Accident Analysis and Prevention 77 (2015) 29–34

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

# Prediction of driving ability: Are we building valid models?

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

[a] New Zealand Brain Research Institute, Christchurch, New Zealand
[b] Psychiatric Service for the Elderly, The Princess Margaret Hospital, Christchurch, New Zealand
[c] Department of Medical Physics and Bioengineering, Christchurch Hospital, Christchurch, New Zealand
[d] Department of Electrical and Computer Engineering, University of Canterbury, Christchurch, New Zealand
[e] Department of Psychology, University of Canterbury, Christchurch, New Zealand
[f] Department of Medicine, University of Otago, Christchurch, New Zealand

ARTICLE INFO

*Article history:*
Received 17 May 2014
Received in revised form 22 December 2014
Accepted 19 January 2015
Available online 6 February 2015

*Keywords:*
Modeling
Regression
Over-fitting
Driving
Prediction

ABSTRACT

The prediction of on-road driving ability using off-road measures is a key aim in driving research. The primary goal in most classification models is to determine a small number of off-road variables that predict driving ability with high accuracy. Unfortunately, classification models are often over-fitted to the study sample, leading to inflation of predictive accuracy, poor generalization to the relevant population and, thus, poor validity. Many driving studies do not report sufficient details to determine the risk of model over-fitting and few report any validation technique, which is critical to test the generalizability of a model. After reviewing the literature, we generated a model using a moderately large sample size ($n = 279$) employing best practice techniques in the context of regression modelling. By then randomly selecting progressively smaller sample sizes we show that a low ratio of participants to independent variables can result in over-fitted models and spurious conclusions regarding model accuracy. We conclude that more stable models can be constructed by following a few guidelines.



## 1. Introduction

A common goal in driving research is to find measures that can be performed in an office setting that accurately predict on-road driving ability, often in terms of a pass/fail dichotomy. Many driving researchers use regression or discriminant analysis models to determine weighted measures to optimize these predictions. These models often have problems due to post-hoc variable selection, over-fitting by the inclusion of too many variables and, perhaps most importantly, by failing to test the generalizability of the classification model to the target population by testing the model against a new sample or by using of resampling techniques such as bootstrapping, *n*-fold, or leave-*n*-out cross-validation approaches.

Model building serves two related purposes. The term "classification" refers to the construction of a model that describes the characteristics of a sample, e.g. categorizing participants into on-road pass and fail groups. The term "prediction" refers to an ideal end point concerning reliable statements about the target population under study, represented by the sample that was recruited. In driving research, the term prediction has often been incorrectly used to refer to classification, which blurs the distinction between these two concepts. Over-fitting occurs when the model is fitted so closely to idiosyncrasies of the sample that it does not generalize well to the population. A number of steps can be taken during model construction to reduce the impact of over-fitting as discussed below.

### 1.1. Variable selection and ratio of participants to independent variables

Babyak (2004) states that selecting variables for a model on the basis of the strength of univariate association with the dependent variable within the sample data set is a common error that can result in over-fitting a model. That is, variables are offered to a model on the basis of a statistically significant association or large effect-size relationship with the dependent variable. Instead, variable selection should be based on *a-priori* reasoning or the

* Corresponding author at: Psychiatric Service for the Elderly, The Princess Margaret Hospital, Cashmere, Christchurch 8022, New Zealand. Tel.: +64 3 3377 997x66218.
** Corresponding author.
*E-mail address:* petra.hoggarth@cdhb.govt.nz (P.A. Hoggarth).

http://dx.doi.org/10.1016/j.aap.2015.01.013
0001-4575/

**Table 1**
Literature review for modelling techniques employed in studies assessing on-road outcome with a primarily cognitively impaired older adult sample.

