Author's response to reviews

Title: Recognizing Acute Delirium As part of your Routine [RADAR]: A validation study

Authors:

Philippe Voyer (philippe.voyer@fsi.ulaval.ca)
Nathalie Champoux (nathalie.champoux@umontreal.ca)
Johanne Desrosiers (johanne.desrosiers@usherbrooke.ca)
Philippe Landreville (philippe.landreville@psy.ulaval.ca)
Jane McCusker (jane.mccusker@mcgill.ca)
Johanne Monette (johanne.monette@mcgill.ca)
Maryse Savoie (maryse.savoie@vac-acc.gouv.qc.ca)
Sylvie Richard (sylvie.richard@fsi.ulaval.ca)
Pierre-Hugues Carmichael (pierre-hugues.carmichael.cha@ssss.gouv.qc.ca)

Version: 5 Date: 10 February 2015

Author’s response to reviews: see over
The authors would like to thank the editor and reviewers for their comments that helped improve the manuscript. For details, please refer to the responses: reviewers’ comments are in italics, responses are in normal and changes to the manuscript are highlighted in yellow in the manuscript itself.

**Reviewer #1:** Judith Dyson

**Comment #1:**
There was one major compulsory issue for me which needs addressing prior to publication. The authors conclude that “RADAR is efficient reliable sensitive” (line 533). This is poorly demonstrated by the statistics cited. For example, line 248 identifies a Kappa ranging from 0.3 indeed this is recognised in lines 365 onward. Is inter-rater reliability insufficient therefore?

**Our reply:**
The sensitivity is very good when compared to the sensitivities of other studies when a test is administered by clinical staff as opposed to research staff. This point is explained in lines 482 to 488.

Concerning the inter-rater agreement, the authors regarded it as very good because the percentages of agreement are high (82 to 98%, line 368) and based on numerous comparisons (52% of all RADAR administration) [Table 3]. The explanation for the lower Kappa can be found in lines 365 to 371 of the manuscript.

**Comment #2:**
The agreement between RADAR and CAM seems low at times (line 374 and table 4).

**Our reply:**
This result is expected since each item of the CAM is scored based on the CAM item assessment supported by the HDS tool. Moreover, the CAM items were tested three times during the day shift. In comparison, during this same time period, the RADAR was completed one to four times. In other words, we compared the level of agreement between a 50-minute assessment (three CAM+HDS) with the RADAR, a 7 to 28 seconds assessment. We also examined whether the number of CAM assessments was related to the percentage of agreement. As shown in Table 4, one CAM administration was more in agreement with the RADAR than were three CAM administrations. With delirium symptoms fluctuation, it is widely known that the more you observe the patient, the more likely it is that you will be able to detect it (Zou et al., 1998). Our result shown in Table 4 corroborates this last-mentioned study.

Comment #3:
PPV is demonstrated to be low also (line 392 and table 5). Acknowledgement of high levels of false-positive (line 497). Yet at times the discussion suggests values are in accordance with general expectations (line 473). My opinion is that this needs a little bit of unpicking. Are these values sufficient and if so according to what parameters? I was left with the impression by the end of the paper that the RADAR is quick and easy (just seven seconds to complete) but that it doesn’t really measure what it is designed to measure. It suspect that some of this incongruence can be addressed by citing usual expectations of the parameters of the tests conducted within the context they have been conducted and support with references. It may be that the concluding remark needs to be more tentative and suggest further work to be done.

Our reply:
In the Conclusion (line 533), when we refer to the efficiency of the RADAR efficiency, we refer to its 7-second administration time. When we refer to its reliability, we refer to its percentage of agreements (line 363) based on 50% of RADAR administrations (Table 3). When we report on its sensitivity, we compare our 73% of sensitivity with sensitivities reported in the manuscript (line 484).

With regard to the PPV, it was indeed a matter of concern. This is why we undertook further analysis which demonstrated that in fact, 100% of the participants with a RADAR positive had an acute cognitive impairment (see lines 401 to 406).

