UNIVERSITY OF BIRMINGHAM

Literature Search on Accuracy of Cognitive Screening Tests as Predictors of Unfitness to Drive

Aggressive Research Intelligence Facility
West Midlands Health Technology Assessment Collaboration

April 2006

For the Drivers Medical Group
DVLA
Swansea

ARIF

WMHTAC
West Midlands HTA Collaboration

About ARIF and the West Midlands Health Technology Assessment Collaboration

The West Midlands Health Technology Assessment Collaboration (WMHTAC) is an organisation involving several universities and academic groups who collaboratively produce health technology assessments and systematic reviews. The majority of staff are based in the Department of Public Health and Epidemiology at the University of Birmingham. Other collaborators are drawn from a wide field of expertise including economists and mathematical modellers from the Health Economics Facility at the University of Birmingham, pharmacists and methodologists from the Department of Medicines Management at Keele University and clinicians from hospitals and general practices across the West Midlands and wider.

WMHTAC produces systematic reviews, technology assessment reports and economic evaluations for the UK National Health Service's Health Technology Assessment (HTA) programme, the National Institute for Health and Clinical Excellence (NICE). Regional customers include Strategic Health Authorities, Primary Care Trusts and regional specialist units. WMHTAC also undertakes methodological research on evidence synthesis and provides training in systematic reviewing and health technology assessment.

The two core teams within WMHTAC are the Aggressive Research Intelligence Facility (ARIF) and the Birmingham Technology Assessment Group (BTAG)

ARIF provides a rapid on-demand evidence identification and appraisal service primarily to commissioners of health care. Its mission is to advance the use of evidence on the effects of health care and so improve public health. The rapid response is achieved by primarily relying on existing systematic reviews of research, such as those produced by the Cochrane Collaboration, the National Institute for Health and Clinical Excellence (NICE), the NHS Centre for Reviews and Dissemination, and the NHS Health Technology Assessment (HTA) programme. In some instances, longer answers to questions are required in which case mini rapid reviews of existing systematic reviews and key primary studies are compiled, typically taking 1-2 months to complete.

Occasionally a full systematic review is required and then topics are referred to BTAG who coordinate the production of systematic reviews for several customers under a number of contracts. ARIF is intrinsically involved in the production of these systematic reviews.

Aggressive Research Intelligence Facility (ARIF)
West Midlands Health Technology Assessment Collaboration (WMHTAC)
Department of Public Health and Epidemiology
University of Birmingham
Birmingham
B15 2TT

arifservice@bham.ac.uk
0121 414 3166

Warning

This is a confidential document.

Do not quote without first seeking permission of the DVLA and ARIF.

The information in this report is primarily designed to give approved readers a starting point to consider research evidence in a particular area. Readers should not use the comments made in isolation and should have read the literature suggested. This report stems from a specific request for information, as such utilisation of the report outside of this context should not be undertaken. Readers should also be aware that more appropriate reviews or information might have become available since this report was compiled.

1 Aims

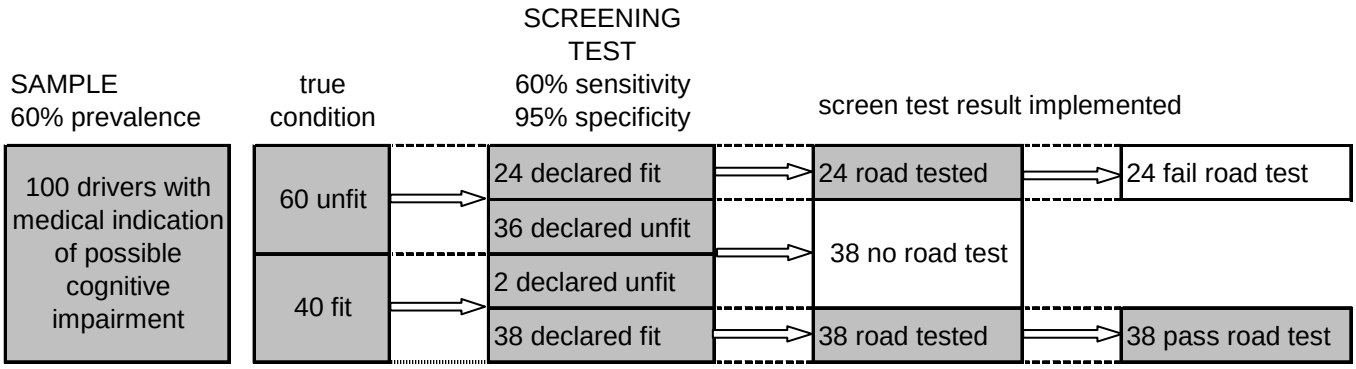
The aim was to identify, appraise and summarize existing reports on the accuracy of cognitive screening tests for determining unfitness to drive in people with medical indication(s) of possible cognitive impairment.

Further details are given in the request submitted by the Drivers Medical Group (Appendix 1 – Details of Request)

2 Background

A screening test for drivers with medical indication of possible cognitive impairment might be capable of identifying those individuals unfit to drive so that an “on road” driving test would be redundant. Such a test has potential advantages of cost and convenience in terms of road tests avoided. For such a test to be useful the first requirement is that it should reach an acceptable level of accuracy in identifying persons unfit to drive. What minimal level of accuracy is considered acceptable is clearly a matter of judgement but should rest on unbiased estimates grounded in well-conducted studies.

Any test that is likely to be useful will be characterised by a high specificity in detecting unfitness to drive. This means that the return of a positive test result for an individual is likely to rule in the presence of unfitness. Figure 1 illustrates this principle by presenting a hypothetical scenario in which the screening test has a specificity of 95% and a sensitivity (that is ability to detect unfitness amongst those who are unfit) of 60% and the test is applied to 100 individuals of whom 60% are unfit to drive. The high prevalence of unfitness to drive among the tested population was chosen to reflect a real world situation.



In this example if the screening test results were to be implemented 38% of on road tests would be avoided. The problem is that 2 of the 38 identified as unfit according to the screen would in fact have passed an on road driving test. The acceptability of incorrect classifications of this kind will be a matter of judgement likely to be influenced by factors additional to the crude rate of such errors. These might include the frequency with which screening is repeated and the probability that a misclassified individual would be correctly classified on subsequent screening and also any knowledge about the average rate of deterioration (or change) in cognitive test scores amongst tested populations.

Two further issues are important in the context of the above example. First, in the calculation of road tests avoided, no account was taken of the uncertainty around the estimates of test accuracy. For example if the estimate of sensitivity had been derived from a study with 50 individuals unfit to drive the 95% confidence interval for sensitivity would be 46% to 74% and the 95% CI of the number of these declared unfit in the example would be 28 to 44. Second, it is only valid to apply the test accuracy values if they were established in a population similar to that which is to be screened; thus it is important that a description of both populations is available so that interpretation of results can be made reliably. Furthermore if a test is to be widely applied it should exhibit high inter-rater and re-test reliabilities.