| Study | *n* | Collinearity assessment | Variable selection method | Participant to variable ratio offered to the model | Automated variable selection method | Type of model validation | Model type | Classification accuracy | Predictive accuracy[a] |
|---|---|---|---|---|---|---|---|---|---|
| Bowers et al. (2013) | 32 | Not reported | Variables moved in and out depending on how they improved the fit of the model to the sample | 8:1 | Not reported | Not validated | BLR | Not reported (Sensitivity 95%, specificity 80%) | N/A |
| Carr et al. (2011) | 99 | Not reported | At least two different models were formed | Not reported | Stepwise–not further defined | Not validated | BLR | "Up to 85%" depending on cut-point | N/A |
| Hoggarth et al. (2013) | 279 | Yes – multicollinearity | All variables included if meeting multicollinearity threshold | 13:1 | Backward stepwise | Leave-one-out cross-validation | BLR | 76% | 72% |
| Innes et al. (2007) | 50 | Not reported | Unable to determine | Not reported | Forward stepwise | Leave-one-out cross-validation | BLR | 94% | 86% |
| | | | | | | | NCRA | 90% | 76% |
| Innes et al. (2011) | 501 | Yes–collinearity and multicollinearity | All variables included if meeting collinearity and multicollinearity threshold | 25:1 | Backward stepwise | Leave-one-out cross-validation | BLR | 78% | 76% |
| | | | | | | | SVM | 100% | 76% |
| | | | | | | | DA | 76% | 75% |
| | | | | | | | PK | 100% | 73% |
| | | | | | | | KP | 81% | 72% |
| Lincoln et al. (2006) | 37 | Not reported | All variables included unless excluded due to skew or kurtosis | Not reported | Not reported | A new sample of 17 participants | DA | 92% | 59% |
| Ott et al. (2013) | 75 | Not reported | Variables moved in and out depending on how they improved the fit of the model to the sample | Not reported | Backward stepwise | Bootstrapping | BLR | 71% | 57% |
| Snellgrove (2000) | 115 | Not reported | Three models were formed | 57.5:1 | Enter | Not validated | BLR | 77% | N/A |

BLR – Binary logistic regression, DA – Discriminant analysis, KP – Kernel product, NCRA – Nonlinear causal resource analysis, PK – Product kernel, SVM – Support Vector Machine.
[a] Predictive accuracy refers to the accuracy of the model following a validation procedure.

sample size should be large enough to accommodate a predetermined selection of variables to lessen the chances of over-fitting. A higher ratio of participants to variables decreases the risk of model over-fitting as it is less likely to be influenced by idiosyncrasies of the sample data (Babyak, 2004; Harrell, 2001; Tabachnick and Fidell, 2013). A minimum of 10 to 15 observations (e.g., participants) per predictor variable was suggested by Babyak (2004). Tabachnick and Fidell (2013) recommend a formula of $N \geq 50 + 8m$ where $m$ is the number of independent variables. For example, a study testing 10 independent variables would require recruitment of 130 participants. These authors suggest that more cases would be needed for expected small effect sizes or when the reliability of the measures is low.

### 1.2. Collinearity between independent variables

Highly correlated variables in a regression analysis leads to regression coefficients with a large standard error, thus making it difficult or impossible to determine the influence of individual variables (Garson, 2010; Tabachnick and Fidell, 2013). Thus, variables with high levels of collinearity should be excluded prior to model construction.

### 1.3. Automated statistical variable selection procedures

Automated selection procedures such as stepwise approaches can increase over-fitting (Babyak, 2004; Harrell, 2001; Tabachnick and Fidell, 2013). Harrell (2001) suggested that if stepwise approaches were introduced today they would be rejected due to violating "every principle of statistical estimation and hypothesis testing" (p. 56). Harrell (2001) also stated that, if stepwise approaches are used, then backwards is preferred to a forwards stepwise approach. Due to the risk of over-fitting, Tabachnick and Fidell (2013) suggested that when automated selection procedures are used the ratio of cases (participants) to independent variables should be 40 to 1, meaning a study with 10 independent variables should recruit 400 participants. Clearly this would be challenging for driving research as most studies use much smaller samples and a majority of driving studies use an automated stepwise approach (see Table 1). Babyak (2004) suggested that all variables should be entered and remain in the model rather than use any type of automated procedure.

