ASSESSING FITNESS TO DRIVE: PRACTICAL TIPS ON CHOOSING THE RIGHT SCREENING TOOLS FOR YOUR PRACTICE

PRINT THIS ARTICLE

***Bruce Weaver, MSc,** assistant professor, Human Sciences Division, Northern Ontario School of Medicine; research associate, Centre for Research on Safe Driving, Lakehead University, Thunder Bay, Ontario*

***Michel Bédard, PhD,** professor, Department of Health Sciences; director, Centre for Research on Save Driving, Lakehead University*

Correspondence may be directed to bweaver@lakeheadu.ca.

Detection and reporting of findings that suggest impaired fitness to drive is a difficult challenge faced by Canadian physicians. The need for a short and valid screening test that can be used in a physician's office is pressing, and will grow as the population ages. To date, none of the proposed tests is universally accepted as adequate to meet the needs of all physicians or their patients. In this uncertain context, busy physicians may have difficulty choosing screening tools that are right for their practices. We offer some practical tips that physicians can apply when appraising both currently available screening tools and those that will be developed in future. Given the general consensus that none of the current tools is completely satisfactory, it is a virtual certainty that more tools will be proposed in the future as work continues in this area.

Seven Questions to Ask When Assessing a Screening Tool for Fitness to Drive

We present below a series of questions that should be asked when one is assessing a potential screening test for fitness to drive. We have attempted to present them in hierarchical order such that one can stop at any point if the answer to one of the questions is unsatisfactory. For example, if the gold standard measure of driving is deemed unsuitable (question 1), then there is no need to consider the remaining questions.

Question 1: How Suitable Is the Gold Standard?

The usefulness of screening test results is directly related to the suitability of the gold standard measure used while developing or validating the test. Furukawa and co-authors provide a good overview of those issues.[1] As they note, the gold standard must be independent of the screening test. To maintain independence, those who categorize patients according to screening test results must be blind to the gold standard results, and the gold standard must be applied to all patients, regardless of their scores on the screening test. Molnar et al. have reviewed the strengths and limitations of various gold standards for fitness to drive (see Table 4 of http://onlinelibrary.wiley.com/doi/10.1111/j.1532-5415.2006.00967.x/pdf).[2]

The most common gold standard for fitness to drive remains a comprehensive on-road driving test, scored by a certified driving examiner or an occupational therapist trained in driving assessment. Arguably, the best case scenario would include independent scoring by two or more examiners, with an a priori process for assessing agreement and resolving discrepancies. This ideal is not likely to be seen very frequently, however. Disadvantages of on-road testing include a lack of correlation with at-fault crashes. Some studies use gold standard measures other than an on-road test. For example, Ball and colleagues have often used crash data as the gold standard when evaluating the Useful Field of View (UFOV) test.[3–5] There are several problems with using crashes as the gold standard. First, crashes are relatively rare events. Second, not all crashes are reported – particularly minor crashes that do not involve the police. Third, and most importantly, being crash free does not imply that one is a safe driver, even if all crashes are reported. One might be crash free simply because other drivers have taken actions to avoid crashes. (This makes the aforementioned lack of correlation between on-road tests and at-fault crashes less damning for on-road tests than it might first appear.)

Another possible gold standard is performance on a driving simulator. There is increasing evidence that driving ability on a simulator is correlated with driving ability on the road.[6] Although the absolute levels of performance may differ in the two situations, those who perform better on the simulator also tend to perform better on the road, and vice versa. Testing on a simulator has some advantages. For example, one has much more control over the testing scenario than on the road in real-world ever-changing traffic conditions. Therefore, the testing scenario can be made exactly the same for every driver. Second, one can examine drivers in a range of situations, some of which would be far too challenging and dangerous to examine on the road in real traffic. In the more challenging conditions, crashes would likely occur much more frequently than in real-world driving, and thus might become a more useful gold standard measure than they usually are. Disadvantages include simulator sickness and variability of testing protocols.

Assuming the gold standard measure is suitable in a general sense, another question is whether it is suitable for the patients in your practice. For example, suppose a study uses as its gold standard an on-road test that includes several episodes of entering and exiting multi-lane divided highways with on- and off-ramps. That gold standard may be very suitable for physicians whose practices are located in densely populated areas where multi-lane highways are encountered frequently by their patients. However, it is arguably less suitable for physicians who practise in remote rural areas, because most of their older patients will rarely, if ever, encounter such highways.