Background information is given in the documentation supplied by the Drivers Medical Group contained in Appendix 1 – Details of Request.

3 Methods

Briefly these were:

- To undertake a search for studies looking at the accuracy of cognitive tests used for screening for unfit drivers
- To initially search for existing systematic reviews on this topic.
- To identify as many primary studies as possible within constraints of time and resources
- To concentrate on primary studies if no suitable reviews were identified
- To appraise the studies with due attention to their methodological quality
- Where appropriate and possible data on relevant outcomes was to be extracted and tabulated.
- Summary and data analysis would depend on information identified.

3.1 Searches

3.1.1 Existing Reviews.

Searches to identify existing systematic reviews on this topic were performed utilising the well-established ARIF search protocol (Appendix 3 – Search strategies)

3.1.2 Primary Studies

Searches were undertaken for primary studies in the Cochrane Library, MEDLINE and EMBASE. The search strategies included text and index terms for cognition, cognitive function, neuropsychological tests and executive function combined with terms for driving. ‘Filters’ for diagnostic tests were also incorporated in MEDLINE and EMBASE. The databases of the Transport and Road Laboratory (TRL) and the National Transport Laboratory (TRIS) were also interrogated.

The detailed search strategies can be found in Appendix 3 – Search strategies page 22.

Searches were predominantly undertaken by an information specialist with additional searches by a research reviewer. Both interacted to ensure searches were conducted appropriately.

An information specialist and a research reviewer scanned the search results for relevance based on information in the title and abstract. Studies that appeared to report accuracy of screening tests for fitness to drive were obtained in full.

Full copy articles were assessed for their match to the report's aim using the criteria outlined below in order to obtain the most informative articles for further scrutiny and reporting.

Design:	Prospective study.
Population:	Drivers with a medical indication or referral that infers possible cognitive impairment.
Screen:	A screening test for unfitness to drive based on one or more measures related to cognitive performance.
Reference standard:	An on road test used to determine true fitness and unfitness to drive.
Outcome:	Studies that report one or more measures of test accuracy or that include sufficient data to allow their calculation.
Exclusion:	Studies which investigated correlations between results of cognitive performance tests and crash frequencies or other surrogates of fitness to drive. Studies of populations with poor vision due to ocular abnormalities or those with sleep pathology such as narcolepsy or sleep apnea. Studies that only report correlation between test performance and on-road performance as an indicator of potential test utility.

Full copy articles were assessed for their match to the question addressed (external validity) and the most informative articles subjected to further scrutiny and reporting.

The reference lists of the most relevant articles were also checked in order to identify further relevant papers.

3.2 Study Quality

There is a theoretical basis and empirical evidence[1] to suggest that studies of diagnostic test accuracy are subject to bias from various sources that may threaten the validity of reported findings.[2,3] Therefore in assessing the included studies attention was focused on the following elements of study design and conduct:

Population: For measures of accuracy to be useful they should be conducted in populations that closely resemble those for whom the test will be used. The study population is best assembled from consecutively recruited referred patients. In particular if studies include "normal" or "healthy volunteer" drivers in their sample then this will

compromise validity of accuracy measures of screening performance for drivers with medical indications of potential cognitive impairment.

Blinding: The screening test should be done by a rater blinded to the result of the reference standard test, and vice versa.

Missing results: All individuals in the study population should receive both screening and reference standard tests. If missing values occur and are randomly distributed between fit and unfit drivers then they may be of little importance. However there is often no way of knowing about the distribution of missing values in which case they may materially influence the accuracy measures reported.

Reference standard: Since licensing depends on the outcome of a binary on-road driving test, the most suitable reference standard is considered to be an on road driving assessment conducted by an experienced examiner or instructor. Studies using a different reference standard (e.g. driving simulator) or those using a non-binary on road test will be considered to have used an imperfect standard and the accuracy measures then reported in terms of "agreement" between screen test and imperfect standard.[4] All individuals in the sample should receive the same reference standard.

Accuracy measures: The outcome measures of choice are sensitivity and specificity. Where these have been reported or are calculable they will be given emphasis. Positive predictive values (i.e. positive test for unfitness) are influenced by prevalence of unfitness in the populations tested. Where PPV has been reported and sensitivity and specificity are not calculable then predictive values will be remarked upon.

3.3 Summarizing study results

Some studies defined a positive test result as one designating unfitness to drive while in others a positive test result was indicative of fitness to drive. This means that "specificity" reported in some studies corresponds to "sensitivity" reported in others (a similar reversal holds for the other pairs of accuracy measures: + / - predictive values and + / - likelihood ratios). For consistency we have calculated test results on the basis that the tests were detecting unfitness; thus if the test designates an individual as unfit this is defined as a positive test result.

Depending on the heterogeneity of studies accuracy measures were combined by standard meta-analyic procedures so as to generate more precise estimates. Where tests and populations were heterogeneous and meta-analysis ill advised accuracy measures were tabulated and sensitivity and specificity values represented in ROC space as appropriate. Emphasis was given to the results of studies conducted in the UK since these are the more likely to have external validity.

No systematic search was conducted for inter-rater or re-test reliability of screening tests, however where these have been considered in the primary studies the outcomes reported will be remarked upon.

Table 1 Major characteristics and results of studies of the accuracy of cognitive screening tests for unfitness to drive.

STUDY COUNTRY	POPULATION INDICATION [N, age, % male] *SELECTION METHOD*	SCREEN TEST	REFERENCE STANDARD TEST	ACCURACY RESULT % [95% CI]	COMMENT *[prevalence of unfitness(%)]*
Nouri 1993[5] UK	Referrals with Stroke who had been driving at least for 10 wks prior to study. [27, mean 58.8 yrs, 85%]. *Recruitment method not reported.*	SDSA (n=27) (also random allocation to GP assessment, n=25)	On road test with instructor. Binary outcome. Prior to index test. Blinding not reported.	SDSA only: Sen 84 [68 -100] Spe 75 [45 -100] PPV 89 [75 -100] NPV 67 [36 - 98]	Small study, blinding uncertain, details of sample selection incomplete, accuracy moderate with wide CI. The GP assessment was inferior to SDSA; SDSA was superior to chance. *[70%]*
Radford 2004[6] UK	Diagnosis with traumatic brain injury. [52, mean 31 yrs, 85%] *Recruited from Mobility Clinic waiting list; inclusion criteria described.*	SDSA. SDSA + 4 others.	On road test with instructor. Assessors were blinded.	SDSA SDSA+4 other Sen 36 [11 - 61]; 64 [39 - 89] Spe 84 [73 - 96]; 95 [88 - 100] PPV 45 [16 - 75]; 82 [59- 100] NPV 78 [65 - 91]; 88 [78 - 98]	Small study, accuracy moderate with wide CI. Costs met by participants, may result in selection bias. Five missing values and so only 52 included in analyses. *[27%]*
Sentinella 2005[7] UK	Stroke patients at least 1 month post stroke. [42, mean 65.2 yrs, 88%]. *Recruited via clinics, advertising and other channels. Inclusion criteria described.*	SDSA	On road test with instructor. Assessors were blinded.	Sen 71 [49 - 92] Spe 50 [30 - 70] PPV 50 [30 - 70] NPV 71 [49 - 92]	15 drop outs from the stroke group leaving 42 participants. One missing value and so 41 analyzed. Recruitment likely to result in selection bias. Inter rater reliability and re-test reliability was also investigated. *[41%]*
Lunberg 2003[8] Scandinavia	Diagnosis of stroke. [97, mean 63 yrs, 90%]. *Recruitment method not reported.*	SDSA; also SDSA "adapted" for Scandinavian drivers.	On road test with instructor. Assessors probably blinded.	SDSA: Sen 61 Spe 76 SDSA modified: a] based on a repeat reference test administered for borderline drivers after a training intervention. b] A retrospective analysis of a randomly selected subset (n=48). Both a] & b] subject to bias and not included here.	Numbers for SDSA not reported so that CIs, prevalence and predictive values cannot be calculated. Two further sets of accuracy results were reported: one subject to bias because all sample did not receive the same reference test, the other a retrospective analysis of random sub-sample. *[Prevalence un-calculable]*