### 1.4. Validation of the classification model

Perhaps the most important step of model building is to test the classification model. Such a test estimates the extent that a model can predict the outcomes of newly recruited cases not part of the original data set. One way to do this is to split the sample to build the model on one subset and test against the held back cases. Tabachnick and Fidell (2013) recommend this technique and suggest an 80/20 split. Conversely, Harrell (2001) suggests that splitting the sample results in lower precision and power. Instead he recommends statistical resampling validation techniques such as boot-strapping, jack-knifing, and leave-*n*-out cross-validation. Steyerberg et al. (2001) found that statistical resampling approaches provide a better estimate of a model's ability to generalize to a new sample than validation using a held-back sample.

### 1.5. Literature review

We reviewed the older-driver literature models of off-road testing used to classify on-road driving assessment outcomes (pass or fail) and assessed the implementation of the above strategies to reduce over-fitting. We focused on studies that primarily included

drivers with suspected or confirmed cognitive impairment, including dementia. Eight studies, including three from the current authors, were found (Bowers et al., 2013; Carr et al., 2011; Hoggarth et al., 2013; Innes et al., 2007, 2011; Lincoln et al., 2006; Ott et al., 2013; Snellgrove, 2000). These studies are summarized in Table 1.

Different methods were used to determine the entry of variables to the classification model. Three studies made more than one model by adding or removing variables to arrive upon the one with the most parsimonious fit (Bowers et al., 2013; Carr et al., 2011; Snellgrove, 2000). One study included all the variables following elimination of those with unacceptable levels of skew or kurtosis (Lincoln et al., 2006). Innes et al. (2007) did not provide information on how variables were selected for entry and Innes et al. (2011) and Hoggarth et al. (2013) included all variables that met a collinearity threshold.

The ratio of participants to variables varied considerably across the studies. It was not possible to determine this ratio for four studies (Carr et al., 2011; Innes et al., 2007; Lincoln et al., 2006; Ott et al., 2013). Of the remaining four studies one offered 57.5 participants per independent variable (Snellgrove, 2000), one offered 25 participants per variable (Innes et al., 2011), one offered 13 participants per variable (Hoggarth et al., 2013), and one offered 8 participants per variable (Bowers et al., 2013). Bowers et al. (2013) constructed several models, with one having a ratio of 4.6 participants per variable.

Two of the eight studies (Bowers et al., 2013; Lincoln et al., 2006) did not report the selection process of the variables used during model construction. One study (Snellgrove, 2000) used an enter approach while the remainder used some type of stepwise procedure (forward or backward). Five studies used a validation procedure to test their classification model and all showed a reduction in the overall accuracy of prediction compared to the initial classification model. Lincoln et al. (2006) recruited a new sample, Hoggarth et al. (2013) and Innes et al. (2007, 2011),) used leave-one-out cross-validation, and Ott et al. (2013) used bootstrapping.

While our literature review has focused on a cognitively-impaired sample, due to the importance of validating models we would also like to mention three articles that report results following model validation in non-cognitively-impaired samples. Wood et al. (2008) used a number of off-road measures to predict on-road driving performance in a sample of 270 non-cognitively impaired drivers aged 70 and over. They validated their model using leave-one-out cross-validation and also tested the model on a held-back sample. Hoggarth et al. (2010) used off-road measures to predict on-road driving ability in a cognitively-intact group of 60 drivers aged 70 and over. They validated their model using leave-one-out cross-validation. Risser et al. (2008) recruited a mixed-age sample of 222 people and constructed a neural network model to predict on-road driving performance. They validated their model using jack-knife, bootstrapping, and testing on a new sample.