Question 2: Are the Study Participants Similar Enough to Your Patients?

If the gold standard is deemed suitable, both generally and specifically for your practice, the next question is whether the study participants are similar enough to the patients in your practice for whom the test will be used. The core issue here is the same as it was for the specific suitability of the gold standard for your practice: The greater the dissimilarity of study participants to your patients, the more inappropriate it becomes to generalize the findings.

One important way in which study participants might differ from your patients is the level of uncertainty regarding fitness to drive. Obviously, physicians typically use screening tests when there is uncertainty. However, studies attempting to validate screening tests sometimes include healthy participants for whom there is little uncertainty. Although *prevalence* of the disease or condition being screened for does not systematically affect estimates of sensitivity and specificity, the so-called *spectrum of disease* (or condition) does. As Montori and colleagues put it, "The sensitivity and specificity of a test, when it is used to differentiate patients who obviously do not have the disease from patients who obviously do, likely overestimate its performance when the test is applied in a clinical context characterized by diagnostic uncertainty."[7] That overestimation is an example of *spectrum bias*.

Question 3: Are Screening Test Properties Reported?

The next question is whether screening test properties are reported – and if they are, whether they have they been reported correctly and completely. In a 2006 article, Ball and co-authors concluded that "high-risk older drivers can be identified through brief, performance-based measures [i.e., Trails B, Motor-Free Visual Perception Test, and UFOV subtest 2] administered in a [Motor Vehicle Administration] setting," but nowhere in the article did they report sensitivity, specificity, predictive values of positive and negative tests, or likelihood ratios.[8] Authors who are promoting a particular tool as a screening test must report screening test properties.[9] All of the screening test properties listed above *should* be reported; but at a minimum, in our view, sensitivity and specificity must be reported.

Another important point, which is often overlooked, is that screening test properties from a validation study are "merely estimates."[10] Therefore, confidence intervals ought to be provided, just as they are for other estimates such as means, odds ratios, and so on. Obviously, if one was considering two screening tests with very similar properties, the relative precision of the estimates for the two tests (with narrower confidence intervals indicating greater precision) would be an important factor in making a choice.

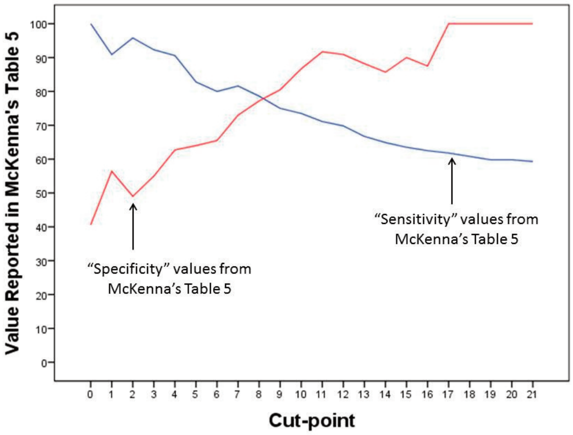

Figure 1. McKenna and colleagues report values described as sensitivity and specificity for every cut-point of their test battery; but closer examination suggests that they actually reported the predictive values of positive and negative tests.[11]

Finally, note that mistakes may occur when screening test properties are computed. For example, McKenna and colleagues report values described as sensitivity and specificity for every cut-point of their test battery; but closer examination suggests that they actually reported PV+ and PV−, the predictive values of positive and negative tests.[11] Therefore, whenever sufficient data are provided, we advise readers to check the calculations (Figure 1).

Question 4: Are the Screening Test Properties Good Enough to Make the Test Useful?

The next question is whether the screening test properties are good enough to make the test useful in your practice. This entails consideration of both the sensitivity and specificity of the test, bearing in mind the relative costs of false positives (e.g., unnecessary social isolation with risk of depression) and false negatives (e.g., the risk of a crash and the risks of injuring or killing the driver or others). Those relative costs may depend on the location of your practice. For example, false positives may be more costly for patients living in sparsely populated rural areas than for those living in densely populated urban areas where more public transportation options are available.

For tests that use two (or more) cut-offs, the percentage of indeterminate cases must also be considered: the smaller the percentage of patients who fall into the indeterminate range, the more useful the screening test. In the article by Dobbs and Schopflocher, approximately 50% of those who were screened with the SIMARD-MD (Screen for the Identification of Cognitively

Impaired Medically At-Risk Drivers, A Modification of the DemTect) fell into the indeterminate category.[12] Bearing in mind that approximately 20% of the subjects in that study were healthy controls with no suspected cognitive impairment (and therefore likely to have negative test results on the SIMARD-MD), the overall percentage with indeterminate results is likely be greater than 50% when the test is applied in a clinical context with diagnostic uncertainty.