STUDY COUNTRY	POPULATION INDICATION [N, age, % male] *SELECTION METHOD*	SCREEN TEST	REFERENCE STANDARD TEST	ACCURACY RESULT % [95% CI]	COMMENT *[prevalence of unfitness(%)]*
Christie 2001[9] UK	Diagnoses of brain damage if they were neurologically stable. [39, 20-55 yrs, 69%] *Recruited at head-injury clinic if accepted incentive to participate.*	Investigated a Test Battery of multiple elements from which 3 were selected for discriminatory test.	On road test with driving advisor. Scores compressed to binary outcome. Blinding not reported.	Sen 56 [23 - 88] Spe 93 [84 -100] PPV 71 [38 -100] NPV 88 [76 - 99]	Small study, accuracy moderate with wide CI. Use of incentive to recruitment may result in selection bias (tested if unrepresentative on basis of severity). *[23%]*
McKenna 2004[10] UK	Diagnoses implicating brain functioning. [142, mean 62 yrs, 81%]. *Recruitment method not reported. Inclusion criteria described.*	Battery of 12 tests. Single composite score derived for each client and cut off score selected for +ve test result.	On road test with instructor. Assessor was not blinded.	Sen 40 [27 - 54] Spe 97 [93 - 100] PPV 91 [80 - 100] NPV 69 [60 - 78]	20 missing values resulted in 122 out of 142 contributing to analyses. Accuracy here calculated from data in table 4[†]. *[43%]*
Schanke 20009[11] Scandinavia	Diagnosis (CT) of brain damage. [55, mean 56.1 yrs, 76%]. *Recruited all those at one hospital who were referred for driving assessment*	Battery of tests (10 elements) administered by one neurophysiologist. Clients classified into 5 categories, no cut off used	On road test with instructor: clients classified into 6 categories. Non binary. Assessors blinded.	Sen 80 [66 - 94] [††] Spe 92 [81 - 100] PPV 92 [81 - 100] NPV 79 [65 - 94]	A single rater produced result of screen test without reliance on formal cut off; this result likely biased by subjective judgments. 18 clients with worst screen classifications were not given on road test. *[54%]*
Hannen 1998[12] Germany[‡]	Diagnosis (CT) of brain damage (trauma, vascular, other) [116, mean 46 yrs, 88%]. *Referred for or asked to be tested.*	Battery of six tests from which 4 were selected to contribute to the discriminatory test.	On road test with instructor. Assessors probably blinded.	Sen 69 [50 - 82] Spe 71 [60 - 82] PPV 63 [50 - 77] NPV 75 [65 - 86]	Well conducted study; 3 missing values. Accuracy moderate with wide CIs. Authors recommend on-road test for all clients. *[42%]*
Galski 1993[13] USA	Diagnoses of traumatic or cerebrovascular brain injury. [106, mean 47 yrs, not reported%]. *Recruitment method not reported.*	Battery of 8 "pre-driver" cognitive tests; (a drive-simulator & "behaviours" during driving also used)	On road test with certified instructor. Blinding not reported.	Pre-driver tests Sen 71 Spec 87	Numbers not reported so that CIs, prevalence and predictive values cannot be calculated. Authors also present Sen and Spec for pre-driver test + behaviour scores and drive-simulator scores. *[Prevalence un-calculable]*

CI = 95% confidence intervals. SDSA = Stroke Drivers Screening Assessment. [†] discrepancy between results in table 5 (N=128) and table 4 (N=122). [††] These values should be viewed skeptically since they were calculated using *post hoc* cut off (between categories 2b and 2c of index test) and assumed ambiguous road tests represented unfitness and that all untested clients would have failed road test.
[‡] German language manuscript.

4 Results

4.1 Reviews identified

No relevant systematic reviews were identified.

Several narrative reviews were identified (e.g. Hogan 2005[14], Lloyd 2001[15]). These discussed the practical and theoretical issues relevant to development and use of screening tests for unfitness to drive but none assessed or reported measures of test accuracy.

4.2 Primary studies identified

Of 250 publications retrieved from searches nine primary studies were found to be directly relevant and to fulfill the inclusion criteria. Five were conducted in the UK,[5-7,9,10] one in Germany,[12] two in Scandinavia,[8,11] and one in USA.[13] The major characteristics of these studies are summarised in Table 1.

Six further studies[16-20] were identified that employed a battery of cognitive tests but these were rejected because the population or reference tests or both were inappropriate or because of a retrospective study design or because no accuracy measure was reported or calculable. They are listed in the bibliography because they offer further insights into the development of screening tests.

4.2.1 Screening tests

The screening tests used varied in their complexity and the apparatus required which ranged from simple picture cards (as with the Stroke Drivers Screening Assessment test) to more complex auditory or visual challenges delivered via computers or other means. Identity of test administrators was unreported except in one study[11] in which a single neurophysiologist delivered the tests. Time required to deliver tests varied from 0.5 to 2 hours or was unreported. In most studies it was stated or apparent that the test could be carried out in the client's home.

The Stroke Drivers Screening Assessment (SDSA) was investigated in 4 studies.[5-8]

A battery of tests was used in six studies.[6,9-13] The approach in these studies was generally as follows:

1. A mixed group of putatively fit (e.g. volunteers) and putatively unfit (e.g. referred) individuals were given driving tasks and each individual assigned an overall score or a score on several elements thought to reflect driving skills.
2. The same individuals were assessed using a batch of cognitive tests.
3. The correlation between individual's driving scores and scores in each of the cognitive tests was then calculated. Alternatively scores in particular elements of driving skill were correlated with cognitive test scores. The resulting correlation coefficients indicated the extent to which each test could explain the variation in driving skills.