Our review of driving studies of participants with cognitive impairment indicates that over-fitting is an under-recognized problem in the field. Many aspects of model design were not reported, thus making it difficult to estimate how likely it was that an individual model was over-fitted. The review also shows that many studies reported no validation procedure, which is necessary to determine how likely it is that the model will generalize to the population.

None of the statistical model building concepts reviewed above are original to this article. However, while this information has been available for some time, it appears to have been only minimally adopted by driving researchers. Thus, the purpose of this paper is to present these concepts in a pragmatic way in a familiar context to make them more accessible for driving researchers. As such, the paper does not seek to determine whether the independent measures reported in the data set are useful for classifying unsafe drivers. The data set used in this study is solely to show how a model can be built to be more robust to over-fitting. Details of the full study from whence this data originates, including the details of the off-road measures and a consideration of their usefulness for detecting unsafe driving, have been previously published (Hoggarth et al., 2013).

To demonstrate the use of model building techniques that reduce the influence of over-fitting a binary logistic regression model was built to determine the utility of sensory-motor and cognitive tests for predicting pass and fail outcomes on an on-road driving assessment in primarily older drivers with confirmed or likely cognitive impairment. First, recommendations listed in Sections 1.1–1.4 were followed to reduce the effects of over-fitting and to test the generalizability of a binary logistic regression model. Second, the effects of offering a high ratio of participants to independent variables was then tested using progressively smaller randomized subsets of sample participants while holding the number of independent variables offered constant. We expected that as the ratio of participants to variables increased then the accuracy of the classification models would increase due to over-fitting; at the same time, the accuracy following the validation procedure was expected to drop as an indication that the models would generalize progressively more poorly to the population.

## 2. Methods

### 2.1. Participants

Participants were 279 referrals (180 male, 99 female; mean age 78.4, range 56–92) to three driving assessment services in New Zealand that specialize in driving assessment for people with medical disorders that may affect driving safety. All participants had diagnosed or suspected Alzheimer's dementia, mild cognitive impairment, unspecified cognitive impairment, or cognitive problems, some associated with suspected or identified cerebral vascular changes including stroke.

### 2.2. Procedure

Participants completed a battery of computerized sensory-motor and cognitive tests which yielded 27 potential independent variables (see Hoggarth et al., 2013 for details). They then completed a medical driving assessment administered by a driving specialist occupational therapist (OT) who assigned each participant with a pass or fail outcome based on driving performance. The OT was blinded to the results of the off-road testing.

### 2.3. Data analysis

A binary logistic regression model was generated using the results of the computerized testing to classify pass/fail outcome on the on-road driving assessment. To minimise over-fitting as discussed in Sections 1.1–1.4 the following processes were employed:

1. Multicollinearity between independent variables was assessed using the 'Collinearity diagnostics' function in SPSS version 11.5.0 (SPSS, Inc., Chicago, IL). These statistics measure the degree of collinearity among all variables entered into the regression. These relationships are independent of the relationship of variables to the dependent variable and, therefore, do not contribute to selection bias by excluding variables that have a low strength relationship with the

**Table 2**
The sensitivities, specificities, and overall classification accuracies of the two models including leave-one-out cross-validation.

| | Enter | | Backward stepwise | |
|---|---|---|---|---|
| | Classification | Prediction (leave-one-out cross-validation) | Classification | Prediction (leave-one-out cross-validation) |
| Sensitivity (%) | 77.4 | 68.4 | 78.7 | 73.5 |
| Specificity (%) | 71.0 | 70.2 | 71.8 | 70.2 |
| Negative predictive value (%) | 71.5 | 64.0 | 73.0 | 67.8 |
| Positive predictive value (%) | 76.9 | 74.1 | 77.7 | 75.5 |
| Accuracy (%) | 74.6 | 69.2 | 75.6 | 72.0 |

independent variable. The lower the tolerance value reported in the table, the more correlated a measure is with one or more of the other variables. Variables with low tolerance values (<0.20) were deleted one at a time and the analysis repeated until all independent variables had tolerance values of >0.20.