Comment #4:
There were a number of small errors and questions begged for me - discretionary revisions. Examples include; DSM in full the first time and referenced.

Our reply:
It has been corrected.

Comment #5:
There seems to be an inaccuracy between lines 76 and 77 where the authors write diagnostic accuracy less than adequate and a specificity of 98 to 100% is cited (which appears perfectly adequate).

Our reply:
The term has been changed to “sensitivity” (line 75).

Comment #6:
I was a bit confused by the barrier of time being cited as a reason for the CAM being inadequate in practice and by way of justifying the RADAR. The authors cite two SRS recommending tools that require administration of 5 to 10 minutes (lines 84-85); earlier (line 69) this is exactly the time the CAM takes. If this is considered too lengthy (despite recommendations) this needs to be clear and the basis of this assertion made explicit.

Our reply:
The CAM takes 5 to 10 minutes (Wong et al., 2010). This text on the CAM is from the training manual:
The CAM is usually rated by a clinical or trained lay interviewer on the basis of an interview with the patient that includes at least a brief cognitive assessment. The Mini Mental State Examination was used in the original validation, but its use is now restricted by copyright law. A more brief assessment, the Modified Mini-Cog Test (Pg. 25) is recommended for quick screening. Generally, the entire CAM rating takes 5-10 minutes to complete.


The point we made from lines 87 to 93 is that, although 5 minutes seems a short time, it takes a significant toll on nursing time. This is why we created the RADAR.

A recent paper from the author of CAM lends support to this point (Marcantonio et al, 2014):

The Confusion Assessment Method (CAM), developed in 1990 (4), has been widely adopted. A recent comparison of diagnostic methods suggests that the CAM is the best-performing bedside delirium assessment tool (5). Although the CAM is widely used in the literature to define delirium (6), it can be challenging to operationalize in the clinical setting because it requires cognitive assessment and substantial interviewer training. Moreover, application of the CAM varies greatly, which can lead to differential performance in detecting delirium (5).


Wong et al. (2010). Does this patient have delirium? Value of bedside instruments. JAMA. 2010;304(7):779-786

Comment #7:
I am not familiar with the role “licenced practical nurse” (line 140) suggest brief definition.

Our reply:
A definition has been added.

Comment #8:
It would be useful to identify the clinical role of the research assistant (192) as they are making clinical judgements.

Our reply:
Details have been added in the Overview of Data Collection and Measures section (lines 210 to 212).

Comment #9:
For this validation study people with a history of psychiatric illness were excluded. Were those with learning disabilities included or excluded in the study?
Our reply:
Yes, they too were excluded. This information has been added in the text (line 181).

Comment #10:
I see the tool was designed for (line 154) and tested with only nurses. Is this a tool that can be used by other healthcare practitioners?

Our reply:
In this study the RADAR was not tested with staff other than nursing staff. However, we will shortly begin two new studies in which we will validate the RADAR when administered by family members at home and by nursing aides in residential care settings.

Comment #11:
Observation of the patient during medicine round informs the scoring of the RADAR. How is it applied if the patient does not need medicines (at all or at one of the times the test is to be completed)?

Our reply:
The RADAR has been validated for administration after giving medications to a patient. If the patient does not take any medication, the value of the RADAR is limited. This point has been added in the limitations section of the Discussion (line 542).

Comment #12:
Given that delirium requires psychiatrist diagnosis, how is the tool to be used? I suspect to monitor change but this was not explicit within the text.

Our reply:
The RADAR is not intended to replace the psychiatrist’s diagnosis. The RADAR is proposed as a screening tool only.

Reviewer #2: Claudia Spies
Comment #13:
The length of the sections should be reconsidered. The introduction is very detailed and epic (=>1,000 words (!)). We suggest a clear and short introduction that outlines the problem and brings up a clear hypothesis and research question.