The predictive values of positive and negative tests should also be examined. In the past, these indices have often been referred to as the *positive predictive value* (PPV) and *negative predictive value* (NPV). We prefer the terms *predictive value of a positive test* (PV+) and *predictive value of a negative test* (PV−) because they make clear that it is the test result, not the predictive value, that is either positive or negative. It is important to bear in mind that unlike sensitivity and specificity, PV+ and PV− are systematically related to prevalence: as prevalence increases, PV+ increases and PV− decreases. If prevalence is different in your practice than in the population used for the study, you will need to estimate PV+ and PV− for your setting using formulas that adjust for prevalence[13]:

$$PV+=\frac{\text{Sensitivity}\times\text{Prevalence}}{\text{Sensitivity}\times\text{Prevalence}+(1-\text{Specificity})\times(1-\text{Prevalence})}$$

$$PV-=\frac{\text{Specificity}\times(1-\text{Prevalence})}{(1-\text{Sensitivity})\times\text{Prevalence}+\text{Specificity}\times(1-\text{Prevalence})}$$

Finally, readers may notice that we have shied away from giving specific targets for sensitivity, specificity, percentage of indeterminate cases, and predictive values. We have done so because there is no statistical formula that can tell us what those target values are. This is a matter that must be discussed by all stakeholders, and society as a whole. The relative costs of false positives (falsely labelling as unsafe those who are truly fit to drive) and false negatives (failing to identify people who are truly unfit to drive) must be weighed carefully. While it is important to keep unsafe older drivers off the road, we must also bear in mind the negative impact of driving cessation on mental health and quality of life.[14] For screening tests that use two cut-points (e.g., the SIMARD-MD), we must also consider the percentage of screened individuals falling into the indeterminate category, and the costs associated with the more intensive screening required for those individuals. As noted earlier, current evidence suggests that at least 50% of individuals screened with the SIMARD-MD will fall in the indeterminate category. As suggested by recent news reports,[15] BC seniors who require further screening as a result of their indeterminate result on the SIMARD-MD are faced not only with significant financial costs, but also with stress due to the possible loss of their license.

Question 5: Have the Test's Properties Been Independently Confirmed?

Another important question to ask when appraising a screening tool for fitness to drive is whether its properties have been independently confirmed. Independent replication and reproducibility of results are core features of the scientific method; but, unfortunately, they often seem to be more honoured in the breach than the observance. On the one hand, editors and reviewers may be reluctant to publish a study that is seen as *just a replication* of something that is *already known*. But even when unsuccessful replications are published, they may be ignored in favour of the one or two positive studies that captured people's imaginations initially. As Ioannidis says, "In some areas, the prevailing mentality until now has been to focus on isolated discoveries by single teams and interpret research experiments in isolation." He continues, "It is misleading to emphasize the statistically significant findings of any single team," and that instead, we must focus on the "totality of the evidence."[16]

Those comments from Ioannidis are certainly consistent with the principles of knowledge translation (KT) laid out by the Canadian Institutes of Health Research (see http://www.cihr-irsc.gc.ca/e/39033.html). At the core of the knowledge-to-action process is the knowledge *funnel* (see http://www.cihr-irsc.gc.ca/images/knowledge_to_action_e.jpg). The knowledge funnel illustrates and emphasizes the need for increasing distillation of research findings before they are ready for use in applied settings. Importantly, in order for meaningful synthesis to occur at the second level of the funnel, there must be a sufficient number of independent and methodologically sound studies in the Knowledge Inquiry section at the top of the funnel. If the number of independent and methodologically sound studies is insufficient, widespread application of the knowledge in applied settings is not warranted. Lest readers think this would never be done, the adoption of the SIMARD-MD by the Canadian province of British Columbia provides one recent example. When British Columbia adopted the SIMARD-MD to screen for cognitive impairment or dementia that might affect fitness to drive, there was only one study in the Knowledge Inquiry section of the knowledge funnel,[12] and concerns have been raised regarding the methodology of that study.[17,18]

Even after the successful development of a tool (such as a screening test) through proper application of the knowledge-to-action process, there should be continued dialogue between users of the tool and the research community. That dialogue is often referred to as *knowledge exchange*. Ideally, information that is fed back to researchers from users of the tool can lead to refinements and improvements.