2. Variable selection and the ratio ofparticipants to variables was addressed by having a large enough sample size that allowed for all measured variables to be entered without resorting to selection strategies (providing variables survived the multicollinearity procedure). We aimed for the minimum ratio suggested by Babyak (2004) of no more than 10participants per variable, which allowed a maximum of 27–28 variables to be entered into our model given the sample size of 279 participants. Using Tabachnick and Fidell's (2013) formula of $N \geq 50 + 8m$,the possible entry of 27 variables would require a sample size of 266, which is well within the number recruited.
3. Both a backwards stepwise and an enterapproach were used to fit the model. Creating the model both ways allowed the resulting models to be compared to identify any obvious differences between the two approaches in this sample.
4. The model was validated using leave-one-out cross-validation (Witten and Frank, 2000) using a script written in MATLAB Version 7.10.0.499 (R2010a, The MathWorks, Inc., Natick, MA). The process involved removing each case individually from the sample, re-training the model on the remaining participants, testing the prediction on the excluded case using the new model, then replacing the case. The procedure is repeated for all cases. In essence, it mimics what would happen if a case was not part of the training data set and, therefore, estimates how the model would perform given a new case from the same population (provided the case is representative of the population).

Secondly, in order to test the assumption that adding a higher ratio of participants to variables could lead to over-fitting, we constructed a series of models using a reducing number of participants of $n = 250$, 200, 150, 100, 90, 80, 70, 60, and 50. To reduce the influence of random differences in the samples, we constructed three randomized samples for each of the sample sizes and averaged the results of the three to generate reported accuracy statistics. Leave-one-out cross-validation was performed on each model and, again, averaged to give an estimate of model generalizability.

## 3. Results

155 of the 279 participants (55.5%) failed the on-road driving assessment. Five variables were deleted due to multicollinearity tolerance values <0.2. The remaining 22 variables were offered to the model in both a backwards stepwise and an enter approach, for a ratio of just under 13 participants per variable, which is greater than the minimums suggested by both Babyak (2004) and Tabachnick and Fidell (2013).

For the enter approach, the model retained all 22 variables to account for 39% of the variance in on-road outcome (Nagelkerke $R^2$). The ROC AUC for the model was .82 (95% CI: .78–.87). Using a default cut-point of 0.5, the model correctly classified 208 of 279 participants (74.6%) with sensitivity of 77.4% and specificity of 71% (Table 2). The averaging of the 279 iterations generated by leave-one-out cross-validation reduced the overall accuracy of the model to 69.2%, sensitivity to 68.4%, and specificity to 70.2%.

For the backwards stepwise approach, the model accepted eight measures of the 22 offered variables and these accounted for 36% of the variance in the on-road outcome (Nagelkerke $R^2$). The ROC area under the curve (AUC) measure for the model was .81 (95% CI: .76–.86). Using a default cut-point of 0.5, the model correctly classified 211 of 279 participants (75.6%) with sensitivity for detecting fails of 78.7%, and specificity of 71.8%. The averaging of the 279 iterations generated by leave-one-out cross-validation reduced the overall accuracy of the model to 72.0%, sensitivity to 73.5%, and specificity to 70.2%. See Table 2 for the classification model and the model following leave-one-out cross-validation. By visual inspection, the classification models for the enter and stepwise approaches appear similar. Following the leave-one-out procedure the enter model has overall accuracy and sensitivity estimates of a few percentage points lower than the stepwise model, although the specificity is identical.