Our reply:
Based on the word counts, the Introduction can certainly be considered long in comparison to other articles. We wanted to explain in detail why we created the RADAR. There are many very good screening instruments that have been published. We wanted to have the chance to illustrate what we view as the limitations in those instruments. Every instrument has strengths, but also its limitations (as does the RADAR). The Introduction allowed us to describe and justify these limitations. Although the Introduction is long, there are no repetitions. We believe the Introduction is fine as is. We appreciate the flexibility that an online journal like BMC Nursing provides and that it allows different lengths for the Introduction.
Comment #14:
As the RADAR-development is part of the methodology, it should be considered to put this whole part into the methods section.

Our reply:
Our reply is to say that we understand the point made by Dr Spies and we even thought about doing this. In the end, we decided otherwise because doing so would have made the manuscript considerably longer. Moreover, we wanted to stay focused on the latest version of the RADAR for which we have numerous results to present plus a large sample size.

We are of the opinion that making a text about the earlier steps of the RADAR available in an "additional file" makes for the perfect compromise. Once again, having these options is definitely something we appreciate about an online journal such as BMC Nursing.

In any event, the earlier steps have been presented at international meetings:


Comment #15:
The authors should be very precise about the facts they use to raise their clinical question: which authors/publications/references support the fact that the CAM/CAM-ICU takes 5-10 minutes in routine? Which other tests are there available? We would suggest a systematic literature search that addresses this question for the commonly used screening tools. The authors should exactly quote which references underline that the CAM/CAM-ICU takes 5 to 10 minutes. Especially on the ICU the CAM-ICU is popular because of its’ fast applicability.

Our reply:
For the 5-10 minutes, we would refer Dr Spies to our response to Comment #6. In the manuscript, whenever we mention the time of 5 minutes, it always refers to the CAM [lines 69], not the CAM-ICU (since this latter is used only in ICU).
Comment #16:
*The other general issue is the reference-standard. It is understandable that research assistants did the CAM as a reference standard but the authors use DSM-IV-TR in their results as their reference. The DSM-IV-TR cannot be put on one level with the CAM as it is a test for delirium detection in clinical-routine as well.*

Our reply:
Regarding the CAM being a detection tool for clinical routine, we refer Dr Spies to our response to comment #6.

As for no putting the DSM-IV-TR on the same level as the CAM, it has been shown that when the CAM is administered several times during a shift, it is more sensitive than a diagnosis made by a psychiatrist (Zou et al., 1998).


We collected data about delirium symptoms with the CAM and the HDS, which are two validated tools. Afterwards, when all symptoms of delirium are rigorously assessed and entered into a database, it is a matter of applying sets of criteria (see references below). However, as mentioned in the limitation section (line 535), the cause of delirium was not determined in this study.


Comment #17:
Problems that evolve when solely using the CAM usually refer to the amount of training that is necessary to perform it with a high validity (especially in routine) and the problem that the CAM only allows a yes/no decision regarding delirium, whereas other scores (ICDSC, Nu-DESC etc.) also allow the detection of “subsyndromal” states (patients that have clinical signs for delirium do not fulfill all necessary criteria for delirium). Why do the authors only refer to the CAM?

Our reply:
We refer only to the CAM because the conclusion of systematic literature reviews indicated that this is the most reliable, validated and commonly-used tool for delirium detection worldwide. We also agree with this conclusion. Plus the other tools have limitations as well, that we want to overcome and that were described in the Introduction (e.g. ICDSC is validated for use in ICU only).


Comment #18:
The methodology is extremely detailed and might be more focused on the clear research questions that have been phrased in the introduction PLUS the development of the score (which is unknown to the potential reader).

Our reply:
In lines 164 to 166 at the end of the Introduction, we have added a sentence to make the text clearer and provide a better transition to the Methods section. We now inform the readers about the reliability and validity test that will be performed. We believe this will make the reading of the Methods easier.

As to the score of the RADAR, it is specified in lines 234-235.
Comment #19:
Patient enrolment: We appreciate the author's efforts to explain their inclusion process but it might increase the intelligibility to put this to a supplementary file and focus on the relevant points: consent was given either by the patient or his/her legal representative (if necessary).

Our reply:
Since we recruited a vulnerable group of patients with dementia, we prefer to be very clear about the method we use for enrolment. We like to be specific about this in the same way as we were in our previously published articles.


