross-validation.

Correspondence should be addressed to Carrie R. H. Innes, Van der Veer Institute for Parkinson's and Brain Research, 66 Stewart Street, Christchurch 8011, New Zealand. E-mail: carrie.innes@vanderveer.org.nz

This work was supported through a Canterbury Medical Research Foundation fellowship and a Road Safety Trust grant. We would like to thank the occupational therapists who undertook the off- and on-road assessments for this study, Julie Severinsen and Jane Gray (Driving and Vehicle Assessment Service, Burwood Hospital, Christchurch), Kevin O'Leary and colleagues (Kevin O'Leary & Associates, Wellington), and Jenny Oxley, Barbara Brook, and colleagues (Organisation of Therapy and Rehabilitation Services, Hamilton).

http://www.psypress.com/qjep
DOI:10.1080/17470218.2011.555821

Neurological disorders often affect an individual's ability to perform complex tasks. This is a particular problem when performance is impaired on a complex task, such as driving, which has the potential to lead to serious injuries or fatalities (Hawley, 2001; Hunt, Morris, Edwards, & Wilson, 1993; Lings & Jensen, 1991; Wood, Worringham, Kerr, Mallon, & Silburn, 2005). Thus, it is important that appropriate methods are available for identifying persons no longer able to safely perform complex tasks, such as driving, due to a neurological disorder.

We have developed a comprehensive battery of computerized sensory–motor and cognitive tests (*SMCTests*™) as an assessment tool in neurology and neurorehabilitation (Jones, Donaldson, Parkin, & Coppage, 1990), with particular application to the assessment of driving abilities in patients with neurological disorders (Heitger et al., 2004; Innes, Jones, Anderson, Hollobon, & Dalrymple-Alford, 2009; Innes et al., 2007; Jones & Donaldson, 1995; Jones, Donaldson, & Parkin, 1989; Jones, Sharman, Watson, & Muir, 1993; Kondraske, 2006). Our aim is to use an off-road assessment of cognitive and sensory–motor functions to predict driving ability and, thus, avoid unnecessary, subjective, and potentially highly risky on-road driving assessments. However, driving is a highly complex task that we, and others, have had limited success in predicting based on off-road tests (Dobbs, 2005; Fischer, Kondraske, & Stewart, 2002; Innes et al., 2007; Innes, Jones, Dalrymple-Alford, & Severinsen, 2009; Nouri & Lincoln, 1992, 1993). For example, in a study of 200 drivers with brain disorders, we were only able to obtain a classification accuracy for pass or fail of 70% using a binary logistic regression (BLR) or a nonlinear causal resource analysis (NCRA) model based on a subset of *SMCTests* measures (Innes, Jones, Dalrymple-Alford, et al., 2009). However, we considered that it might be possible to increase the predictive accuracy through the use of complex nonparametric kernel-based modelling techniques such as support vector machines (SVM; Burges, 1998; Hsu, Chang, & Lin, 2003) or classifiers built using kernel product (KP) or product kernel (PK) estimators (Chen, 2009; Cooley & MacEachern, 1998; Jebara, Kondor, & Howard, 2004).

While the classification accuracy of a model indicates how well the model represents the data it is modelling, it is crucial to determine how well a model generalizes to an independent group. Throughout this paper, "classification" is used to describe the method of using a single sample to both train and test a model of performance. "Prediction" is used to describe the method of testing a model on an independent sample or of estimating such through statistical procedures such as k-fold or leave-one-out cross validation.

Our aim was to determine whether kernel-based modelling techniques (SVM, KP, and PK) could improve the predictive accuracy of *SMCTests*-based models over and above the performance of other modelling techniques such as discriminant analysis (DA), BLR, or NCRA (Vasta & Kondraske, 1994).

Method

Participants

A consecutive sample of 509 people with brain disorders were recruited to the study. Data from 8 participants were removed from the final analysis due to incomplete or missing data, leaving a final sample of $n = 501$ (374 males and 127 females). Study participants had been referred to one of three driving assessment services based in the New Zealand cities of Christchurch, Hamilton, and Wellington. All referrals wished to return to, or continue, driving despite a medical condition that might have affected their driving ability. Referrals for driving assessment were made by general practitioners, District Health Board practitioners, or the Accident Compensation Corporation. Referrals were required to be free from any unrelated diagnosed psychiatric illness and to have use of their lower limbs. Referrals either held a current full driver's licence or had held one prior to the brain disorder. Ethical approval for this study was obtained from the New Zealand Multi Region Ethics Committee.

Apparatus

Study participants used a steering wheel, indicator stick, accelerator, and brake pedals to respond to computer-generated test stimuli displayed on a screen with a visual angle of ±11.3 deg. The *SMCTests* program, run by an assessor on a separate monitor, generated the tests, analysed performance, stored biographical and test data in a database, and printed performance summaries.

