Comparison of a linear and a non-linear model for using sensory–motor, cognitive, personality, and demographic data to predict driving ability in healthy older adults

Petra A. Hoggarth, a, b, *, Carrie R.H. Innes, a, c, John C. Dalrymple-Alford, a, b, d, Julie E. Severinsen, e, Richard D. Jones, a, b, c, d, f

a Van der Veer Institute for Parkinson's and Brain Research, 66 Stewart Street, Christchurch 8011, New Zealand
b Department of Psychology, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand
c Department of Medical Physics and Bioengineering, Christchurch Hospital, Private Bag 4710, Christchurch, New Zealand
d Department of Medicine, University of Otago, PO Box 4345, Christchurch 8140, New Zealand
e Department of Occupational Therapy, Burwood Hospital, Private Bag 4708, Christchurch, New Zealand
f Department of Electrical and Computer Engineering, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand

Article history:
Received 12 August 2009
Received in revised form 8 March 2010
Accepted 22 April 2010

Keywords:
Driving
Aging
Fitness to drive
Sensory–motor assessment
Cognitive assessment

1. Introduction

Drivers aged 70 and above have higher injury and death rates compared to middle-aged drivers when measured by distance travelled, per trip, or per number of licensed drivers (Langford and Koppel, 2006; Organisation for Economic Co-operation and Development, 2001; Tefft, 2008). Factors linked to unsafe driving in older adults include cerebrovascular disease, visual attention deficits, and cognitive deficits associated with dementia (Ball and Owsley, 1991, 1993; Cooper et al., 1993; Dobbs et al., 1998; Johansson et al., 1996; McGwin et al., 2000; McKnight and McKnight, 1999; Meuleners et al., 2006). Given that 25% of the population of OECD countries will be aged 65 years and over by 2050 (Organisation for Economic Co-operation and Development, 2001), the detection of at-risk older drivers is an increasingly important public health concern.

There are no universally accepted requirements for older adult driver licensing, with countries and states varying widely in testing and licence renewal procedures. Until recently, New Zealand drivers aged 80 and biennially thereafter completed compulsory on-road testing in order to maintain their driver's licence, with the opportunity to re-sit the assessment following a failing grade. This compulsory on-road testing was abolished in 2006 due to claims that the policy was ageist. New Zealand drivers are now required to procure a 'medical fitness to drive' certificate from their doctor at ages 75, 80, and biennially thereafter. Variations in assessing older
drivers’ competence prompted the current study, which examined whether formal off-road measures were an efficient way to identify at-risk older drivers. During the previous system of compulsory on-road testing, a study following 39,318 drivers found that the risk of involvement in a crash in the following two years rose 33% for each time the test had to be sat in order to receive a passing grade (Keall and Frith, 2004). An Australian study found drivers who self-reported a crash over the previous five-year period made significantly more errors in road observation, blind spot checks, braking and accelerating, and approaching hazards during an on-road driving assessment compared to drivers who reported no crashes (Wood et al., 2009). On-road driving assessments may, therefore, provide valid estimations of older people’s driving safety in real-world situations.

Measures of cognitive, sensory–motor and personality domains, as well as demographic data have previously shown utility in classifying and predicting driving outcomes in older adults. Neuropsychological tests associated with on-road driving assessment outcome in primarily healthy older drivers include lower scores on the Useful Field of View (UFOV) (De Raedt and Ponjaert-Kristoffersen, 2000; Stav et al., 2008), selective attention (De Raedt and Ponjaert-Kristoffersen, 2000; Risser et al., 2008), and measures of cognitive flexibility (De Raedt and Ponjaert-Kristoffersen, 2000). Similarly, on-road crashes are associated with lower cognitive flexibility in a 12-month retrospective self-report study (De Raedt and Ponjaert-Kristoffersen, 2000) and lower global cognitive status in five- and six-year retrospective studies of police-reported crashes (Owsley et al., 1991; Sims et al., 1998). Lower scores on the UFOV were related to the incidence of six-year retrospective police-recorded at-fault crashes (Sims et al., 1998), three-year prospective officially recorded crashes (Owsley et al., 1998), and the frequency of five-year retrospective officially recorded crashes (Owsley et al., 1991). Sensory and motor classifiers of on-road assessment performance include movement perception and response (De Raedt and Ponjaert-Kristoffersen, 2000; Sommer et al., 2008; Wood et al., 2008), reaction time (Risser et al., 2008; Sommer et al., 2008), rapid pace walking (Stav et al., 2008), and postural sway (Wood et al., 2008).

The relationship between personality traits and driving has been investigated primarily in young adult samples. A non-linear neural network classification model utilized emotional stability, accepted level of risk, and social responsibility as classifiers of on-road driving assessment outcome in a sample with a mean age of 39 years (Sommer et al., 2008). In college students, higher scores on the 14-item Driving Anger Scale (Deffenbacher et al., 1994) have been associated with increased self-reported risky driving behaviour (Dahlen and White, 2006; Deffenbacher et al., 2003, 2002; Schwebel et al., 2006), and low scores on the personality construct of conscientiousness, and high scores on sensation-seeking have been associated with higher rates of both self-reported and simulated risky driving behaviour (Schwebel et al., 2006). The only study that examined drivers aged 75 and over found that higher scores on sensation-seeking were related to self-reports of higher numbers of driving violations and tickets (Schwebel et al., 2007). The effects of personality traits on driving behaviour may prove especially relevant to drivers without significant cognitive and sensory–motor impairment.