Comment #20:
The validation of a new test usually requires testing against the reference-standard. In case of delirium these are the DSM criteria (currently in the 5th edition). The CAM is a screening test as well. The authors should carefully describe why they have chosen the CAM as a reference standard for their setting although they doubt its generalizability in the introduction.

Our reply:
Regarding the reference standard, we invite Dr Spies to read our response to Comment #16. We would like also to point out that we do not question the generalizability of the CAM. In fact, we only mention positive things about the CAM except that it is too long for use on a daily basis.

As for the DSM-V comment, at the time we wrote our research proposal, the DSM-V was not yet published. Besides, it is usually not recommended to change variables, research questions, or research method during a project. Nonetheless, we did further analysis and when applying the criteria of DSM-V to our sample, we ended up having exactly the same number of delirium with the exact same participants. In other words, DSM-V criteria have not changed the results presented in Table 6.

This information has been added to the Limitations section of the Discussion (lines 538 to 541).
Comment #21:
Furthermore it is absolutely correct that the CAM has been designed as a structured test for checking DSM-III (three) (the authors write that on p. 7; l. 235) but afterwards the manuscript always refers to DSM-IV-text-revised (four, text revised). This is important for the entire manuscript: the authors use the CAM as the reference-standard NOT the DSM-IV-TR criteria. Only, a psychiatrist or special trained research staff should use these criteria.

Our reply:
We refer Dr Spies to our response to Comment #16.

Comment #22:
The authors should explain the sentence “During the course of the study, RAs assessed samples of patients and residents simultaneously and independently for inter-rater reliability (n = 37/193; 19% of total sample). Percentage of agreement varied from 72% to 100%, while kappa values ranged from 0.30 to 1, depending of the item tested” (p.7, ll. 246-248) in detail: how was the kappa-agreement calculated? Do the authors refer to different CAM-domains?

Our reply:
The sentence has been clarified in the text (lines 250 to 253). It now reads:

During the course of the study, RAs assessed a sample of patients (n = 37/193; 19% of total sample) simultaneously and independently to check inter-rater reliability on the CAM. Percentage of agreement on each item of the CAM varied from 72% to 100%, while kappa values ranged from 0.30 to 1.

Comment #23:
The authors mention the cut-off (1/3 RADAR-items positive) in the section “statistical analysis” for the first time. Did they try other cut-offs as well? Why not opening a section where RADAR is explained in detail – why this important information to the supplementary file? (in short: how have items been chosen, how is it applied, what is the cut-off)

Our reply:
With only one item rated positive, RADAR sensitivity is 73%. If we apply a more restrictive definition of what a positive RADAR is, the sensitivity would decline. Had we had a 100% sensitivity, we could have looked at other cut-offs. But as the results stand, further tests were not deemed necessary. For instance, if we had stipulated that 2-items of the RADAR had to be checked in order to have a RADAR+, these would be the results:
Table 6. Concurrent validity of RADAR+ (two items scored +) compared with a DSM-IV-TR criterion-defined delirium according to the number of daily RADAR administrations.

<table>
<thead>
<tr>
<th>Number of RADAR administrations (N)</th>
<th>Sensitivity % [95% CI]</th>
<th>Specificity % [95% CI]</th>
<th>Positive Predictive value % [95% CI]</th>
<th>Negative Predictive value % [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to 4 (193)</td>
<td>34.8 [16.4-57.3]</td>
<td>89.4 [83.8-93.6]</td>
<td>30.8 [14.3-51.8]</td>
<td>91.0 [76.8-87.9]</td>
</tr>
<tr>
<td>1-2 (80)</td>
<td>33.3 [9.9-65.1]</td>
<td>94 [85.6-98.4]</td>
<td>50.0 [15.7-84.3]</td>
<td>88.9 [79.3-95.1]</td>
</tr>
<tr>
<td>3-4 (113)</td>
<td>36.4 [10.9-69.2]</td>
<td>86.3 [78.0-92.3]</td>
<td>22.2 [6.4-47.6]</td>
<td>92.6 [85.4-97.0]</td>
</tr>
</tbody>
</table>

CI: Confidence interval
Comment #24:
The results section should be more focused on the main results. In general the length of the result section seems appropriate but the authors validated a score: The validation-data itself seem somewhere lost in the rest of information. Which statistical test did the authors use for repeated measurements?