Sensory–motor and cognitive tests (SMCTests)

The subset of sensory–motor and cognitive tests (*SMCTests*) used in this study comprised a visuomotor reaction time and speed test (Ballistic Movement), two visuomotor coordination tests (Sine Tracking and Random Tracking), a complex sustained attention test (Complex Attention), a visual scanning test (Arrows Perception), a combined visuomotor coordination and visual scanning test (Divided Attention), and a visuomotor planning test (Planning). *SMCTests* are briefly described below and in detail elsewhere (Christchurch Neurotechnology Research Programme, 2006; Heitger et al., 2004; Innes, Jones, Anderson, et al., 2009; Innes et al., 2007; Jones, 2006; Jones & Donaldson, 1995; Jones et al., 1989; Jones et al., 1993).

Ballistic Movement measures the reaction time and maximum speed at which a participant can turn the steering wheel to move an arrow out of a box and across a pass-line in response to an unpredictable signal (latency 3–7 s).

The two visuomotor coordination tracking tests, Sine and Random Tracking, measure the accuracy (mean absolute error) with which a participant can track a laterally moving target (preview of 8 s) using the steering wheel to move a horizontally moving arrow. The tracking target is either a sine wave (Sine Tracking) or a random wave (Random Tracking).

Complex Attention assesses ability to sustain attention over an extended period of time. Participants must turn the steering wheel from left to right repeatedly to maintain an arrow in a box on the same side of the screen as a green light symbol is being presented. Variability in reaction times is analysed to identify lapses in concentration.

Divided Attention assesses ability to divide attention between two simultaneously performed separate activities. Random Tracking is combined with a simultaneous visual scanning task (Arrows Perception). While the participant tracks the random target, consecutive sets of four arrows are displayed. The participant has to maintain accurate tracking of the target, while determining whether the arrows are pointing in the same direction or not. Subjects were tested separately on Random Tracking and Arrow Perception in order to obtain baseline performance on the component tasks.

Planning assesses ability to use accurate timing and judgement as an indicator of planning ability. The participant is presented with a screen showing a bird's eye view of a road and surrounds. When the participant presses the accelerator, the road and surrounds scroll down the screen. The blue car must drive as far as possible in 6 min while avoiding all hazards.

On-road assessment

An on-road assessment provided the criterion measure ("gold standard") of driving ability. Performance during the on-road assessment was evaluated by an occupational therapist and an independent driving instructor. The occupational therapist knew the age and reason for referral for all participants. However, both the occupational therapist and driving instructor were blinded to performance on *SMCTests*. The driving instructor was seated in the front passenger seat and was responsible for giving directions to the subject and for maintaining the safety of the vehicle. The occupational therapist, experienced in driving assessment and rehabilitation of persons with brain disorders and/or disabilities, was seated in the rear of the car.

All assessments began with the subject's ability to control the initial starting and stopping of the vehicle being assessed. Subjects were then asked to drive to a residential suburb that experiences little traffic during the day but includes controlled (give-way and stop-sign controlled) and uncontrolled intersections. Subjects were then asked to drive in increasingly busy and complicated traffic situations. Traffic hazards included single-lane roundabouts, dual-lane roundabouts, dual-lane roads, controlled intersections (give-way, stop-sign, and traffic-light controlled), uncontrolled intersections, and changes in speed zone (i.e., 50 km/hr, 60 km/hr, and 80 km/hr sections). Assessments were approximately 45 min in duration. However, if the occupational therapist or driving instructor considered that their safety or the safety of the vehicle or other road users was at risk at any stage during the assessment, the assessment was terminated, and the driving instructor drove the vehicle back to the starting point.

On-road driving performance was scored as a pass or a fail. Assessment was defined by four areas of driving deemed necessary for safe and able driving: search, hazard identification, controls, and observation of traffic regulations, each subdivided into specific components. Driving performance was also scored using a driving scale (outlined in Innes et al., 2007). A driving score was determined by mutual agreement by the two assessors. If a subject failed the on-road assessment, they were given a driving score of 0–5. Subjects who passed were given a driving score of 6–10. Exact scores were defined by the number of observed driving errors, whether these were considered major or minor, and whether the participant was able to subsequently correct errors. A copy of the driving scale is available at www.neurotech.org.nz/files/Driving_Scale.pdf (Innes & Jones, n.d.).

Data analysis

As none of the *SMCTests* data were normally distributed (Shapiro–Wilk *W* test and Lilliefors probabilities, $p < .05$), and several measures were ordinal, nonparametric techniques were used to analyse the data at group level. Mann–Whitney *U* analysis was undertaken to determine significant differences in off-road test performance between the referrals who passed the on-road assessment and those who failed. The Cohen-type effect size statistic for rank-transformed variables (Hopkins, 2000) was used to evaluate the magnitude of differences in off-road performance between the pass and fail referral groups.