4. The scores in those cognitive tests that explained most of the variance in driving ability (good positive correlation) were then combined (e.g. employing logistic regression methods). Different tests were given various weightings according to their power in explaining variation in driving skills. The paradigm of parsimony in test combination was then used to determine the particular combination that gave the best predictive performance when applied to the putatively unfit drivers (who were independently classified as truly fit or truly unfit using an on road test). Alternatively scores on all tests in the battery were combined using weightings indicated by their correlation coefficients.

A problem with this general approach is that the resulting screening test developed will be deeply rooted in the properties of the population used in its development, so that generalizing these accuracy measures to other populations of interest is unlikely to be valid. Tests developed in this way need to be challenged with other sample populations to find out if their performance holds up. However this approach may identify combinations of tests that are suitable for particular medically-defined groups of clients.

4.2.2 Reference standard tests and blinding

The reference standard used in all studies to determine true fitness or unfitness to drive was an on road driving test with an experienced driving instructor. The tests lasted between 0.75 and 2 hours or the duration was unreported (5 studies). Other details of the tests were usually provided (such a number of items scored in order to determine the test result); no single accepted standard test system exists and so it is likely the reference standard varied between studies and especially between studies conducted in different countries. In one study[8] one version of accuracy estimates rested on an analysis in which a sub-sample of the population received different reference test to the rest.

Assessors were blinded to the screening result in four studies,[6,7,11,12] not blinded in one 41 and blinding was not reported in three.[5,9,13]

4.2.3 Populations

Populations studied were small ranging from 27[5] to 142[10] and so were unlikely to deliver precise estimates of screen test accuracy. Six studies included brain-damaged patients with various etiologies especially traumatic injury or cerebro-vascular events, but three studies[5,7,8] exclusively concerned stroke patients. Only one study recruited consecutively referred patients (Schanke[11] studied all patients referred to one hospital) and methods used to select patients were generally poorly reported. The prevalence of unfitness to drive in study populations varied between 23% and 70% (un-calculable for two studies).

4.2.4 Missing values

Except for the smallest study (Nouri 1993[5]) and one other[8] all studies exhibited missing values. In most studies these represented a substantial proportion of the total sample. Missing values were almost exclusively due to a proportion of clients not undergoing the on road reference test because screening had indicated that they would represent a danger to themselves and to others while driving.

4.2.5 Screen test accuracy: sensitivity and specificity

The test accuracy values reported or calculated from study data were all imprecise and characterised by wide confidence intervals. Specificity values in different studies ranged from 50% to 97% and sensitivities from 36% to 84%. The trade off between specificity and sensitivity in the individual studies is illustrated in the receiver operating characteristic graph shown in Figure 1. No studies combined good performance in both measures of accuracy. The values shown for Schanke 2000[11] are considered as measures of agreement with an imperfect standard and should be viewed with caution (see Appendix 4).

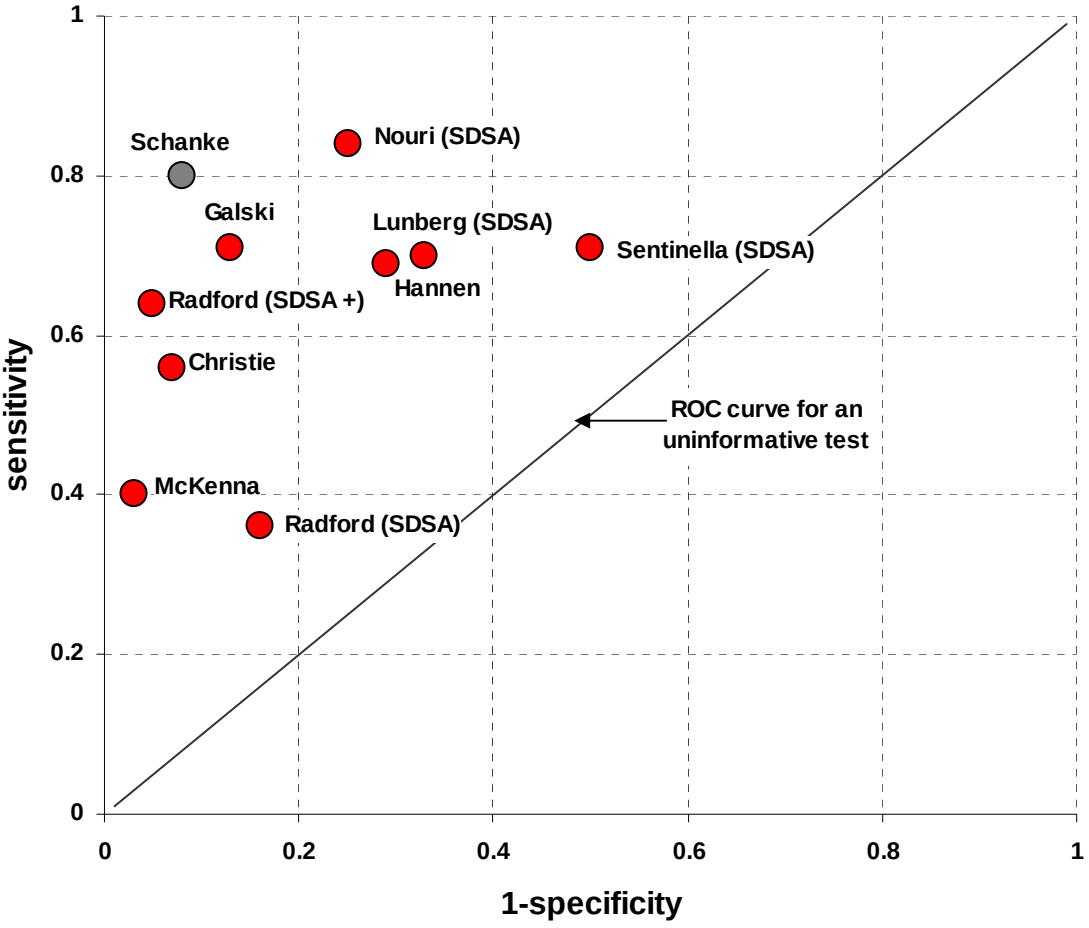


Figure 1 Receiver operating characteristics graph for screening studies.

Of studies conducted in the UK that of McKenna 2004[10] returned the best specificity value (> 97%) but this was at the expense of a moderate sensitivity in which false negatives outnumbered true positives. These values were calculated from data in table 4 of the study. Table 5 reports different sensitivity and specificity values but these are in fact positive and negative predictive values. The relationship between sensitivity and specificity in this study can be illustrated by plotting the data from table 4 in a ROC graph (Figure 2). This shows that to achieve a sensitivity of > 60% specificity must be sacrificed to less than 90% meaning that of fit

drivers screened 10% would be falsely designated unfit to drive. These values are associated with considerable uncertainty and are likely to overestimate test accuracy because of a lack of blinding in administering reference and screening tests.