Additional classification models were constructed for progressively smaller randomized samples of participants (Table 3). Each

**Table 3**
The classification and leave-one-out accuracies of progressively decreasing ratios of participants to variables for models that forced all variables into the model (enter approach) and those that used backwards stepwise elimination.

| | | Enter (i.e., no variable selection) | | Backward stepwise variable selection | | |
|---|---|---|---|---|---|---|
| Sample size[a] | Ratio of participants to variables | Mean accuracy of classification models (% ± SD) | Mean accuracy for leave-one-out cross-validation models (% ± SD) | Mean number of measures accepted into model (Post-hoc ratio of participants:variables) | Mean accuracy of classification models (% ± SD) | Mean accuracy for leave-one-out cross-validation models (% ± SD) |
| 250 | 11.4:1 | 75.5 ± 0.8 | 67.6 ± 1.4 | 8.3 (30:1) | 75.9 ± 1.5 | 69.7 ± 3.7 |
| 200 | 9.1:1 | 74.7 ± 2.3 | 68.3 ± 1.6 | 7.7 (26.0:1) | 72.5 ± 3.0 | 65.7 ± 2.6 |
| 150 | 6.8:1 | 74.2 ± 1.7 | 67.8 ± 4.1 | 6.7 (22:1) | 73.8 ± 1.7 | 70.4 ± 3.0 |
| 100 | 4.5:1 | 84.7 ± 1.5 | 69.3 ± 10.5 | 8.0 (13:1) | 79.3 ± 3.1 | 67.7 ± 13.1 |
| 90 | 4.1:1 | 84.5 ± 2.9 | 58.9 ± 5.9 | 7.7 (12:1) | 76.4 ± 3.8 | 66.7 ± 4.4 |
| 80 | 3.6:1 | 79.2 ± 5.9 | 59.6 ± 1.4 | 6.7 (12:1) | 78.8 ± 5.8 | 59.6 ± 6.4 |
| 70 | 3.2:1 | 91.8 ± 7.2 | 57.6 ± 3.0 | 5.7 (12:1) | 83.2 ± 3.4 | 58.6 ± 6.2 |
| 60 | 2.7:1 | 90.6 ± 4.2 | 60.0 ± 4.4 | 9.0 (7:1) | 86.7 ± 5.0 | 56.1 ± 11.3 |
| 50 | 2.3:1 | 97.3 ± 4.6 | 52.7 ± 9.2 | 10.3 (5:1) | 96.7 ± 5.8 | 55.3 ± 6.4 |

[a] Three random samples were constructed for each sample size. Results reported are the mean values for three samples.

model was offered the same 22 variables that were offered to the $n = 279$ model. The models were formed both with backwards stepwise and an enter all 22 variables approach.

For both the enter and backwards elimination models, the accuracy of the classification models increased as the ratio of participants to variables decreased. By contrast, the accuracy of the leave-one-out cross-validation models decreased. This divergence became apparent around the 100 participant mark, when the ratio was 4.5 participants per variable. Also evident in Table 3 is the increase in variability in the accuracy of the models as the ratio of participants to variables decreased, as shown in the standard deviations. This suggests that the models became less stable when over-fitted by using a lower number of participants to variables, consistent with the reduced generalizability shown following leave-one-out cross-validation. Visual inspection of the accuracy statistics for the enter and backwards stepwise models do not suggest any particular bias toward the stepwise approach over-fitting the data in this sample.

## 4. Discussion

The leave-one-out cross-validation for the model utilizing all of the discussed over-fitting reduction strategies produced only a small decrement in accuracy of the classification models for both the backwards stepwise and enter approaches, i.e., dropping from 75.6% to 72.0% and 74.6% to 69.2% respectively. This indicates that these classification models were only minimally over-fitted. Classification accuracies increased as expected when progressively more over-fitted models were formed by decreasing the number of participants-to-variables ratio. Conversely, these increasingly over-fitted models performed more poorly in leave-one-out cross-validation, which indicates that the high classification accuracies were due to the spurious effects inherent to over-fitting. No convincing differences were found in the accuracies of the backwards stepwise and enter models in either the $n = 279$ or increasingly over-fitted samples.