The subset of seven *SMCTests* provided 24 key measures. However, to prevent problems with collinearity, intercorrelations between test measures were calculated to identify test measures showing a low tolerance ($<.2$), which indicates multicollinearity of that variable with other variables, or that were highly correlated ($r \geq .80$) with another measure.

Five nonparametric modelling techniques were used to determine the classification and predictive value of performance on *SMCTests* for on-road driving ability at the individual level for referrals—binary logistic regression, nonlinear causal resource analysis, and three kernel methods (product kernel density, kernel product density, and support vector machine). In addition, a common parametric modelling technique, discriminant analysis, was used for comparison.

Discriminant analysis (DA) is a parametric classification technique used to classify input data into two or more mutually exclusive groups. DA works on the same principle as a multivariate analysis of variance, where, based on a matrix of variances and covariances, variables are assessed to determine whether they discriminate between groups (i.e., whether the mean of a variable differs between groups; StatSoft, 2003). A backward stepwise method was used with our data to select the optimal set of *SMCTests* variables for discriminating between the on-road assessment fail and pass groups. Amongst other considerations, DA assumes that the data form part of a normal distribution. It is important to note that the number of variables offered to the model must follow the rule of thumb that there be at least 5, if not 10 times, as many cases as independent variables in a regression analysis to minimize the risk of overfitting the model to the sample (Hosmer & Lemeshow, 2000; Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996; Tabachnick & Fidell, 2001). This rule must be applied *prior* to implementation of the stepwise regression. In the current study, enough data were obtained to apply a conservative 1:20 ratio of independent variables to participants prior to the stepwise regression.

Binary logistic regression (BLR) is a nonparametric classification method for the case where the dependent variable is dichotomous (i.e., pass or fail on an on-road driving assessment). BLR was used to estimate the probability of an on-road assessment fail based on an exponential function of *SMCTests* variables and weightings. As with DA, a backward stepwise method was used to select the optimal set of *SMCTests* variables for predicting on-road assessment outcome. Variables that explain a significant amount of the variance in the dependent variable are weighted along with the other entered variables to form an equation of best fit. The model uses a stepwise elimination procedure to remove variables that do not explain a significant amount of variance in the dependent measure. As with DA, a conservative ratio of one independent variable to 20 cases was applied prior to implementation of the BLR stepwise regression. In contrast to DA, BLR makes no assumptions about the distributions of the variables and can use categorical, ordinal, and interval data.

Nonlinear causal resource analysis (NCRA) is a performance prediction method based on the resource economic performance modelling constructs of general systems performance theory and the elemental resource model (Kondraske, 2006). With NCRA, the minimum resource level required to achieve a given level of performance on a high-level task is determined for each test function and is plotted as a resource demand function (RDF) curve (Fischer et al., 2002; Kondraske, 2006). RDF curves were created for key performance measures from each of the *SMCTests* tests. A major benefit of NCRA is that it can then determine the specific test function that maximally limited each subject's performance on the high-level driving task. NCRA-predicted scores for each referral were compared with observed Driving Scale scores (0−10) in order to determine the accuracy of the NCRA model predictions. For classification purposes, Driving Scale scores <6 were classified as fail while scores ≥ 6 were classified as pass.

Kernel-based classification methods can be used to solve nonlinear problems by mapping nonlinear observations into a higher dimensional space where a linear or nonlinear classifier can then be used—this is known as the Kernel Trick, which is based on theories developed by Vapnik (1998). Two of the kernel-based classification approaches used with our data involved modelling the probability density function for each group (i.e., pass or fail on an on-road driving assessment) for each test measure (Parzen, 1962), then using the Bayes classifier to estimate the probability that an observation came from the pass or fail group. While several density estimation techniques could be used, two nonparametric density estimators were implemented: product kernel density estimator (PK) and kernel product density estimator (KP; Chen, 2009; Cooley & MacEachern, 1998; Jebara et al., 2004). As KP makes a stronger assumption than PK that the features (i.e., test measures) are independent, KP may not be as suitable for representing data that have a strong dependence between feature variables.

Finally, a sparse kernel machine (support vector machine, SVM) approach to classification was implemented. SVM is a classification method that uses hyperplanes in multidimensional space to maximally separate data into defined categories (i.e., pass or fail on an on-road driving assessment). The margin of the separating hyperplane is the distance from the hyperplane to the nearest data point of either category. The data points that lie along the margins of the separating hyperplanes are called the support vectors. When data are not completely separable, an extra cost is assigned within the equation for errors. Our data were modelled using a nonlinear radial basis function kernel. See Hsu et al. (2003) for a practical guide to support vector classification or Burges (1998) for a more detailed description of the underlying theory behind SVM.