Driving simulators have also been used to measure on-road driving outcomes. Scores on the STISIM Drive™ driving simulator have been shown to account for 65.7% of the variance in an on-road assessment in a sample with a mean age of 73 years (Lee et al., 2003a). Additionally, simulator performance classified a group of older drivers into 12-month retrospective self-reported crash incidence groups with a sensitivity of 91.4% and specificity of 82.3% (Lee et al., 2003b). However, despite continued improvements, up to 10% of older drivers experience simulator sickness to the extent that they cannot complete the assessment (Lee et al., 2003a; Schwebel et al., 2007), which limits the application of simulators in the older population.

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 the use of statistical procedures such as boot strapping or leave-one-out cross-validation.

The current study utilized a range of off-road measures across the sensory–motor, cognitive, and personality domains to construct and test classification models of on-road driving assessment scores in a sample of older drivers with no overt neurological impairment. Additionally an off-road computerized assessment battery that includes sensory–motor and cognitive tests (SMCTests™) was used. This test battery has been found useful for prediction of driving ability in people with brain disorders, primarily stroke (Innes et al., 2009a, 2007). The study also compares the classification and prediction accuracy of non-linear causal resource analysis (NCRA) and binary logistic regression (BLR). BLR is the common method employed to assess predictors of Pass/Fail driving performance, and uses measures of central tendency to find a limited number of key tests to classify group members. NCRA uses absolute values of test scores to find the most useful test for classifying each individual. Non-linear techniques have shown promise in classification and prediction of driving performance (Fischer et al., 2002; Innes et al., 2007; Risser et al., 2008; Sommer et al., 2008). The differing focus of the two modelling techniques is expected to lead to two very different models for predicting on-road driving performance, with the relative utility of each model compared and discussed.

2. Methods

2.1. Participants

A convenience sample of current older drivers was recruited from churches, recreational groups, word of mouth, and advertisements placed in two free local health magazines in Christchurch (population 369,000). Sixty participants (50% male) aged 70–84 years (mean = 76.7) were recruited, with 10 men and women in each of three age ranges (70–74, 75–79, and 80+ years); 93% identified their ethnicity as New Zealand European. The sample reported an average of 55.1 years driving experience (range 31–69 years), with males reporting more years of driving (58.8 versus 51.5 years, two-tailed Mann–Whitney U test, z = −3.82, p < .001). Men also reported driving more km per year (males median = 8693 versus 5894, two-tailed Mann–Whitney U test, z = −2.81, p = .005). Exclusion criteria included a history of moderate to severe brain injury, diagnosed neurological or cognitive disorder, severe musculoskeletal disease, and acute psychiatric disorder. No participant scored below 27 on the MMSE (mean = 28.8), suggesting none had significant cognitive impairment. Participants were free to continue driving irrespective of the outcome of their assessment and received NZ$ 50 compensation for participation. All undertook a 3-h off-road testing session that included a computerized sensory–motor and cognitive test battery, personality measures and standardized cognitive tests. On-road driving assessment was performed approximately 14 days (range 2–41 days) later. The study was approved by the Upper South A Regional Ethics Committee, Canterbury, New Zealand, and all participants gave informed consent.

2.2. Off-road assessment

2.2.1. Sensory–motor and cognitive tests (SMCTests™)

A subset from the SMCTests battery was used, measuring reaction time, ballistic movement, visuomotor tracking, visual search, complex attention, divided attention, and plan-
Fig. 1. Four screen shots of SMCTests tests as they appeared to participants. (a) Divided Attention, which incorporates both the Arrows Perception and Random Tracking tests; (b) Visual Search; (c) Complex Attention; and (d) Planning.

ning (Innes et al., 2009b). Test stimuli were presented on a computer monitor and responses recorded using a system that included a steering wheel, direction indicators, and foot pedals. Detailed specifications of the SMCTests battery are available from the User Manual (Christchurch Neurotechnology Research Programme, 2006), available online: www.neurotech.org.nz/files/CanDAT_SMCTests_User_Manual.pdf.

2.2.1. Sensory–motor function tests. The Footbrake and Clutch test records reaction and movement times for releasing the accelerator pedal and depressing brake and clutch pedals in response to green and red-light stimuli. Ballistic Movement records reaction time, movement time and peak velocity when rapidly moving the steering wheel following a visually presented cue. Sine Tracking and Random Tracking measure visuomotor coordination by recording the mean absolute error in mm of the tip of a vertically pointing arrow relative to a target when participants track 2D sinusoidal and random targets (with 8-s previews), respectively, using the steering wheel (Fig. 1a shows the central line and vertical arrow stimuli used for Random Tracking) (Jones, 2006). Sine Tracking and Random Tracking are performed twice each and alternated with one another: Sine Tracking trial 1, Random Tracking trial 1, Sine Tracking trial 2, Random Tracking trial 2.

2.2.1.2. Higher cognitive function tests. The Arrows Perception test requires participants to verbally respond whether four simultaneously presented horizontal arrows presented on screen are pointing in the same or different directions (Fig. 1a; four arrows on periphery of the screen) with reaction time and number of correct responses recorded. Arrows Perception gathers information on visual search speed and decision-making ability. Divided Attention consists of concurrent testing of the Arrows Perception and Random Tracking tests with participants asked to verbally respond regarding arrow directions and follow the random line target using the steering wheel (Fig. 1a). Divided Attention measures how well participants are able to concentrate on two competing activities using two separate response types (physical tracking of the target line using a steering wheel, and verbal responses regarding the direction the arrows are pointing). Visual Search requires participants to detect a left- or right-turning arrow target from an array of 70 road signs stimuli and to rotate the steering wheel in the direction the arrow is pointing (Fig. 1b), with mean response time and the number of correct responses recorded. Visual Search is sensitive to visual scanning speed and also to decision-making in regard to which way the steering wheel is turned. Complex Attention requires participants to move an arrow out of a box and across the screen using the steering wheel as quickly as possible following changing green-light stimuli (Fig. 1c). Recorded measures are reaction and movement times, and lapses (when the arrow was not moved out of the box following the stimulus change) and invalid trials (when the arrow was not within the box when the stimulus changed). Complex Attention requires that participants focus on relevant cues, discount irrelevant cues, and is sensitive to lapses in attention which are detected in the invalid and lapse response errors. Planning involves the presentation of a driving scene in plan view with participants instructed to ‘drive’ the car along a road using the steering wheel, accelerator and brake pedals (Fig. 1d). Obstacles to be negotiated include curves in the road, paint hazards, and intersections. Measures include number of paint hazards hit, number of collisions with other cars, safety margins between cars while crossing intersections, and number and duration of road position errors (including driving off the road). Planning can best be thought of as a complex task which requires the use of three types of apparatus (steering wheel, indicator stalk, and accelerator and brake pedals) to follow a series of rules in a unique environment (e.g., stopping at intersections, indicating appropriately while overtaking obstacles on the road, all while avoiding collisions with other vehicles). This test is not intended to be a simulation of actual driving.
2.2.2. Demographic measures and road knowledge