By showing the effects of increasingly over-fitting models by using lower ratios of participants to variables we have demonstrated that it is essential for researchers to report the number of variables *initially offered* to the model, not the number that are *accepted*. Only four of the eight studies reviewed in the Introduction reported the number of variables offered to the model. Secondly, as many driving studies have fewer than 100 participants, our study suggests that over-fitted models may be common, which has serious repercussions for the field. However, it is important to note that finding indications of possible over-fitting in a study does not necessarily invalidate it. Notwithstanding, the more potential there is for over-fitting, the more cautiously the results should be interpreted.

Innes et al. (2011) investigated the accuracy of a sample of models including commonly used discriminant analysis and logistic regression, along with computationally complex models that had not been previously employed in driving research: support vector machine, product kernel density, and kernel product density. The authors employed the over-fitting reduction strategies suggested above, but did use a backwards elimination variable selection procedure. The three computationally complex models produced very high classification accuracies, two as high as 100%, while discriminant analysis and logistic regression produced accuracies of 76% and 78% respectively. Following leave-one-out cross-validation, the accuracies of the complex models all fell to within the same range for estimated prediction as the simpler models (all ranging between 72% and 76%). This indicates that it is possible to build classification models with stunningly high accuracies that do not generalize well even when following over-fitting reduction strategies and highlights the need to validate models using a resampling protocol, even in the most ideal circumstances.

In addition to considerations of how models are built, it is also necessary to consider what statistics are presented for interpretation and whether these remain true when the model is applied to a new sample. While sensitivity and specificity are values that detect true positive and negative outcomes and are routinely reported, the usefulness of a model is also determined by the number of false positives and negatives it produces, otherwise known as the positive and negative predictive values. While sensitivity and specificity are not sensitive to the base rate of the dependent variable, positive and negative predictive values are. Labarge et al. (2003) provide an example of how given a base rate of 10% and a test with 80% sensitivity and 90% specificity, the percentage of predicted positive cases that actually are positive (positive predictive value) is equal to only 47%. What these statistics show is that a model developed on a particular sample will not generalize well to another population if the base rate of the dependent variable is different. This essentially negates attempts to construct a universal model that works for a number of clinical groups without taking into consideration the base rate of the outcome variable and adjusting the model cut points accordingly. We would like to see more driving researchers report the positive and negative predictive values of their models.

There are a number of steps that driving researchers can take in the construction of classification models. One suggestion that could be easily included as part of driving research protocols include making use of *a-priori* variable selection where possible, or at least a strict limit on the number of measures offered to the model. This last step is dependent on the sample size. Knowing that we should only be offering a certain ratio of participants to variables may assist in making more pragmatic decisions about what we test and the reasons for it. All models that have incorporated over-fitting reduction strategies should be tested with a resampling validation procedure such as bootstrapping or leave-*n*-out cross-validation, as this affords the best estimate of how the model may generalize to a new sample. Because of this we recommend that the post-validation accuracies are the ones that should be reported as the final model, rather than the initial and likely inflated classification accuracies. Even if over-fitting reduction steps have not been followed in the construction of the models, resampling techniques can test the generalizability of existing models. We suggest that driving research articles report the steps of model construction so that readers can more readily assess whether the accuracy reported is more likely due to the way the model is constructed, or to the actual utility of the model to predict driving ability in the real world.

## Acknowledgements

Petra A. Hoggarth made substantial contributions to conception and design of the study, performed off-road testing of 60 participants, and analyzed and interpreted data. She was the primary writer of the article. Carrie R.H. Innes made substantial contributions to the conception and design of the study, collected data for 219 participants, and analyzed and interpreted data. She reviewed and revised the article through to the final stage. John C. Dalrymple-Alford made substantial contributions to the conception and design of the study and helped with analysis of the data. He reviewed and revised the article through to the final stage. Richard D. Jones made substantial contributions to the conception and design of the study and helped with analysis of the data. He reviewed and revised the article through to the final stage. Funding for this project was received from the Canterbury Medical Research Foundation, the Accident Compensation Corporation, and the Road Safety Trust. None of the sponsors of this research played any role

in the design, methods, participant recruitment, data collections, analysis or interpretation of the data, or preparation, review or approval of the manuscript.