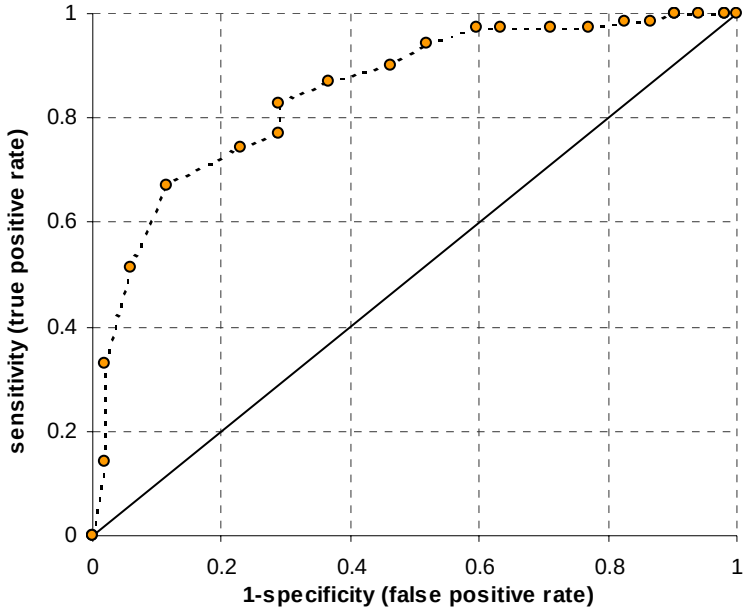


Figure 2 Receiver operating characteristic curve for screening test of McKenna 2004.

Three UK studies (Sentinella 2005,[7] Nouri 1993[5]and Radford 2004[6]) and a Scandinavian study[8] investigated the performance of the Stroke Drivers Screening Assessment. The sensitivity and specificity values were relatively poor (Figure 1), they were associated with considerable uncertainty and were noticeably different to each other. Some of the difference in results may be explained in part by use of different cut-offs for distinguishing unfit from fit drivers and also by differences in populations (Radford[6] traumatic brain injury, the other two stroke patients). Accuracy of screening was considerably improved in Radford's study by combining the SDSA with 4 additional test elements so that specificity reached 95%. Sentinella[7] remarked that the inter-rater and re-test reliability of the SDSA gave some cause for concern, however the investigation again was hampered by very small numbers of participants. The study of Nouri[5] reported the highest sensitivity of any of the studies, however specificity was poor indicating that a quarter of fit drivers would be designated unfit by the test. The study recruited only 27 individuals to SDSA and confidence intervals were consequently very wide.

The non-UK studies reported sensitivity and specificity values that fell in about the same range as the UK studies. Hannen 1998[12] found a battery of cognitive tests to be only moderately discriminatory when administered to a mixed population of stroke and other brain damaged patients. The tests of Galski1993[13] produced a sensitivity and specificity of 71% and 87%; the data presented did not allow the prevalence of unfitness to be calculated and so the proportion incorrect test results is not known. These authors improved the accuracy values of the screening test by incorporating scores on "behaviours" observed during the on-road-driving test; the values have not been reported here since they would be subject to incorporation bias.

4.2.6 Screen test accuracy: predictive values.

Several studies emphasized and most reported predictive values as measures of screening test accuracy; of studies that did not report the values they could be calculated form the published data in all cases except for the study of Galski 1998 and Lundberg.[8,13]

Positive and negative predictive values (PPVs and NPVs) are sensitive to prevalence of unfitness to drive in the screened population and this may vary according to setting. Prevalence varied between 23% and 70% in the eight studies reviewed here. Figure 3 depicts the PPVs reported in seven studies of screening tests and illustrates how these would change under the influence of altered prevalence. PPV for four of the studies (Radford,[6] Sentinella,[7] Hannen[12] and Christie[9]) lie on steep curves relating PPV to prevalence indicating that these PPVs would be particularly sensitive to shifts in prevalence.

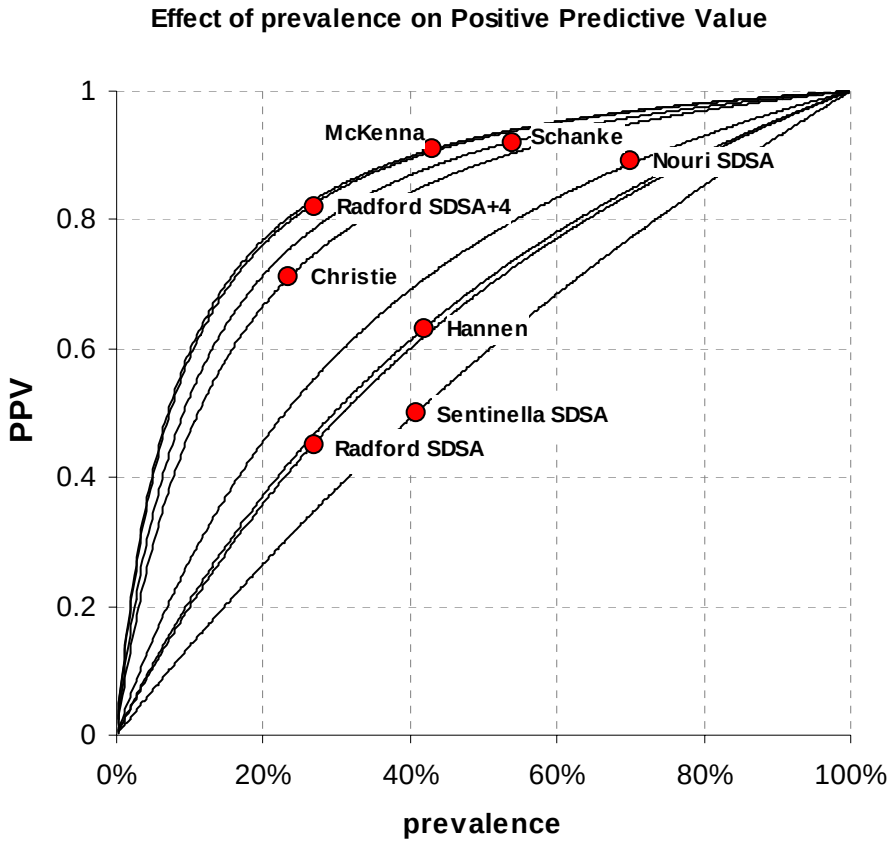


Figure 3 Positive predictive values reported in studies of screening tests.

Figure 4 depicts the NPVs reported in seven studies of screening tests and illustrates how these would change under the influence of altered prevalence. NPVs for all studies lie on steep curves relating PPV to prevalence indicating that these NPVs would be particularly sensitive to shifts in prevalence.