Participants provided information regarding years of education and driving, and completed a modified version of the Road Sign test (Land Transport Safety Authority, 2002) which requires the participant to identify six different road signs and state the appropriate action a driver should take for each.

2.2.3. Standardized psychometric and personality tests

Neuropsychiatric status was assessed with the 30-item Geriatric Depression Scale (GDS) (Aging Clinical Research Center, n.d.) and the Beck Anxiety Inventory (Beck, 1990). The 14-item Driving Anger Scale was used to measure propensity to become angry in driving situations (Deffenbacher et al., 1994) and was administered twice—once by each participant prior to the first assessment, and the second time at the first assessment appointment in the presence of the examiner. We could not find a sensation-seeking scale with questions we considered appropriate for older people. Instead we used the 44-item Big Five Inventory, which has not been previously used in samples of older adults in relation to driving. In this case, the inclusion of the Big Five Inventory was exploratory in nature. The inventory measures five personality dimensions: extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience (John and Srivastava, 1999).

2.2.4. Standardized cognitive tests

The Wechsler Test of Adult Reading (WTAR) (Wechsler, 2001) was used to estimate IQ. A standardized version of the Mini-Mental State Examination (MMSE) (Molloy and Standish, 1997) and the Dementia Rating Scale-2 (DRS-2) (Mattis et al., 2001) established global cognitive status. Trail Making Tests A and B (TMT A and TMT B) (Brainmetric, n.d.) were used to assess visual scanning and mental flexibility.

2.2.5. Driving frequency

Participants kept a log of odometer readings before and after each driving trip for a one-week period prior to the off-road testing session. Details of longer driving excursions over the previous 12 months were elicited at the first assessment appointment and kilometres travelled calculated using tables of travel distances and Google Maps (http://maps.google.co.nz/). These were combined with extrapolated driving log records to form an estimate of driving exposure over the previous 12 months.

2.3. On-road driving assessment

On-road assessments were conducted by an experienced driving occupational therapist and a driving instructor both from the Driving and Vehicle Assessment Service at Burwood Hospital, Christchurch. On-road assessors were blind to the results of all off-road testing. Participants were able to use their own cars (automatic or manual) for the driving assessment, as older drivers are more likely to Pass on-road driving assessment if they use their own car (Lundberg and Hakamies-Blomqvist, 2003). The driving instructor sat in the passenger seat, provided directions, and maintained safety of the vehicle while the occupational therapist sat in the rear and observed driving performance. All participants travelled the same 45-min public road route with an equal number of left and right turns. Road conditions included single-lane roundabouts, dual-lane roundabouts, dual-lane roads, controlled intersections (yield and stop signs, and traffic light controlled), uncontrolled intersections, and changes in speed zone (i.e., 50 km/h, 60 km/h, and 80 km/h sections). Driving ability was rated as a consensus Pass or Fail score. A driving scale score was then assigned by the occupational therapist using an 11-item ordinal driving scale where scores of 0–5 could be given to those in the Fail range and scores 6–10 given to those in the Pass range (Innes et al., 2007). This scale was designed to give a continuous measure of how well a person performed in the on-road assessment. Driving scale scores were assigned by the number of observed driving errors, whether these were considered major or minor, and whether the participant was able to correct errors once they were pointed out.

2.4. Data analysis

2.4.1. Binary logistic regression

Binary logistic regression is a non-parametric statistic used when the dependent variable is dichotomous—in this case Pass or Fail on the on-road assessment. BLR takes a number of entered variables and builds a parsimonious equation that explains how the variables relate to the dependent dichotomous 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. In essence, the model decides whether a variable is useful for describing the dependent variable and, in stepwise and elimination procedures, the model expels variables that do not explain a significant amount of variance in the dependent measure. As regression models can become overly fitted to the sample data, to minimize over-fitting we only offered variables to the model that were related to the on-road driving assessment Pass or Fail group outcome. Variables had to fulfill at least one of the following criteria: (1) a significant \( p \leq .05 \) difference in scores between Pass and Fail groups as evidenced by Mann–Whitney U tests for non-normally distributed data and t-tests for normally distributed data, (2) a significant Spearman correlation between a test measure and scores on the 0–10 Driving Scale score, or (3) a receiver operating characteristic (ROC) for predicting Pass and Fail scores with an area under the curve (AUC) of .60 or higher. If two selected variables correlated at 0.8 or higher, we would only offer the variable most highly related to the dependent variable to the model.