In summary, if a tool has been developed without any regard for the knowledge-to-action process, and if there is no independent confirmation of its properties, let the user beware.

Question 6: Is There Any Conflict of Interest?

Question 5 stressed the importance of independent replication and reproducibility of the results supporting the use of a screening tool. Independent replication and reproducibility become even more important when financial or other interests come into play. As Ioannidis puts it, "The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true."[16] Therefore, when appraising a screening tool for fitness to drive, it is essential to ask if there is any real or apparent conflict of interest, and who stands to profit if a screening process is adopted by a government or other organization.

Once again, British Columbia's adoption of the SIMARD-MD provides a relevant example. In that province, the SIMARD-MD is administered to any drivers suspected of cognitive impairment or dementia that might affect their fitness to drive. The SIMARD-MD uses two cut-points. Those who score in the mid-range (31–70) are required to undergo further assessment to be carried out by DriveABLE, a private for-profit company. British Columbia adopted the SIMARD-MD on the strength of one article by Dobbs and Schopflocher.[12] What may not be obvious to all readers of that article is the relationship between its first author and DriveABLE. The following is a quotation from the "Declaration of Conflicting Interests" section of the article: "The CEO and President of DriveABLE™ Assessment Centres, Dr Allen Dobbs, is the spouse of the first author (B.D.). B.D. has no shares in or financial relationship to DriveABLE™ Assessment Centres. Dr Allen Dobbs was not involved in this research. D.S. declares no competing interests." Readers will need to judge whether this represents a possible conflict of interest.

Question 7: How Acceptable Would the Test Be to Your Patients?

The UK National Screening Committee (http://www.screening.nhs.uk/) lists several criteria that should be met before a screening program is launched. One of those criteria is that the test should be "acceptable to the population." Acceptability of the test is clearly an important consideration when screening for fitness to drive. As Dalchow and co-authors argue, "Clients may be more motivated to perform tests that they believe are actually related to the driving task. Thus, scores on these types of tests may be more reflective of actual performance. Administering tests such as those from the Neuropsychological Assessment Battery (e.g., Driving Scenes, Map Reading) with colored stimuli may also maintain interest. Finally, clients may be more willing to accept 'failed' test results if the test is perceived to directly relate to driving abilities."[19]

Dalchow et al. focus exclusively on the face validity of screening tests for fitness to drive – in fact, they treat the terms *face validity* and *acceptability to clients* as synonymous. While face validity is a very important component of a test's acceptability to patients, it is not the only factor. Cost to patients is also important. This can refer not only to financial costs but also to other types of costs – for example, how much time and effort are required to complete the test, or how much embarrassment is entailed in answering incorrectly.

Summary

Assessment of older patients' fitness to drive is one of the more difficult challenges faced by physicians. Many screening tests have been proposed, but to date, none of them has proved universally acceptable in meeting the needs of physicians or their patients. We have provided a list of seven questions that physicians should ask about any screening procedure they are considering for use in their practice:

1. How suitable is the gold standard?
2. Are the study participants similar enough to your patients?
3. Are screening test properties reported?
4. Are the screening test properties good enough to make the test useful?
5. Have the test's properties been independently confirmed?
6. Is there any conflict of interest?
7. How acceptable would the test be to your patients?

The questions are arranged in hierarchical order such that one can stop at any point when the answer to a question is unsatisfactory. We hope that these practical tips will be useful to physicians as they continue to wrestle with this difficult issue.

Acknowledgements

The idea for this article was spawned during discussions with Drs. Frank Molnar and Malcolm Man-Son-Hing. Beyond participating in the original discussion, Dr. Man-Son-Hing also commented on earlier drafts of this article, and we thank him for that. We also thank an anonymous reviewer for many helpful suggestions on how to make the article more accessible and useful to physicians.

Dr. Michel Bédard is a member of the Candrive research team (www.candrive.ca). Candrive is conducting a large multi-site study to try to identify screening tools for unsafe driving. This could be perceived by some as a potential conflict of interest as it might unconsciously influence our interpretations of current proposed screening tools. Being cognizant of this concern, we have made every effort to remain as balanced and unbiased as possible in our interpretations.

This article was peer reviewed.