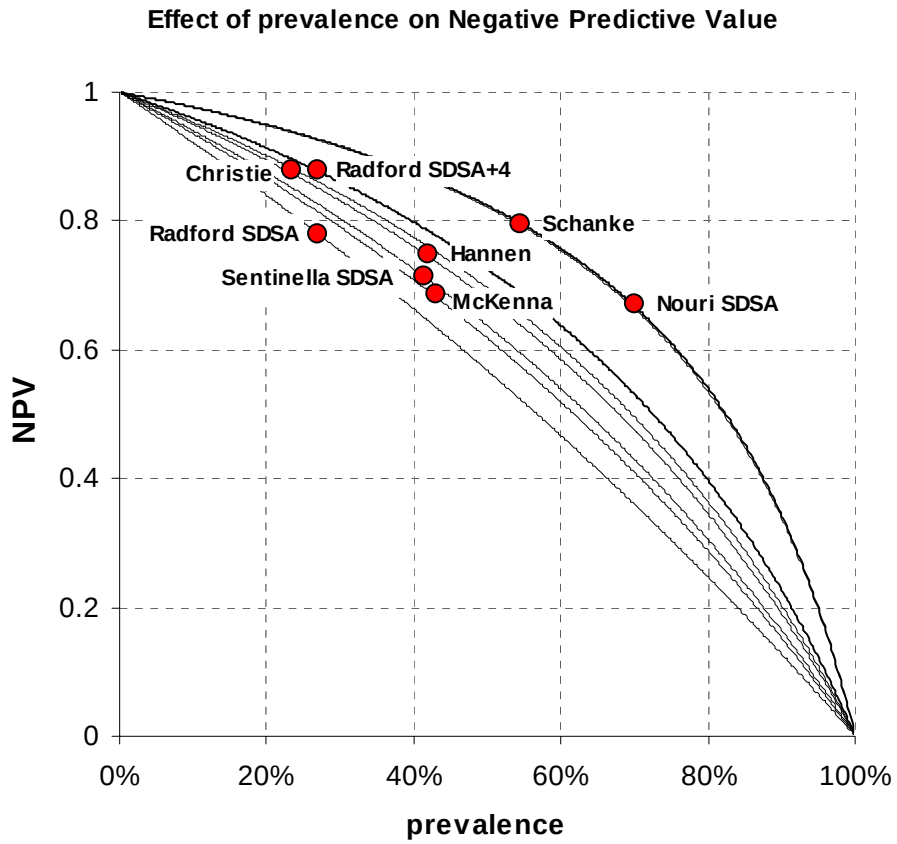


Figure 4 Negative predictive values reported in screening studies.

4.2.7 Test reliability and patient-acceptability

In most studies the number of administrators of the index test was unreported. Apart from one study[7] test performance in terms of inter-rater reliability and test re-test reliability were either not investigated or remained unreported. Sentinella[7] found inter-rater reliability in interpretation of scores in the SDSA was poor despite high reliability in test element scores. These results were based on a sample of only three patients. Test re-test reliability was assessed in a sample of six patients four of whom were clearly fit to drive; reliability reached reasonable levels (the Intra Class Correlation statistic varied between different elements of the test). Reliance on these results is hampered by the small sample sizes employed.

Neuro-physiological and or cognitive tests may be less acceptable to clients that an on-road test of fitness to drive. Such tests can be unpopular with patients (e.g. Nagels et al 2005[21]). A Medline search for studies of test-acceptability amongst cognitively impaired-patients retrieved 171 references but none of these proved relevant. It appears unlikely that research on this question has been performed.

5 Conclusion

The amount of evidence bearing on the accuracy of screening tests designed to determine fitness to drive amongst individuals with a medical indication of possible cognitive impairment appears to be meager. No systematic review was identified and only eight primary studies were readily retrieved by the broad search strategy employed here.

Those studies that have been conducted have recruited only small numbers of individuals and the results are associated with considerable statistical uncertainty. Furthermore in most cases the accuracy achieved by the tests was moderate. The tests and populations used in these studies were too heterogeneous to allow the use of meta-analytic methods that would have improved precision of accuracy estimates.

The tests with better accuracy performances in terms of test sensitivity and specificity have been developed using procedures that use correlation analysis to winnow out elements of test batteries that result in the greatest discriminatory power for the population under examination. This means test performance is rooted in the properties of that particular population. Until tests developed in this way are challenged with further population samples their validity remains in question. The SDSA test was examined in 3 UK studies; on its own this test appears to perform differently in different populations or when administered by different personnel. None of the other tests retrieved by the search conducted for this report have been challenged in more than one sample population.

Methods for selection of the study populations used in developing the screening tests were poorly reported and there are reasons to believe selection bias was introduced by use of incentives or participation via advertisement or invitation. Only one study recruited consecutive patients but unfortunately accuracy of the screen in this study was limited by the non-binary nature of both the reference standard and the screen test and by the fact that a large proportion of the sample did not receive the reference standard.

It is quite possible that the amount of variance in open road driving performance that could be accounted for by tests of the sort used in the studies summarized here is insufficient to allow a screening test to completely replace the on road driving test for people with possible cognitive impairment. Certainly the evidence presented in these published studies does not refute this conclusion and the authors of one of the nine studies summarized above concluded that, on the basis of the screening results, an on road test should be offered for all clients. There may be scope for improvement and or expansion of cognitive and neuro-physiological tests with commensurate greater accuracy in screens that incorporate these, however they will need testing in larger populations than used in the studies summarized here in order to reach a reasonably precise estimate of their accuracy.

5.1 Limitations of report

This is not a systematic review but a rapid assessment for relevant literature. Although the search strategies were broad and comprehensive for both systematic reviews and primary studies it cannot be guaranteed that relevant studies were not omitted. However brief citation checking of relevant articles did not identify further studies.

6 References

6.1 Main References

1 Rutjes AW, Reitsma JB, Di NM, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006; **174**(4):469-476.

2 Whiting PF, Westwood ME, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Med Res Methodol* 2006; **6**(1):9.

3 Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, *et al*. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Fam Pract* 2004; **21**(1):4-10.

4 FDA. Statistical guidance on reporting results from studies evaluating diagnostic tests: draft guidance for industry and FDA reviewers. *http://www fda gov/cdrh/osb/guidance/1428 pdf* 2006;1-19.

5 Nouri FM, Lincoln NB. Predicting driving performance after stroke.[see comment]. *BMJ* 1993; **307**(6902):482-483.

6 Radford KA, Lincoln NB, Murray-Leslie C. Validation of the stroke drivers screening assessment for people with traumatic brain injury. *Brain Injury* 2004; **18**(8):775-786.

7 Sentinella J, Read L. The validity and reliability of the stroke drivers screening assessment. *TRL Report PPR061* 2005;1-70.

8 Lundberg C, Caneman G, Samuelsson SM, Hakamies-Blomqvist L, Almkvist O. The assessment of fitness to drive after a stroke: the Nordic Stroke Driver Screening Assessment. *Scandinavian Journal of Psychology* 2003; **44**(1):23-30.

9 Christie N, Savill T, Grayson G, Ellison B, Newby G, Tyerman A. The assessment of fitness to drive after brain injury or illness. *TRL Report 485* 2001;1-17.

10 McKenna P, Jefferies L, Dobson A, Frude N. The use of a cognitive battery to predict who will fail an on-road driving test. *British Journal of Clinical Psychology* 2004; **43**(Pt 3):325-336.