2.4.2. Non-linear causal resource analysis

Non-linear causal resource analysis is an approach based on the resource economic performance modelling constructs of general systems performance theory and the elemental resource model which state that a suboptimal amount of a necessary resource applied to a task will result in substandard performance regardless of the utilization of other relevant resources (Kondraske, 2006). Using NCRA, the minimum resource required to achieve a certain level of performance on a high-level task (the dependent variable) is plotted for each variable as a resource demand function (see Fig. 2 as an example). Each participant’s final predicted score is assigned based on his or her single poorest test score compared to others in the sample; there is no weighting of test scores. Unlike regression models which produce an equation for assigning the dependent measure score, the output of NCRA consists of a resource demand function for each entered variable. In addition, the NCRA software program provides a predicted dependent score for each participant and identifies the measure which has limited the person’s predicted score. The NCRA model predicts a score on a continuous measure. For example, an outcome measure may have a possible range of 0–10 and may be further broken down into two groups, e.g., 0–5 equals a Fail score and 6–10 equals a Pass score. The NCRA model assigns the 0–10 score but has no knowledge of what constitutes a Pass or a Fail. In the context of driving assessment, the essence of NCRA analysis can be summarized by: if no-one in the training data with a foot reaction time of 800 ms had a score above 4 on the driving scale, a new person with a foot reaction time of 800 ms would also be predicted to have a driving score of no higher than 4 based upon that single measure, irrespective of whether they performed better on other tests.
will not predict low scores for future cases who receive similar low reaction speeds.

2.4.3. Leave-one-out cross-validation

The relatively small size of the study sample precluded prediction validation using an independent test data set. Instead, leave-one-out cross-validation was used to assess the stability of the model. Leave-one-out cross-validation consists of removing each case individually from the analysis, re-training the model on remaining participants, and then testing the prediction on the excluded case using the new model (Witten and Frank, 2000). The procedure is repeated for all cases and prediction rates averaged across all iterations. As classification models are by definition optimized to the specific characteristics of the study sample, it was expected that the accuracy of the BLR and NCRA classification models would be lower for prediction than for classification.

2.4.4. Choice of cut-points for reporting accuracy

Inspection of the ROC curve coordinates for each model allows for the selection of criterion cut-points for classifying Pass and Fail outcomes. Three different cut-points were inspected for each model. First used were the default cut-points for a Fail score of 0.5 for the BLR and <6 for NCRA. Another cut-point was chosen that represented the highest value of sensitivity and specificity when averaged together. For the third cut-point we wished to maximize the sensitivity of the model for classifying Fails, and chose a cut-point that allowed for a minimum sensitivity of 80%, meaning that 80% of the observed Fail group were correctly detected by the model. We were also interested in the negative predictive value of each cut-point. This value represents the proportion of participants predicted to Pass who were actual Passes and not misdiagnosed Fails. Avoiding misclassifying Fails as Passes is a prime goal of driver screening, therefore ideally we would like this value to be as high as possible, with a goal of at least 80%.

3. Results

Sixteen participants (27%) failed the on-road driving assessment (7 males, 9 females; Fisher's Exact Test, two-tailed p = .77). Tables 1 and 2 show descriptive statistics for Pass and Fail groups with associated effect sizes for the differences between these groups. Cronbach’s α was calculated for several ordinal scales in order to determine the internal consistency of the measures: Geriatric Depression Scale α = .79; Beck Anxiety Inventory α = .81; Driver Anger Scale first administration α = .91; Driver Anger Scale second administration α = .92; Big Five Inventory Extraversion subscale α = .80; Big Five Inventory Conscientiousness subscale α = .84; Big Five Inventory Neuroticism subscale α = .70; Big Five Inventory Openness to Experience subscale α = .73.

3.1. Binary logistic regression

Nine variables showed a relationship with either the Pass/Fail score or the 0–10 Driving Scale score. Five were SMCTests measures (Random Tracking error—trials 1 and 2, Sine Tracking error—trials 1 and 2, Complex Attention reaction time standard deviation) and two were cognitive measures (longer completion times on both TMT A and TMT B). The remaining two variables were the demographic measures of age grouping (with older age groups having increased numbers of on-road Fails) and occupation code, with those in the Fail group, on average, having jobs of lower socioeconomic status. Random Tracking errors on trials 1 and 2 were correlated (r = .87), so only Random Tracking error—trial 1 was entered as it had a higher ROC AUC. The remaining eight variables were entered into the BLR model using a backwards elimination procedure. The model accepted one SMCTests measure – Random
Tracking error—trial 1 – and one cognitive test measure—TMT B completion time, which accounted for 25% of the variance in the on-road outcome (Nagelkerke $R^2$). The ROC AUC for the BLR model was .76, which is higher than the AUC of .50 for no discrimination ($z = 3.48; p < .001$) (Fig. 3).

Using a criterion value of 0.5 the model correctly classified 46 out of 59 participants (78%) into on-road Pass or Fail groups. The sensitivity to classify fails was 25% (4/16 correctly classified), sensitivity to classify Passes indeed received a Pass on the on-road assessment. The cut-point for the highest average of sensitivity and specificity value (cut-point = 0.25) correctly classified 44 of 60 participants (73.3%) into on-road Pass or Fail groups with a sensitivity of 81.3%, specificity of 59.1%, and negative predictive value of 89.7%.

The culmination of the 60 iterations generated by leave-one-out cross-validation reduced accuracy at the 0.5 cut-point to 75.0% (45/60 correctly classified), sensitivity to 12.5% (2/16 correctly predicted as Fail), and increased specificity to 97.7% (1/44 incorrectly predicted as Fail). Applying leave-one-out results to the 0.25 cut-point reduced overall accuracy from 73.3% to 65.0%, sensitivity from 68.8% to 50.0%, and specificity from 75.0% to 70.5%. Using the more sensitive 0.17 cut-point dropped overall accuracy from 65.0% to 58.3% with a sensitivity drop of 81.3–68.8% and specificity drop of 59.1–54.5%.