## References


Babyak, M.A., 2004. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom. Med. 66 (3), 411–421.

Bowers, A.R., Anastasio, R.J., Sheldon, S.S., O'Connor, M.G., Hollis, A.M., Howe, P.D., Horowitz, T.S., 2013. Can we improve clinical prediction of at-risk older drivers? Accid. Anal. Prev. 59, 537–547.

Carr, D.B., Barco, P.P., Wallendorf, M.J., Snellgrove, C.A., Ott, B.R., 2011. Predicting road test performance in drivers with dementia. J. Am. Geriatr. Soc. 59 (11), 2112–2117.

Garson, G.D., 2010. Multiple Regression. Retrieved 9 November 2013. from http://faculty.chass.ncsu.edu/garson/PA765/regress.txt

Harrell, F.E., 2001. Regression Modeling Strategies: with Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer-Verlag New York, Inc., New York.

Hoggarth, P.A., Innes, C.R.H., Dalrymple-Alford, J.C., Jones, R.D., 2013. Predicting on-road assessment pass and fail outcomes in older drivers with cognitive impairment using a battery of computerized sensory-motor and cognitive tests. J. Am. Geriatr. Soc. 61 (2), 2192–2198.

Hoggarth, P.A., Innes, C.R.H., Dalrymple-Alford, J.C., Severinsen, J.E., Jones, R.D., 2010. 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. Accid. Anal. Prev. 42 (6), 1759–1768.

Innes, C.R.H., Jones, R.D., Dalrymple-Alford, J.C., Hayes, S., Hollobon, S., Severinsen, J., Anderson, T.J., 2007. Sensory-motor and cognitive tests can predict driving ability of persons with brain disorders. J. Neurol. Sci. 260 (1–2), 188–198.

Innes, C.R.H., Lee, D., Chen, C., Ponder-Sutton, A.M., Melzer, T.R., Jones, R.D., 2011. Do complex models increase prediction of complex behaviours? Predicting driving ability in people with brain disorders. Q. J. Exp. Psychol. 64 (9), 1714–1725.

Labarge, A.S., McCaffrey, R.J., Brown, T.A., 2003. Neuropsychologists' abilities to determine the predictive value of diagnostic tests. Arch. Clin. Neuropsychol. 18 (2), 165–175.

Lincoln, N.B., Radford, K.A., Lee, E., Reay, A.C., 2006. The assessment of fitness to drive in people with dementia. Int. J. Geriatr. Psychiatry 21 (11), 1044–1051.

Ott, B.R., Davis, J.D., Papandonatos, G.D., Hewitt, S., Festa, E.K., Heindel, W.C., Carr, 2013. Assessment of driving-related skills prediction of unsafe driving in older adults in the office setting. J. Am. Geriatr. Soc. 61, 1164–1169.

Risser, R., Chaloupka, C., Grundler, W., Sommer, M., Häusler, J., Kaufmann, C., 2008. Using non-linear methods to investigate the criterion validity of traffic-psychological test batteries. Accid. Anal. Prev. 40, 149–157.

Snellgrove, C.A., 2000. Cognitive Screening for the Safe Driving Competence of Older People with Mild Cognitive Impairment or Early Dementia. Australian Transport Safety Bureau, Canberra.

Steyerberg, E.W., Harrell, F.E., Borsboom, G.J.J.M., Eijkemans, M.J.C., Vergouwe, Y., Habbema, J.D.F., 2001. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J. Clin. Epidemiol. 54 (8), 774–781.

Tabachnick, B.G., Fidell, L.S., 2013. Using Multivariate Statistics, 6th ed. Pearson Education, Inc., Boston.

Witten, I.H., Frank, E., 2000. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco.

Wood, J.M., Anstey, K.J., Kerr, G.K., Lacherez, P.F., Lord, S., 2008. A multidomain approach for predicting older driver safety under in-traffic road conditions. J. Am. Ger. Soc. 56, 986–993.