Leave-one-out was used to estimate the true error rate of the models in an independent test set. In leave-one-out cross-validation, each case is left out in turn while the remaining data are used to train the model; the model is then tested on the single case that was left out (Witton & Frank, 1999). This method has the advantage that the greatest possible amount of data can be used to train each model. If the population in the training set is representative of the population that the predictive model will be used with, cross-validation provides a sound estimate of the predictive accuracy that would be achieved with a separate test dataset (Witton & Frank, 1999). In addition, the results of the leave-one-out cross-validation were compared to the results of a repeated stratified 10-fold cross-validation. Ten-fold cross-validation is similar to leave-one-out except instead of leaving one case out, the data are split into 10 groups (stratified for proportion of pass and fail), and each group is left out in turn while the remaining 90% of data are used to train the model; the model is then tested on the 10% of data that were left out. The 10-fold cross-validation was repeated 100 times by using a random number generator to sort and split the data into 10 different groups before repeating the training and testing. Unfortunately, there is no method for automated 10-fold or leave-one-out cross-validation of the NCRA model, and, thus, the predictive accuracy of the NCRA model could not be estimated.

DA and BLR both employ a backward stepwise method for variable selection, which has been shown in our earlier data to increase the generalizability of the model to independent data, compared to an Enter model (personal communication), presumably by minimizing overfitting. However, the kernel methods do not have an in-built method for variable selection. Thus, two model-independent methods for ranking the importance of variables were used to identify a reduced set of 10 *SMCTests* variables in order to assess the impact of a reduced number of input variables on KP, PK, and SVM model classification and prediction accuracy. The first reduced set of variables was selected by identifying the 10 *SMCTests* variables with the largest Cohen-type effect size statistic for rank-transformed variables for separating pass and fail groups. The second reduced set of variables was selected by performing receiver-operating characteristic (ROC) curve analysis on each predictor and using the area under the curve to identify the top 10 ranked variables.

The performance of the six modelling techniques for classification and prediction of driving assessment outcome was assessed based on the area under the ROC curve (ROC AUC), accuracy, sensitivity (% detection of on-road assessment fails), specificity (% detection of on-road assessment passes), positive predictive value (PPV), and negative predictive value (NPV). PPV (also known as the “predictive value of a positive test”, PV+) provides a measure of the proportion of referrals predicted to fail who did fail. NPV (also known as the “predictive value of a negative test”, PV−) provides a measure of the proportion of referrals predicted to pass who did pass. Comparison of ROC curves to test the statistical significance of the difference between the areas was undertaken using the method of DeLong, DeLong, and Clarke-Pearson (1988; MedCalc Software 11.3.1).

Results

All 501 referrals recruited to the study had a definite or suspected brain disorder as follows: 163 suspected or probable dementia, 153 stroke, 113 traumatic brain injury, 27 Parkinson’s disease, 9 brain tumours, and 36 other neurological disorders. Referrals had a mean age of 66.3 years (*SD* 18.9, range 17−92).

Of the 501 referrals, 207 (41%) failed the on-road driving assessment, and 294 (59%) passed. A total of 50% of fails ($n = 103$) had their on-road assessment terminated early due to serious safety concerns. A total of 6% of failed on-road assessments ($n = 12$) were terminated within 20 min, 26% ($n = 54$) between 20−30 min, and 18% ($n = 37$) between 30−40 min.

Mann−Whitney *U* analysis showed that there were differences in off-road test performance between the group of referrals who passed the on-road assessment and those who failed on each of the 24 key *SMCTests* measures (Table 1). Cohen-type effect sizes of the difference between pass and fail groups were all significant ($p < .05$) and ranged from 0.37−1.20. Cohen-type effect sizes were highest for measures of visuomotor

Table 1. *SMCTests performance difference between the on-road assessment pass and fail groups based on Mann–Whitney U analysis*