All 60 iterations of the leave-one-out cross-validation contained Random Tracking error—trial 1 and all but two contained TMT B completion time, indicating a degree of stability for the relationship of these variables to on-road driving outcome in the sample.

### Table 1
Demographic, psychiatric, and personality measures in the Pass and Fail groups.

<table>
<thead>
<tr>
<th>Test measure</th>
<th>Pass group (n=44)</th>
<th>Fail group (n=16)</th>
<th>Cohen’s $d$</th>
<th>$p$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (1 = male, 2 = female)</td>
<td>Mean (SD, range)</td>
<td>Mean (SD, range)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years)</td>
<td>1.48 (0.51, 1–2)</td>
<td>1.56 (0.51, 1–2)</td>
<td>0.17*</td>
<td>.56</td>
</tr>
<tr>
<td>Age grouping</td>
<td>1.91 (0.80, 1–3)</td>
<td>2.25 (0.86, 1–3)</td>
<td>0.42*</td>
<td>.15</td>
</tr>
<tr>
<td>Visual acuity left eye</td>
<td>10.16 (8.79, 4–60)</td>
<td>8.88 (3.42, 5–18)</td>
<td>0.04</td>
<td>.89</td>
</tr>
<tr>
<td>Visual acuity right eye</td>
<td>8.50 (4.22, 4–18)</td>
<td>8.94 (4.70, 5–24)</td>
<td>0.20</td>
<td>.51</td>
</tr>
<tr>
<td>Handedness (1 = right, 2 = left)</td>
<td>1.09 (0.29, 1–2)</td>
<td>1.00 (0, 1–1)</td>
<td>0.44*</td>
<td>.22</td>
</tr>
<tr>
<td>Years of education</td>
<td>13.18 (3.22, 8–19)</td>
<td>13.56 (2.71, 9–18)</td>
<td>0.14*</td>
<td>.65</td>
</tr>
<tr>
<td>Occupation code</td>
<td>2.73 (1.65, 1–7)</td>
<td>3.38 (2.00, 1–8)</td>
<td>0.38*</td>
<td>.20</td>
</tr>
<tr>
<td>Distance driven in the past year (1000 km)</td>
<td>7.31* (18.03, 2.18–122.57)</td>
<td>7.18* (6.02, 0.62–27.03)</td>
<td>0.27*</td>
<td>.34</td>
</tr>
<tr>
<td>Years of driving</td>
<td>54.98 (7.78, 31–69)</td>
<td>55.56 (6.61, 40–66)</td>
<td>0.08</td>
<td>.79</td>
</tr>
<tr>
<td>Geriatric Depression Scale</td>
<td>3.66 (3.43, 0–14)</td>
<td>4.38 (4.18, 0–12)</td>
<td>0.12*</td>
<td>.64</td>
</tr>
<tr>
<td>Beck Anxiety Inventory</td>
<td>3.86 (4.07, 0–15)</td>
<td>5.00 (5.21, 0–21)</td>
<td>0.25*</td>
<td>.40</td>
</tr>
<tr>
<td>Driving Anger Scale, 1st administration</td>
<td>33.16 (9.58, 15–53)</td>
<td>31.75 (8.85, 15–44)</td>
<td>0.15</td>
<td>.61</td>
</tr>
<tr>
<td>Driving Anger Scale, 2nd administration</td>
<td>32.32 (11.13, 16–61)</td>
<td>28.25 (7.36, 15–44)</td>
<td>0.40*</td>
<td>.20</td>
</tr>
<tr>
<td>Big Five personality factors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extraversion</td>
<td>26.61 (5.53, 16–39)</td>
<td>23.50 (5.94, 11–35)</td>
<td>0.55</td>
<td>.06</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>38.43 (3.80, 30–45)</td>
<td>36.75 (4.91, 25–44)</td>
<td>0.40</td>
<td>.17</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>37.30 (5.01, 28–45)</td>
<td>34.81 (7.13, 22–45)</td>
<td>0.45</td>
<td>.14</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>18.52 (5.41, 10–32)</td>
<td>18.13 (3.48, 12–24)</td>
<td>0.08</td>
<td>.74</td>
</tr>
<tr>
<td>Openness to experience</td>
<td>35.18 (5.44, 24–47)</td>
<td>35.63 (6.60, 25–48)</td>
<td>0.08</td>
<td>.79</td>
</tr>
</tbody>
</table>

* Cohen’s effect size for rank-transformed variables (Hopkins, 2004).

* 1 = 70–74 years, 2 = 75–79 years, 3 = 80 plus years.

* Scoring is expressed in metric with the number listed being the denominator in the fraction, e.g. 6/6 would be expressed as ‘6’ and is equivalent to 20/20 vision as measured in feet.

* 1 = managers to 8 = labourers.

* Median score presented due to highly skewed distribution.

3.2. Non-linear causal resource analysis

Fifty of 75 variables fulfilled criteria to be entered into the NCRA model. Tests and measures which limited the scores of the 16 drivers who failed the on-road assessment were km per year, visual acuity, TMT A, TMT B, Dementia Rating Scale-2, Beck Anxiety Inventory, Random Tracking trial 1, Sine Tracking error 1, Complex Attention, Divided Attention, Arrows Perception, Ballistic Movement, and Planning. The MMSE and SMCTests Footbrake and Clutch were the only tests that did not limit any of the 60 participants to either Fail or Pass scores. The resource demand curve for the MMSE was flat, meaning that variation in the score was not related on to the Driving Scale score. The lowest scores for Footbrake and Clutch measures were from participants who Passed the on-road assessment, thus no person could be limited to a predicted Fail score. The ROC AUC of .88 was higher than for no discrimination ($z = 8.96; p < .001$) (Fig. 3). There was no significant difference between the ROC AUC of the NCRA and BLR models ($z = 1.488; p = .137$).