11 Schanke AK, Sundet K. Comprehensive driving assessment: neuropsychological testing and on-road evaluation of brain injured patients. *Scandinavian Journal of Psychology* 2000; **41**(2):113-121.

12 Hannen P, Hartje W, Skreczek W. Evaluating driving ability after brain damage. Neuropsychological diagnosis and driving test. *Nervenarzt* 1998; **69**(10):864-872.

13 Galski T, Bruno RL, Ehle HT. Prediction of behind-the-wheel driving performance in patients with cerebral brain damage: a discriminant function analysis. *American Journal of Occupational Therapy* 1993; **47**(5):391-396.

14 Hogan DB. Which older patients are competent to drive? Approaches to office-based assessment.[see comment. *Canadian Family Physician* 2005; **51**:362-368.

15 Lloyd S, Cormack CN, Blais K, Messeri G, McCallum MA, Spicer K, *et al*. Driving and dementia: a review of the literature. *Canadian Journal of Occupational Therapy - Revue Canadienne d Ergotherapie* 2001; **68**(3):149-156.

16 De Raedt R, Ponjaert-Kristoffersen I. Short cognitive/neuropsychological test battery for firsttier fitness-to-drive assessment of older adults. *Neuropsychology, Development, & Cognition Section D, The Clinical Neuropsychologist* 2001; **15**(3):329-336.

17 Kantor B, Mauger L, Richardson VE, Unroe KT. An analysis of an older driver evaluation program. *Journal of the American Geriatrics Society* 2004; **52**(8):1326-1330.

18 Mazer BL, Korner-Bitensky NA, Sofer S. Predicting ability to drive after stroke. *Archives of Physical Medicine & Rehabilitation* 1998; **79**(7):743-750.

19 Fox GK, Bashford GM, Caust SL. Identifying safe versus unsafe drivers following brain impairment: the Coorabel Programme. *Disability & Rehabilitation* 1992; **14**(3):140-145.

20 Fox GK, Bowden SC, Bashford GM, Smith DS. Alzheimer's disease and driving: prediction and assessment of driving performance. *Journal of the American Geriatrics Society* 1997; **45**(8):949-953.

21 Nagels G, Geentjens L, Kos D, Vleugels L, D'hooghe MB, Van AP, *et al*. Paced visual serial addition test in multiple sclerosis. *Clinical Neurology & Neurosurgery* 2005; **107**(3):218-222.

7 Appendices

7.1 Appendix 1 – Details of Request

ARIF REQUEST FORM

Date of Request	1 / 12 / 05
Lead Medical Adviser Issuing request	Name – Dr Graham Wetherall Secretary to Neurological Disorders and Psychiatric Panels
Contact details	Drivers Medical Group DVLA Sandringham Park Swansea Vale Llansamlet Swansea SA7 OAA

1. Without worrying about the structure of the question, state in full the nature and context of the problem.

We need to know:
What tests of cognitive function predict driving ability in psychiatric illnesses.

2. Please give a background to the question. Why has DMG raised this problem?

We need to introduce a simple screening test, which would be of benefit in the driver licensing process without putting individuals through a practical driving assessment. It would be helpful to have an understanding of the tests of executive function as predictors of safe driving.

3. Giving references where appropriate, briefly detail the sources you have used to obtain background information on the *options* and *issues*, which might be important for the problems, you describe.

(a) Minutes of Secretary of State for Transport's Honorary Medical Advisory Panel on Driving and Disorders of the Nervous System held on:

22 November 2004
23 March 2005

(b) 'Fitness to Drive and Cognition', The British Psychological Society, 2001

(c) 'The use of a cognitive battery to predict who will fail an on-road driving test', British Journal of Clinical Psychology (2004)

(d) 'Driver landmark and traffic sign identification in early Alzheimer's disease' EY Uc et al, J Neurol Neurosurg Psychiatry 2005;76:764-768

4. Please give name and contact details of any expert or clinical contact e.g. relevant Panel Chairman/expert Panel member.

Professor Malcolm Lader (Chairman)
OBE PhD DSC MD FRCPsych FMEDSci
Institute of Psychiatry
Denmark Hill
London SE5 8AF

[redacted]

Dr P Divall (Panel Member)
Consultant in Old Age Psychiatry
St Martin's Hospital
Midford Road
Bath BA2 5RP

[redacted]

Dr Judith Morgan (Deputy Panel Secretary)
[redacted]

5. What is the nature of the target population of the issue detailed above? E.g. age, profile, vocational drivers, young drivers, other co-morbid features.

Whilst the emphasis is likely to be on the older age group, an across the board profile is required to include the younger age groups more likely to suffer serious psychiatric illness rather than degenerative brain disease or head injury. The influence of treatment with medication on driving ability would be helpful.

6. What are the outcomes you consider particularly important in relation to the question posed? What decisions rest on these outcomes?

Identify tests, which are effective and relatively simple to apply. Executive function may be a marker across a range of psychiatric illnesses, including dementia, depression, psychosis and learning disability. Greater screening power to enable us to make licensing decisions more quickly. Possibly without the costs incurred with practical driving tests.

What is the latest date that an ARIF response would be of value 1 / 4 / 2006

Please either:

Fax this form to: 0121 414 7878 marking FAO ARIF

E-mail as a word document or pdf attachment to: [redacted]

Post to:- **Dr David Moore**
Senior Research Reviewer and Analyst
Aggressive Research Intelligence Facility
West Midlands Health Technology Assessment Collaboration
Department of Public Health
University of Birmingham
Edgbaston
Birmingham
B15 2TT

Please ring 0121 414 3166 or 6767 if you have any queries, or you want to check the progress with your request.

7.2 Appendix 3 – Search strategies

7.2.1 ARIF Reviews Protocol

SEARCH PROTOCOL FOR ARIF ENQUIRIES

(July 2005)

In the first instance the focus of ARIF's response to requests is to identify systematic reviews of research. The following will generally be searched, with the addition of any specialist sources as appropriate to the request.

A. Cochrane Library

- Cochrane Reviews
- Database of Abstracts of Reviews of Effectiveness (DARE)
- Cochrane Central Register of Controlled Trials (CENTRAL)
- Health Technology Assessment (HTA) database

B. ARIF Database

- An in-house database of reviews compiled by scanning current journals and appropriate WWW sites. Many reviews produced by the organisations listed below are included.

C. NHSCRD (WW Web access)

- DARE
- Health Technology Assessment Database
- Completed and ongoing CRD reviews

D. Health Technology Assessments and evidence based guidelines(WW Web access)

- NICE appraisals and work plans for TARs, Interventional Procedures and Guidelines programmes (NCCHTA work pages:www.ncchta.org/nice/)
- Office of Technology Assessment
- NHS Coordinating Centre for Health Technology Assessments
- Canadian Co-ordinating Office for Health Technology Assessment
- New Zealand Health Technology Assessment
- Wessex STEER Reports
- Agency for Healthcare Research and Quality (AHRQ)
- National Horizon Scanning Centre
- SIGN (Scottish Intercollegiate Guidelines Network)

E. Clinical Evidence

F. Bandolier

G. TRIP Database

H. Bibliographic databases

- Medline - systematic reviews
- Embase - systematic reviews
- Other specialist databases.