References

1. Furukawa TA, Strauss S, Bucher HC, Guyatt G. Diagnostic tests. In: Guyatt G, Rennie D, Meade MO, Cook DJ, eds. Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice, 2nd edition. New York: The McGraw-Hill Companies, Inc.; 2008: 419–38.
2. Molnar FJ, Patel A, Marshall SC, et al. Clinical utility of office-based cognitive predictors of fitness to drive in persons with dementia: a systematic review. J Am Geriatr Soc 2006;54(12):1809–24.
3. Ball K, Owsley C, Sloane ME, et al. Visual attention problems as a predictor of vehicle crashes in older drivers. Invest Ophthalmol Vis Sci 1993;34:3110–23.
4. Owsley C, Ball K, Sloane ME, et al. Visual/cognitive correlates of vehicle accidents in older drivers. Psychol Aging 1991;6:403–15.
5. Rizzo M, Reinach S, McGehee D, Dawson JD. Simulated car

Key Points

- *Detection and reporting of findings that suggest impaired fitness to drive is a difficult challenge faced by Canadian physicians.*
- *The need for a short and valid screening test that can be used in a physician's office is pressing, and will grow as the population ages.*
- *The usefulness of screening test results is directly related to the suitability of the gold standard measure used while developing or validating the test.*
- *While it is important to keep unsafe older drivers off the road, we must also bear in mind the negative impact of driving cessation on mental health and quality of life.*

crashes and crash predictors in drivers with alzheimer disease. Arch Neurol 1997;54:545–51.

6. Mullen N, Charlton J, Devlin A, Bédard M. Simulator validity: behaviors observed on the simulator and on the road. In: Fisher DL, Rizzo M, Caird JK, Lee JD, eds. Handbook of driving simulation for engineering, medicine, and psychology. Boca Raton (FL): CRC Press; 2011:13–18.
7. Montori VM, Wyer P, Newman TB, et al.; Evidence-Based Medicine Teaching Tips Working Group. Tips for learners of evidence-based medicine: 5. The effect of spectrum of disease on the performance of diagnostic tests. Can Med Assoc J 2005;173(4):385–90.
8. Ball KK, Roenker DL, Wadley VG, et al. Can high-risk older drivers be identified through performance-based measures in a department of motor vehicles setting? J Am Geriatr Soc 2006;54(1):77–84.
9. Bedard M, Weaver B, Darzins P, Porter MM. Predicting driving performance in older adults: we are not there yet! Traffic Inj Prev 2008;9(4):336–41.
10. Harper R, Reeves B. Reporting of precision of estimates for diagnostic accuracy: a review. BMJ 1999;318(7194):1322–3; http://www.bmj.com/content/318/7194/1322.1.full.
11. McKenna P, Jefferies L, Dobson A, Frude N. The use of a cognitive battery to predict who will fail an on-road driving test. Br J Clin Psychol 2004;43:325–36.
12. Dobbs BM, Schopflocher D. The introduction of a new screening tool for the identification of cognitively impaired medically at-risk drivers: the SIMARD. A modification of the DemTect. J Prim Care Comm Health 2010;1(2):119–27.
13. Altman DG, Bland JM. Diagnostic tests 2: predictive values. BMJ 1994;309(6947):102; http://www.bmj.com/content/309/6947/102.1.pdf%2Bhtml.
14. Mullen N, Bédard M. The end of driving. In: Odell M, editor. Older Road Users: Myths and Realities; A Guide for Medical and Legal Professionals. Tucson (AZ): Lawyers & Judges; 2009:281–92.
15. CBC News. Access to DriveABLE test concerns elderly BC drivers. Vancouver (BC): CBC Vancouver, 2012 Mar 16; http://www.cbc.ca/news/canada/british-columbia/story/2012/03/16/bc-elderly-driveable-test.html.
16. Ioannidis JPA. Why most published research findings are false. PLoS Med 2005;2(8):e124; http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.0020124.
17. Bédard M, Weaver B, Man-Son-Hing M, et al. The SIMARD screening tool to identify unfit drivers: are we there now? J Prim Care Comm Health 2011;2(2):133–5.
18. Hogan DB, Bédard M. Papers that might change your practice: review of the introduction of a new screening tool for the identification of cognitively impaired medically at-risk drivers. Can Geriatr J 2011;14(2):51–4; http://cgjonline.ca/index.php/cgj/article/view/12/31.
19. Dalchow JL, Niewoehner PM, Henderson RR, Carr DB. Test acceptability and confidence levels in older adults referred for fitness-to-drive evaluations. Am J Occup Ther 2010;64(2):252–8.