As with the BLR model, three cut-points from the ROC curve were inspected. The first cut-point was based on those predicted.
Table 2
Performance-based tests including cognitive tests and SMCTests measures in the Pass and Fail groups.

<table>
<thead>
<tr>
<th>Test measure</th>
<th>Mean (SD, range)</th>
<th>Fail group (n = 16)</th>
<th>Cohen’s d</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Road Sign test, number correct</td>
<td>11.09 (1.18, 7–12)</td>
<td>10.94 (1.00, 10–12)</td>
<td>0.23</td>
<td>.43</td>
</tr>
<tr>
<td>Mini-Mental State Examination</td>
<td>28.80 (1.00, 27–30)</td>
<td>28.75 (0.96, 28–30)</td>
<td>0.11</td>
<td>.70</td>
</tr>
<tr>
<td>TMT A (s)</td>
<td>33.34 (10.09, 19–59)</td>
<td>40.44 (13.22, 21–69)</td>
<td>0.58</td>
<td>.06</td>
</tr>
<tr>
<td>TMT B (s)</td>
<td>91.00 (38.32, 29–254)</td>
<td>121.31 (54.24, 61–271)</td>
<td>0.80</td>
<td>.01</td>
</tr>
<tr>
<td>Wechsler Test of Adult Reading, estimated IQ</td>
<td>110.23 (10.09, 79–125)</td>
<td>109.06 (9.91, 91–121)</td>
<td>0.15</td>
<td>.61</td>
</tr>
<tr>
<td>Dementia Rating Scale-2, AEMSSb</td>
<td>11.11 (2.69, 5–17)</td>
<td>10.13 (2.50, 5–13)</td>
<td>0.37</td>
<td>.20</td>
</tr>
</tbody>
</table>

Footbrake and Clutch test
- Reaction time (ms) 299.68 (43.09, 238–447) 299.50 (39.99, 249–388) 0.29 .92
- Movement time (ms) 301.11 (75.36, 175–509) 298.19 (60.93, 200–438) 0.02 .95

Ballistic Movement test
- Reaction time, grand mean (ms) 354.03 (51.22, 275–497) 357.65 (50.16, 298–475) 0.07 .80
- Movement time, grand mean (ms) 239.99 (57.63, 153–434) 232.85 (52.98, 181–378) 0.16 .59
- Peak velocity, grand mean (ms) 944.27 (181.24, 603–1355) 953.61 (164.28, 716–1265) 0.05 .86

Tracking test
- Sine Tracking trial 1, error (mm) 15.24 (6.18, 6.74–31.29) 19.36 (9.13, 10.13–42.15) 0.46 .12
- Sine Tracking trial 2, error (mm) 9.65 (5.17, 3.99–31.63) 11.77 (6.04, 6.05–27.57) 0.52 .09
- Random Tracking trial 1, error (mm) 8.57 (3.40, 3.94–17.39) 12.52 (6.90, 5.68–31.93) 0.71 .02
- Random Tracking trial 2, error (mm) 8.17 (3.30, 2.79–16.79) 10.00 (5.54, 4.00–26.66) 0.50 .10

Arrows test
- Number correct (out of 12) 11.73 (0.50, 10–12) 11.38 (0.81, 9–12) 0.53 .06

Divided Attention test
- Tracking error (mm) 8.92 (2.21, 5.49–15.38) 9.85 (3.57, 6.44–20.94) 0.28 .34
- Arrows correct (out of 12) 11.50 (0.73, 9–12) 11.38 (0.96, 9–12) 0.05 .85

Visual Search
- Reaction time (s) 4.76 (0.78, 3.10–6.80) 4.95 (0.59, 4.00–6.00) 0.26 .39
- Number correct (out of 20) 15.48 (2.39, 9–20) 14.75 (2.49, 10–18) 0.30 .31

Complex Attention test
- Reaction time (s) 435.34 (80.12, 298–625) 465.19 (108.25, 329–763) 0.26 .36
- Movement time (s) 303.66 (74.76, 190–471) 302.88 (81.27, 212–456) 0.07 .80
- Reaction time SD (s) 160.82 (126.82, 20–541) 229.88 (200.95, 26–663) 0.25 .38
- Movement time SD (s) 69.41 (89.70, 13–457) 63.50 (4.54, 6.44–20.94) 0.50 .10
- Number of lapse errors 0.01 (0.21, 0–1) 0.06 (0.21, 0–1) 0.07 .79
- Number of invalid trials 0.23 (0.61, 0–3) 0.15 (0.0, 0) 0.61 .09

Planning test
- Number of hazards hit 2.50 (1.41, 0–5) 2.49 (1.26, 1–5) 0.07 .82
- Number of crashes 1.02 (1.25, 0–5) 1.25 (1.24, 0–4) 0.23 .43
- Duration of positional faults (s) 6.16 (16.40, 0–19.20) 6.48 (4.89, 2.50–22.20) 0.09 .99
- Intersection safety margin (mm) 40.59 (13.05, 13–63) 35.56 (14.74, 4–59) 0.30 .31
- Lateral road position error (mm) 2.68 (0.30, 2.00–3.30) 2.73 (0.37, 2.00–3.30) 0.16 .57
- Distance travelled (m) 4.77 (0.46, 3.00–5.20) 4.60 (0.62, 2.70–5.20) 0.29 .32

\* Cohen's effect size for rank-transformed variables (Hopkins, 2004).
\b AEMMS, age and education-adjusted MOANS scaled score.

as a Driving Scale score of <6 predicted as a Fail. The model using this cut-point correctly classified 52 out of 60 participants (86.7%) into Pass and Fail groups with a sensitivity for classifying Fails of 75.0% (12/16 correctly classified as Fail), specificity of 90.9% (4/44 incorrectly classified as Fail), and negative predictive value of 87.1%. The cut-point for the highest average of sensitivity and specificity values (cut-point = 5.7) correctly classified 90.0% of participants into Pass and Fail groups with a sensitivity of 75.0%, specificity of 95.5%, and negative predictive value of 91.3%. The cut-point with a minimum sensitivity of 80% (cut-point = 6.01) correctly classified 66.7% of the participants into Pass and Fail groups with a sensitivity of 81.3%, specificity of 61.4%, and negative predictive value of 90.0%.