I. Contacts

- Cochrane Collaboration (via Cochrane Library)
- Regional experts, especially Pharmacy Prescribing Unit, Keele University (&MTRAC) and West Midlands Drug Information Service (url: www.ukmicentral.nhs.uk) for any enquiry involving drug products

7.2.2 Primary studies search strategies

Cochrane Library 2006 Issue 1

#1 cognition in All Fields in all products
#2 psychological next test* in All Fields in all products
#3 cognition next disorder* in All Fields in all products
#4 memory in All Fields in all products
#5 attention in All Fields in all products
#6 visual next perception in All Fields in all products
#7 cognitive next function* in All Fields in all products
#8 executive next function in All Fields in all products
#9 cognitive next impairment* in All Fields in all products
#10 cognitive next dysfunction* in All Fields in all products
#11 MeSH descriptor Cognition explode all trees in MeSH products
#12 MeSH descriptor Neuropsychological Tests explode all trees in MeSH products
#13 MeSH descriptor Cognition Disorders explode all trees in MeSH products
#14 MeSH descriptor Memory Disorders explode all trees in MeSH products
#15 (#1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10 OR #11 OR #12 OR #13 OR #14)
#16 driver* OR driving OR drive* OR road in All Fields in all products
#17 MeSH descriptor Automobile Driving explode all trees in MeSH products
#18 (#16 OR #17)
#19 (#15 AND #18)

Ovid MEDLINE(R) 1966 to February Week 4 2006
1 cognition/
2 cognition disorders/)
3 neuropsychological tests/
4 memory disorders/
5 Visual Perception/
6 cognitive dysfunction$.mp.
7 cognitive function$.mp.
8 executive function$.mp.
9 cognitive disabilit$.mp.
10 cognitive impairment.mp.
11 or/1-10

12 automobile driver examination/
13 automobile driving/
14 (driver or driving or drivers).mp.
15 or/12-14
16 11 and 15
17 limit 16 to "diagnosis (optimized)"
18 limit 16 to "diagnosis (sensitivity)"
19 from 18 keep 1-205

EMBASE 1980 to 2006 Week 09

1 exp cognition/
2 Cognitive Defect/
3 cognition/
4 NEUROPSYCHOLOGICAL TEST/
5 Memory Disorder/
6 cognitive dysfunction$.mp.
7 cognitive function$.mp.
8 executive function$.mp.
9 cognitive disabilit$.mp.
10 cognitive impairment$.mp.
11 1 or 2 or 4 or 5 or 6 or 7 or 8 or 9 or 10
12 or/2-11
13 Driving Ability/ or Car Driving/ or Driver/ or Car Driver/
14 (driver or driving or drivers).mp.
15 or/13-14
16 11 and 15
17 limit 16 to "diagnosis (optimized)"
18 limit 17 to human
19 limit 16 to (human and "diagnosis (specificity)")
20 from 19 keep 1-37

7.3 Appendix 4. Accuracy values for Schanke 2000

Schanke[11] did not report accuracy values. The authors resisted the necessity of introducing a binary reference test, instead preferring to classify some drivers as “ambiguous” regarding their fitness to drive. Similarly the authors did not decide on a cut off category for the screening test. The sensitivity and specificity shown in fig 1 for this study could only be calculated by using the authors’ data and applying binary decisions on the basis of test results. This is clearly an unsatisfactory process since decisions on cut off choices are made in full knowledge of both sets of results (introducing “incorporation bias” where the element of an index test forms an element of the reference test, or vice versa). The sensitivity and specificity values shown are almost certainly overestimates of test accuracy and should only be viewed as degree of agreement between the index test and an imperfect reference standard.

7.4 Appendix 5. Email communication 01/03/2006

From: Chris Hyde
Sent: 01 March 2006 10:03
To: [redacted]
Cc: Martin Connock; David Moore; Anne Fry-Smith; Sue Bayliss; Ann Massey; [redacted]
Subject: RE: ARIF request: What tests of cognitive function predict driving ability in psychiatric illness

Dear Malcolm, Paul, Heather and Sue

Thanks for your replies on this

Based on the feed back we'll proceed with the plan as stated below but taking into account two issues raised:

1. The need to consider not just people with dementia, but also other conditions which might impair executive function like depression and schizophrenia

2. The need to consider the feasibility of any tests or batteries of tests identified as potentially useful

Concerning the latter, I am not completely confident that the research literature will help, but there are often snippets of information which can help build up a picture of whether the tests would be acceptable in practice. Further feasibility is an area where I think clinical and patient experience may actually have as much if not more to offer than research literature. In contrast assessing how often a new test or battery of tests predicts a particular outcome defintiely does require some insight from a properly conducted piece of research.

Wll be back in touch in a months time

Please feel free to contact me again if anything interesting crops up in the interim

Best wishes

Chris Hyde

-----Original Message-----
From: Chris Hyde
Sent: 20 February 2006 16:30
To: [redacted]
Cc: Martin Connock; David Moore; Anne Fry-Smith; Sue Bayliss; Ann Massey
Subject: ARIF request: What tests of cognitive function predict driving ability in psychiatric illness

Dear Graham

Martin Connock and I have been assigned to this request, due for delivery 1/4/06

We've both had a chance to review in detail the request form and ancillary information you provided

which was excellent and pretty clear

Just wanted to check with you and provide an opportunity for your panel chair and expert to feed in if they feel necessary about our initial approach.

In my view the task is made much easier by the existing overview of by The British Psychological Society, "Fitness to drive and cognition". They set out the key issues and I can see no benefit in us doing a further general overview covering exactly the same ground, even if 5 years further on.

I would thus suggest that our task is to do something which builds specifically on the work of the The British Psychological Society.

Again I would suggest the obvious target is to identify, appraise and summarise as many test evaluations as we can of the type exemplified by Pat McKenna (The use of a cognitive battery to predict who will fail an on-road driving test. B J of Clinical Psychology 2004;43:325-336) comparing new tests or test combinations with the results of on-road driving tests. Such studies are so clearly relevant to your decision that it is essential you ensure your guidance is compatible with **all** their results, even if they do not provide the whole answer.

I'm sure it will be obvious that we won't be able to do a truly systematic review of all the test evaluations. However, I am confident we will be able to substantially extend the range of studies you will have to draw on in your decision making process (or reduce the probability that such studies exist to a very low level)

Would be very happy to discuss this further, emphasising the latest we would ideally like to leave starting our searches would be 1/3/06. A teleconference at some stage over the next week might also be very useful if a number of people wanted to contribute/challenge the suggested initial line of enquiry.

Best wishes

Chris Hyde

Senior Lecturer in Public health

University of Birmingham

[redacted]