SMCTests measure	*Pass referrals (n = 294)* *Median*	*Fail referrals (n = 207)* *Median*	*Effect size of pass vs. fail*[a]
Random Tracking: mean error (mm) Run 2	7.5	14.3	1.20
Divided Attention: tracking mean error (mm)	8.3	14.5	1.15
Complex Attention: reaction time (ms)	495	689	1.09
Sine Tracking: mean error (mm) Run 2	10.0	16.4	1.05
Random Tracking: mean error (mm) Run 1	7.8	14.4	1.04
Planning: duration of lateral position errors (s)	4.7	13.8	1.04
Ballistic Movement: reaction time (ms)	369	459	0.99
Sine Tracking: mean error (mm) Run 1	14.2	21.7	0.94
Complex Attention: movement time (ms)	346	465	0.94
Planning: safety margin at intersections (mm)	39.0	23.0	0.85
Complex Attention: movement time *SD* (ms)	54	97	0.80
Divided Attention: nonresponses	0.0	0.0	0.80
Ballistic Movement: peak velocity (mm/s)	805	666	0.79
Ballistic Movement: movement time (ms)	263	319	0.78
Planning: hazards hit	1.0	3.0	0.77
Complex Attention: reaction time *SD* (ms)	139	269	0.74
Complex Attention: lapses	0.0	1.0	0.72
Planning: crashes with other vehicles	0.0	1.0	0.69
Divided Attention: arrows correct	12.0	11.0	0.63
Complex Attention: invalid trials	0.0	0.0	0.60
Arrows Perception: nonresponses	0.0	0.0	0.49
Planning: lateral position error (mm)	2.9	3.2	0.46
Planning: distance travelled (m)	4.0	3.6	0.43
Arrows Perception: arrows correct	12.0	12.0	0.37

[a]Effect size calculated using a Cohen-type effect-size statistic for rank-transformed variables (Hopkins, 2000); $p < .05$.

coordination, divided attention, and sustained attention.

Referrals who failed the on-road assessment were older than those who passed (median 79 years vs. 63 years, Mann–Whitney $U = 15{,}193$, $p < .001$).

Due to the large data set obtained for this study, all key *SMCTests* measures could be included in the analysis while still maintaining a conservative 1:20 ratio of independent variables to participants. However, to avoid problems with collinearity, 5 of the 24 key measures were removed from the classification analysis due to low tolerance (<0.2) or intercorrelation ($r \geq .80$) with other measures. Measures removed were Ballistic Movement peak velocity, Random Tracking absolute mean error Run 2, Sine Tracking absolute mean error Run 2, Arrows Perception number of correct trials, and Divided Attention Arrows Perception number of correct trials.

The 19 *SMCTests* measures plus age were offered to the backward stepwise DA analysis, which produced a model with five measures—Divided Attention Tracking mean absolute error, Complex Attention reaction time, Age, Planning intersection safety margin, and Planning hazards hit. The area under the ROC curve (ROC AUC) of the DA model was 0.84. Based on the ROC curve, the optimal cut-point for the model that gave the highest mean sensitivity and specificity was determined to be 0.45. Based on the optimized cut-point, the DA model correctly classified 380 of the 501 referrals (i.e., 76% accuracy) as an on-road pass or fail.

The backward stepwise BLR analysis produced a model with eight measures—Divided Attention Tracking mean absolute error, Divided Attention Arrows Perception nonresponses, Complex Attention reaction time, Complex Attention number of invalid trials, Planning duration of lateral position faults, Planning number of hazards hit, Planning intersection safety margin, and age. The ROC AUC of the BLR model was 0.84. Based on an optimized cut-point of 0.44, the BLR model correctly classified 389 of the 501 referrals (i.e., 78% accuracy) as an on-road pass or fail.

NCRA produced resource demand function curves for the 20 measures to model the data. The ROC AUC of the NCRA model was 0.80. Based on an optimized RDF curve-predicted driving scale score of less than 7.8 being fail, the model correctly classified 371 of the 501 referrals (i.e., 74% accuracy).

The ROC AUC of the PK model was 0.99. Based on an optimized cut-point of 0.41, the PK model used all 20 measures to correctly classify 500 of the 501 referrals (i.e., 99.8% accuracy).

The ROC AUC of the KP model was 0.87. Based on an optimized cut-point of 0.69, the KP model used all 20 measures to correctly classify 407 of the 501 referrals (i.e., 81% accuracy).

The ROC AUC of the SVM model was 0.98. Based on an optimized cut-point of 0.48, the SVM model used all 20 measures to correctly classify 499 of the 501 referrals (i.e., 99.6% accuracy).

Compared to a default cut-point of 0.50, using an optimized cut-point only improved the accuracy of the DA, BLR, KP, PK, and SVM classification models by 1%. However, an optimized cut-point improved the classification accuracy of the NCRA model from 66% to 74%.

Based on the optimized cut-points determined by the classification model ROC curve analysis, leave-one-out cross-validation analysis estimated that the five models would correctly predict 72–76% of an independent test set to pass or fail an on-road assessment (DA 75%, BLR 76%, PK 73%, KP 72%, SVM 76%). The accuracy of the leave-one-out prediction models was the same whether optimized cut-points or the default cut-point of 0.50 were used. Stratified 10-fold cross-validation with 100 repeats gave very similar but slightly lower estimated accuracies for the five models (DA 74%, BLR 75%, PK 72%, KP 71%, SVM 73%) than leave-one-out cross-validation and at a much higher computational cost.