The combined results of the 60 iterations generated by leave-one-out cross-validation reduced overall accuracy of the cut-point for the highest average of sensitivity and specificity values to 90.0% of the sample. The criterion point that would Pass (44 passed, 73.3% of the sample). The rate that would have been achieved by predicting that every driver would Pass (44 passed, 73.3% of the sample). The criterion point of the BLR model can be shifted in order find the most appropriate balance of sensitivity or specificity and will depend on factors such as whether the test would be used as a screen to detect people more likely to Fail. The two other cut-points examined showed predictable trade-offs between sensitivity and specificity. To be used in a practical setting, considerations over the appropriate cut-point to use would depend on factors such as the cost of more comprehensive driving assessment, and the percentage of Passes that would...
initially be flagged for further, unnecessary testing. It is clear that
the sensitivity and specificity of the BLR model at the three differ-
ent cut-points are not high enough for the model to be used as the
sole determinant of driving ability.

One of the tests selected by the BLR model, Random Tracking,
measures visuomotor planning and execution, with lower accu-

racy scores related to an increased likelihood of an on-road Fail
outcome. Random Tracking trial 1 is performed after Sine Track-
ing trial 1, which many people find difficult initially. This is usually
resolved by the end of the trial. Thus, Random Tracking trial 1’s
ability to classify driving ability may reflect either difficulties with
visuomotor control or with delayed learning of the tracking task
that extends past the first tracking trial. The other test selected by
the BLR model, TMT B, consists of visual scanning, sequencing, and
task-switching, with greater time to completion associated with a
Fail score. Lower scores on TMT B could indicate the presence of
undetected cognitive impairment in the group.

NCRA utilized 13 tests to correctly classify 86.7% of the sample
at a cut-point of <6 into on-road Pass or Fail groups. NCRA allows
any single test in the battery to limit a person’s predicted perfor-
mance. A test which limits no-one to a Fail score in the classification
set could limit a person to a Fail score given a new sample, which
means the significance of the 13 tests used in the NCRA classifi-
cation should not be interpreted in the same way as the two tests
found useful in the BLR model. Shifting the cut-point for detecting
Fails to the highest average balance of sensitivity and specificity
produced a sensitivity of 75.0% and specificity of 95.5%, and a cut-
point for a higher sensitivity of 81.3% had a specificity of 61.4%. As
with the BLR model, a large drop in specificity occurred when the
sensitivity was raised to capture at least 80% of those who failed
the assessment. Once again, the NCRA model alone does not have
high enough levels of both sensitivity and specificity in order to be
used as the sole determinant of driving ability.

As expected, the accuracy of both the BLR and NCRA models
were reduced following leave-one-out cross-validation, and this
held for all three cut-points for both the BLR and NCRA models.
The BLR model suffered primarily in sensitivity for predicting Fails
with specificity less affected. The NCRA model showed both a drop
in sensitivity as well as large reductions in specificity. These drops
in specificity occur because removing a participant from the model
leads to redrawing of resource demand curves in which that per-
son’s score made up part of the curve’s lower boundary. This can
lead to a Fail classification when the person’s score is tested against
redrawn resource demand curves (see Fig. 4 for an explanation of
this). For this reason the NCRA model appears substantially less sta-
bility than the BLR, although this only became apparent following the
leave-one-out cross-validation. This instability may be addressed
by using a larger training set but it would be necessary to devise
a method for eliminating low outliers to avoid the accumulation of
spurious flat resource demand curves which then offer no dis-


Fig. 4. Example of the effects on a resource demand function curve when a boundary
score is removed. The solid line shows a hypothetical resource demand curve drawn
to fit the distribution of scores for a variable. The circled data point shows a subject
with an observed test measure score of 8 and a Driving Scale score of 6 (on-road
Pass). When the circled data point is withdrawn during the leave-one-out analysis
the demand function curve is redrawn as indicated by the dashed line. When the
circled data point is re-entered as a test case against the new model, an observed
test measure score of 8 now aligns with a Driving Scale score of 4. This participant’s
score has now gone from 6 (Pass) to 4 (Fail).

influence of more general personality factors but no differences
were found between Pass and Fail groups for the BLR model, and
scores on the Big Five Inventory were not suitable for entry into the
NCRA model. Knowledge of road rules was not comprehensively
investigated, although knowledge of road signs and their related
driving actions (Road Sign test) did not discriminate between
groups. Although age differences are often found in those who Pass
versus those who Fail an on-road assessment, the current study
had a restricted age range of 70 and above which would have con-
tributed to no significant differences being found. If the lower age
bound was extended by a decade or more, average age differences
between Pass and Fail groups would be expected to be found.

Beyond the outcomes of classification and estimated prediction
rates, practical considerations influence whether a BLR or NCRA
model could be used in an applied setting. BLR is designed to choose
a small group of tests which best explain a binary outcome. In the
current study only two variables entered the BLR model, meaning
the assessment could be completed in 15 min. Assessment using
NCRA requires the inclusion of all tests since any test could poten-
tially limit the score of a future case. In the current study this would
stretch assessment time to 2.5–3 h and require several neuropsy-
chological tests that require administration training and substantial
hand-scoring. NCRA may allow for a more sensitive prediction than
BLR as it allows for many more tests to be available for the model
without the disadvantages of over-fitting that occur when numer-
ous measures are forced into a regression model. Tests that are
not useful for classification in NCRA are simply not used by the
model, and the addition of more tests with relationships to driving
can only improve classification. Any higher sensitivity afforded by
NCRA, however, requires a substantially greater effort to achieve
and we are currently investigating how NCRA works with a larger
sample.

Comparing classification performances between studies is diffi-
cult, as samples, independent and dependent variables, Pass and
Fail rates and utilized cut-points all affect the accuracies achieved.
However, comparing gross outcomes with studies that have used
similar methods may suggest whether the current procedures
have advantages over those used in previous studies. The variance
accounted for by the BLR classification model (25%) is lower than in two other studies incorporating neuropsychological tests (64% and 44% respectively) (De Raedt and Ponjaert-Kristoffersen, 2000; Stav et al., 2008). However, these studies differed in that they included participants at higher risk for unsafe driving: people referred for assessment following one or more crashes (De Raedt and Ponjaert-Kristoffersen, 2000) and people scoring below 24 on the MMSE (Stav et al., 2008). Variance cannot be accounted for in NCRA but the sensitivity of this model for detecting fails is near the 83–97% range found by previous authors who have used non-linear techniques to classify on-road driving ability in clinical populations (Innes et al., 2007; Risser et al., 2008; Sommer et al., 2008).

A limitation of the study is the sample size of 60 participants (achieved power of 80% for an effect size of $d = 0.83$ with $n = 16$ on-road Fail and $n = 44$ on-road Pass outcomes). In Tables 1 and 2, effects in the moderate range did not approach a priori defined levels of significance to be entered into the BLR model. As the study sample was not recruited from a strictly representative population of older drivers, and it was voluntary to take part, we cannot say for certain that the results would be replicated in a general population sample. For example, people who experience heightened anxiety related to driving may have been less likely to volunteer for the study and, thus, the sample could be biased to more confident drivers. Another limitation relates to the on-road driving assessment used in the study, as it has not been investigated for reliability or validity, although this limitation is far from unique. Korner-Bitensky et al. (2006) surveyed the driving assessment methods of 144 North American and Canadian driving assessors. Ninety-four percent of respondents routinely used on-road assessments as part of their evaluation, 24% used a standardized scoring system, and only 10% used a pre-defined cutoff score to define driving competency. Only two respondents reported using a standardized road test. Standardized on-road assessments do exist, such as the Driving Performance Evaluation and the Washington University Road Test. Some standardized assessments have been tested for inter-rater and test–retest reliability, with the former usually found to be moderate to high, and the latter in the moderate range (Fitten et al., 1995; Hagge, 1994; Hunt et al., 1997; Janke and Eberhard, 1998; Romanowicz and Hagge, 1995). Investigations into the validity of standardized road tests have found some associations to real-world crashes or infringements (Fitten et al., 1995; Keall and Frith, 2004; Romanowicz and Hagge, 1995), although due to the low base rates of crashes in particular, power is low for detecting statistically significant associations. Other methods to test on-road assessment validity have been based on finding differences in group performance in expected directions, such as differences in error scores or Pass and Fail results between novice and experienced drivers (Hagge, 1994; Romanowicz and Hagge, 1995). Measuring validity against prospective real-world negative driving outcomes could be ideal but there are ethical problems in allowing persons considered to be unsafe to continue driving in order to assess whether they have increased rates of real-world accidents and offences compared to those who Pass an on-road assessment. Owsey et al. (1991) have suggested that at-fault crashes would be a more useful criterion for determining driving safety compared to on-road assessments. Although this measure would be useful to include in research studies, the low base-rate of crashes even in people with demonstrable impairments would make history of recent crashes a poor measure for determining the driving safety of an individual in clinical settings. There is a certain amount of natural justice in allowing a person to demonstrate driving ability through the task of driving. Driving assessments of individuals require a thorough assessment which should incorporate demonstrated risk factors in the decision-making task. Accumulated risk factors alone would likely not be accepted as sufficient for decisions regarding driver safety. In the current study, all participants maintained their driver’s licences which provides a unique opportunity to follow the group prospectively to investigate the relationship between on-road Pass or Fail status and real-word crashes and traffic infringements, of which preliminary results are available (Hoggarth et al., 2009).

This study supports prior findings that non-linear methods can be at least comparable to traditional approaches for prediction and understanding of driving behaviour, although the low rates of sensitivity for predicting Fails in the current study may suggest that sensory–motor and cognitive measures of impairment may not be suited to a population with few, if any, measurable risk factors for unsafe driving. There are several areas in the driving research literature the authors believe require attention. Researchers are encouraged to compare the utility of linear and non-linear techniques for the classification and prediction of driving ability, to validate on-road assessment outcomes against recorded and/or self-reported crashes and driving offences, and to test classification models on independent samples, or at least to use statistical modelling techniques such as boot strapping or leave-one-out cross-validation to investigate the stability and generalizability of models.

Conflicts of interest statement

The authors believe there are no actual or potential conflicts of interest present that would lead to biasing of data collection, interpretation, or presentation.

Acknowledgements

The authors wish to thank driving instructor Neville Holmes for administering the on-road driving assessments. The authors thank Professor George Kondraske for providing the NCRA software used in this project.

Role of the funding source: Funding for this project was received from the Transport Research and Educational Trust Board, Accident Compensation Corporation, Christchurch Neurotechnology Research Programme, and University of Canterbury. None of the sponsors of this research played any role in the design or conduction of the study, collection, management, analysis, or interpretation of the data, or preparation or approval of the manuscript for submission. We have no other disclosures of financial interest.

References