The 10 variables with the largest Cohen-type effect sizes were used as a reduced set of input variables for the PK, KP, and SVM models. With this reduced set of input variables, the classification accuracy of the PK model remained unchanged (PK 99.8%). However, there was reduced classification accuracy in the remaining models (KP 77%, SVM 89%), as well as reduced leave-one-out cross-validation accuracy across all three models (PK 70%, KP 71%, SVM 72%). The 10 variables with the largest ROC AUCs were used as a second reduced set of input variables. With this reduced set of input variables, the classification accuracy of the PK model remained unchanged (PK 99.8%) while the accuracy of the other models was reduced (KP 76%, SVM 97%). Leave-one-out cross-validation accuracy was also reduced across all three models (PK 66%, KP 68%, SVM 72%).

A summary of the classification and prediction performance of the models, in terms of ROC AUC, sensitivity, specificity, NPV, PPV, and overall accuracy, is provided in Table 2. The comparative discrimination of the models for distinguishing between those who pass or fail the on-road driving assessment was assessed by testing the statistical significance of the difference between the areas under the ROC curves (DeLong et al., 1988). The discrimination of the SVM and PK models was superior to the other models at the classification level ($p < .001$). However, at the prediction level, the discrimination of SVM was not different to that of any of the other models. The BLR model was superior to PK and KP ($p < .001$) but not different to SVM or DA. DA was superior to KP ($p = .041$) but not different to any other model.

There was no difference in the mean ages of female or male referrals. However, females were

Table 2. *Performance of models for classification and prediction of on-road driving based on SMCTests performance and age and using optimized cut-points*

		ROC AUC (95% CI)	*Sensitivity (%)*	*Specificity (%)*	*PPV (%)*	*NPV (%)*	*Accuracy (%)*
Classification	PK	0.99 (0.99–1.00)	99.5	100.0	100.0	99.7	99.8
	SVM	0.98 (0.96–1.00)	99.5	99.7	99.5	99.7	99.6
	KP	0.87 (0.84–0.91)	64.7	92.9	86.5	78.9	81.2
	BLR	0.84 (0.81–0.88)	72.9	81.0	72.9	81.0	77.6
	DA	0.84 (0.80–0.87)	78.7	73.8	67.9	83.1	75.8
	NCRA	0.80 (0.76–0.84)	69.1	77.6	68.4	78.1	74.1
Leave-one-out cross-validation	PK	0.78 (0.74–0.82)	66.7	77.9	68.0	76.8	73.3
	SVM	0.82 (0.79–0.86)	74.4	76.9	69.4	81.0	75.8
	KP	0.77 (0.72–0.81)	57.5	81.6	68.8	73.2	71.7
	BLR	0.83 (0.79–0.86)	69.1	81.0	71.9	78.8	76.0
	DA	0.83 (0.79–0.86)	76.8	72.8	66.5	81.7	74.5

Note: ROC AUC = receiver-operating characteristic curve, area under the curve. CI = confidence interval. PPV = positive predictive value. NPV = negative predictive value. PK = product kernel density estimator. SVM = support vector machine. KP = kernel product density estimator. BLR = binary logistic regression. DA = discriminant analysis. NCRA = nonlinear causal resource analysis.

more likely to fail the on-road driving assessment than males (50% females vs. 39% males failed), $\chi^2(1) = 4.82$, $p = .028$. On average, the males had higher upper-limb peak velocities, had more accurate visuomotor tracking, had faster reaction and movement times on a complex sustained attention task, spent less time off-road, and had greater safety margins leading to fewer crashes with hazards and other vehicles in Planning. The male performance advantage was observed across more test measures than had been observed in previous studies in normal subjects (Innes, Jones, Anderson, et al., 2009). This notwithstanding, the Cohen-type effect sizes of the differences were generally small to moderate (0.24–0.65), and the addition of sex as a variable did not improve the accuracy of the classification models.

Referrals with suspected or probable dementia were more likely to fail the on-road driving assessment than those referred with other brain disorders (58% vs. 33% failed), $\chi^2(1) = 28.68$, $p < .001$. There was no difference in brain disorder type between males and females. However, the referrals with suspected or probable dementia were older than the referrals with other brain disorders (median 80 years vs. 63 years, Mann–Whitney $U = 11{,}932$, $p < .001$). The addition of brain disorder type as a variable did not improve the accuracy of any of the classification models.

Discussion

As driving is a very complex behaviour, we had hypothesized that more complex models may provide more accurate prediction of driving ability in people with brain disorders over and above previously used simpler modelling techniques such as binary logistic regression. However, while two of the kernel-based modelling methods (SVM and PK) were very accurate at modelling on-road driving outcome at a classification level, the kernel-based modelling methods were not more accurate than the other modelling techniques when leave-one-out cross-validation was used to estimate how well the models would generalize to independent test data. It is likely that the degrees of freedom that give the kernel-based models their strength probably led to substantial overfitting of the data.

By identifying hyperplanes that separate the data with the maximum margin, SVM models are designed to minimize structural risk. In basic terms, structural risk is the trade-off between empirical risk (i.e., measured mean error rate on the training set, which can vary depending on underfitting or overfitting the data) and actual risk (i.e., true mean error in an independent set; Burges, 1998). However, while SVM models are expected to be somewhat resilient to the problem of overfitting (Vapnik, 1999), our data show that the SVM classification model did not generalize well. Two model-independent methods for ranking the importance of variables were used to identify a reduced set of 10 *SMCTests* variables and to assess whether a reduced number of input variables might reduce any overfitting in the KP, PK, and SVM model classification models and thereby increase the generalization of the models for prediction. However, the models based on a reduced set of input variables resulted in unchanged or reduced classification and prediction accuracy. This indicates that reducing the number of explanatory variables available did not improve predictive accuracy via a reduction in overfitting in the classification models.

A recent paper reported that a repeated 10-fold cross-validation estimator may be the best method for estimating the accuracy of a classification model when an independent test set was not available (Kim, 2009). Overall, we found that 10-fold cross-validation with 100 repeats gave an almost identical range of estimated accuracies for the five models compared to leave-one-out cross-validation.

In addition to the overall accuracy of the models, the positive and negative predictive values also give important information about model performance. It is important that a driver screening assessment does not lead to a high proportion of people required to undergo unnecessary further assessment but who ultimately are assessed as safe to drive. The leave-one-out cross-validated model PPVs ranged from 67−72%, which indicates that, if used as a screening tool in an independent group of referrals, 28−33% of people predicted to fail would ultimately pass an on-road assessment. This may be a reasonable proportion of false positives if at the same time the model is able to identify the people who are truly unsafe to continue driving. However, the leave-one-out cross-validated model NPVs ranged from 73−82%, which indicates that 18−27% of people predicted to pass would actually fail the on-road assessment. Thus, if used as a screening tool without further assessment, this would mean than 18−27% of people allowed to return to driving would actually be unsafe drivers.

This study highlights the danger of relying on classification accuracy to compare models and emphasizes the need to assess the generalizability of a model in an independent data set or via cross-validation. This conclusion is supported by extensive previous work in statistics and machine learning, where it has been shown mathematically that the empirical classification accuracy is, on average, an overestimate of the generalization accuracy (Kohavi, 1995; Picard & Cook, 1984).

We have determined that, for this type of data, the estimated discrimination of the relatively simple BLR model is approximately the same as that of the DA and the more complex SVM models and is superior to that of the other more complex KP and PK models. Furthermore, leave-one-out cross-validation of model performance indicates that, irrespective of the complexity of the model, none of the models is accurate enough to be used as a complete screening tool without need for further assessment to determine driving safety.

Original manuscript received 30 September 2010
Accepted revision received 10 January 2011
First published online 26 April 2011

REFERENCES

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2,* 121–167.

Chen, C. (2009). *Classification via product kernel and kernel product density estimators with application for predicting driving ability of persons with brain disorders*. Unpublished BSc (Hons) thesis, University of Canterbury, Christchurch, New Zealand.

Christchurch Neurotechnology Research Programme (2006). *Canterbury Driving Assessment Tool (CanDAT^{TM}) incorporating SMCTests^{TM} Version 5.0—user's manual*. Christchurch, New Zealand: Canterbury District Health Board.

Cooley, C. A., & MacEachern, S. N. (1998). Classification via kernel product estimators. *Biometrika*, *85,* 823–833.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44,* 837–845.

Dobbs, A. R. (2005, June). The development of a scientifically based driving assessment and standardization procedures for evaluating medically at-risk drivers. Paper presented at the *Proceedings of the Canadian Multidisciplinary Road Safety Conference XV*. Fredericton, New Brunswick, Canada.

Fischer, C. A., Kondraske, G. V., & Stewart, R. M. (2002, October). Prediction of driving performance using nonlinear causal resource analysis. *Proceedings of the Second Joint Engineering in Medicine and Biology Society/Biomedical Engineering Society Conference*, *2*, Houston, TX, USA, pp. 2473−2474.

Hawley, C. A. (2001). Return to driving after head injury. *Journal of Neurology, Neurosurgery, and Psychiatry*, *70,* 761–766.

Heitger, M. H., Anderson, T. J., Jones, R. D., Dalrymple-Alford, J. C., Frampton, C. M., & Ardagh, M. W. (2004). Eye movement and visuomotor arm movement deficits following mild closed head injury. *Brain*, *127,* 575–590.

Hopkins, W. G. (2000). *A new view of statistics*. Internet Society for Sport Science. Retrieved 16 August, 2004, from www.sportsci.org/resource/stats/

Hosmer, D., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York, NY: Wiley & Sons.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). *A practical guide to support vector classification*. Retrieved 3 April, 2010, from www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Hunt, L., Morris, J. C., Edwards, D., & Wilson, B. S. (1993). Driving performance in persons with mild senile dementia of the Alzheimer type. *Journal of the American Geriatrics Society*, *41,* 747–752.

Innes, C. R. H., & Jones, R. D. (n.d.). Driving Scale. Retrieved June 21, 2009, from http://www.neurotech.org.nz/files/Driving_Scale.pdf

Innes, C. R. H., Jones, R. D., Anderson, T. J., Hollobon, S. G., & Dalrymple-Alford, J. C. (2009). Performance in normal subjects on a novel battery of driving-related sensory-motor and cognitive tests. *Behavior Research Methods*, *42,* 284–294.

Innes, C. R. H., Jones, R. D., Dalrymple-Alford, J. C., Hayes, S., Hollobon, S., Severinsen, J., et al. (2007). Sensory-motor and cognitive tests predict driving ability of persons with brain disorders. *Journal of the Neurological Sciences*, *260,* 188–198.

Innes, C. R. H., Jones, R. D., Dalrymple-Alford, J. C., & Severinsen, J. (2009, June). Prediction of driving ability in people with dementia- and non-dementia-related brain disorders. *Proceedings of the International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*. Big Sky, MT, USA, pp. 342−348.

Jebara, T., Kondor, R., & Howard, A. (2004). Probability product kernels. *Journal of Machine Learning Research*, *5,* 819–844.

Jones, R. D. (2006). Measurement of sensory-motor control performance capacities: Tracking tasks. In J. D. Bronzino (Ed.), *The biomedical engineering handbook: Biomedical engineering fundamentals* (3rd ed., Vol. 1, pp. 77:1–77:25). Boca Raton, FL: CRC Press.

Jones, R. D., & Donaldson, I. M. (1995). Fractionation of visuoperceptual dysfunction in Parkinson's disease. *Journal of the Neurological Sciences*, *131,* 43–50.

Jones, R. D., Donaldson, I. M., & Parkin, P. J. (1989). Impairment and recovery of ipsilateral sensory-motor function following unilateral cerebral infarction. *Brain*, *112,* 113–132.

Jones, R. D., Donaldson, I. M., Parkin, P. J., & Coppage, S. A. (1990). Impairment and recovery profiles of sensory-motor function following stroke: Single-case graphical analysis techniques. *International Disability Studies*, *12,* 141–148.

Jones, R. D., Sharman, N. B., Watson, R. W., & Muir, S. R. (1993). A PC-based battery of tests for quantitative assessment of upper-limb sensory-motor function in brain disorders. *Proceedings of 15th International Conference of IEEE Engineering in Medicine and Biology Society*, *15,* 1414–1415.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, *53,* 3735–3745.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, *2,* 1137–1143.

Kondraske, G. V. (2006). The elemental resource model for human performance. In J.D. Bronzino (Ed.), *The biomedical engineering handbook: Biomedical* *engineering fundamentals* (3rd ed., pp. 75:01–75:19). Boca Raton, FL: CRC Press.

Lings, S., & Jensen, P. B. (1991). Driving after stroke: A controlled laboratory investigation. *International Disability Studies*, *13,* 74–82.

Nouri, F. M., & Lincoln, N. B. (1992). Validation of a cognitive assessment: Predicting driving performance after stroke. *Clinical Rehabilitation*, *6,* 275–281.

Nouri, F. M., & Lincoln, N. B. (1993). Predicting driving performance after stroke. *British Medical Journal*, *307,* 482–483.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, *33,* 1065–1076.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, *99,* 1373–1379.

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, *79,* 575–583.

StatSoft (2003). *StatSoft textbook*. Retrieved, April 19, 2010, from www.statsoft.com/textbook/stathome.html

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). New York, NY: HarperCollins.

Vapnik, V. N. (1998). *Statistical learning theory*. New York, NY: John Wiley & Sons.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10,* 988–999.

Vasta, P. J., & Kondraske, G. V. (1994, November). Performance prediction of an upper extremity reciprocal task using non-linear causal resource analysis. *Proceedings of International Conference of IEEE Engineering in Medicine and Biology Society*, Baltimore, MD, USA, pp. 305–306.

Witten, I. H., & Frank, E. (1999). *Data mining*. San Francisco, CA: Morgan Kaufmann.

Wood, J. M., Worringham, C., Kerr, G., Mallon, K., & Silburn, P. (2005). Quantitative assessment of driving performance in Parkinson's disease. *Journal of Neurology, Neurosurgery, and Psychiatry*, *76,* 176–180.