Our reply:
The results section with its many subheadings appears clear to us. Unfortunately, we do not understand the comment about repeated measurements.

Comment #25:
- The authors should also mention hyperactive and mixed delirium as RADAR items solely refer to hypoactive states.

Our reply:
This is a very important comment and we are delighted that Dr Spies asked it. The question of hyperactive delirium was inadvertently omitted in our manuscript and we very much regret that. The fact is that the second item of the RADAR captures not only “inattention” but also “agitation”. In the earlier version of RADAR there was a separate item related to agitation, but it did not increase the sensitivity of the scale. After some digging into our database and further analysis, we realised that it was the second item of the RADAR that captured best both the agitation and hyperactivity of the participants. We have now included these results in Table 4.
### Table 4. Convergent validity: Item of RADAR vs. corresponding CAM item

<table>
<thead>
<tr>
<th>RADAR items</th>
<th>CAM</th>
<th>Number of CAM assessments on which the CAM item is rated</th>
<th>% of agreement</th>
<th>Kappa [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item #1</td>
<td>CAM 4</td>
<td>1</td>
<td>83.9 (162/193)</td>
<td>0.36 [0.19-0.53]</td>
</tr>
<tr>
<td>“…was the patient drowsy?”</td>
<td>Level of consciousness</td>
<td>2</td>
<td>76.7 (148/193)</td>
<td>0.25 [0.11-0.39]</td>
</tr>
<tr>
<td></td>
<td>Hypoalert</td>
<td>3</td>
<td>65.8 (127/193)</td>
<td>0.15 [0.05-0.26]</td>
</tr>
<tr>
<td>Item #2</td>
<td>CAM 2</td>
<td>1</td>
<td>51.8 (100/193)</td>
<td>0.08 [0.03-0.12]</td>
</tr>
<tr>
<td>“…trouble following your instructions?”</td>
<td>Inattention</td>
<td>2</td>
<td>43.5 (84/193)</td>
<td>0.08 [0.03-0.12]</td>
</tr>
<tr>
<td></td>
<td>CAM 4 and 8</td>
<td>1</td>
<td>85 (164/193)</td>
<td>0.16 [-0.02-0.33]</td>
</tr>
<tr>
<td>“…trouble following your instructions?”</td>
<td>Hyper- alert/agitation</td>
<td>2</td>
<td>85 (164/193)</td>
<td>0.16 [-0.02-0.33]</td>
</tr>
<tr>
<td>Item #3</td>
<td>CAM 9</td>
<td>1</td>
<td>76.6 (147/192)</td>
<td>0.40 [0.25-0.54]</td>
</tr>
</tbody>
</table>
“…movements slowed down?”

<table>
<thead>
<tr>
<th>Psychomotor retardation</th>
<th>2</th>
<th>76.0 (146/192)</th>
<th>0.43 [0.29-0.57]</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td>75.0 (144/192)</td>
<td>0.42 [0.28-0.56]</td>
</tr>
</tbody>
</table>

CAM: Confusion Assessment Method
Comment #26:
- The authors should carefully discuss the 7-second result. The authors reveal that RADAR takes 7 seconds. How can that be the case? You should also mention the time that is necessary to gather the information to answer the questions? The 7 seconds seem rather the time it takes to make three crosses on a medical chart than the time it takes to acquire the clinical information that is necessary to answer these questions. We ask the authors to discuss this carefully.

Our reply:
The items of the RADAR are answered after the nurse has handed out the medication to the patient. During this interaction, the nurse simply has to pay attention to the patient, no other tasks are demanded of the nurse (see lines 156 to 162). And yes, it does take only 7 seconds on average to score the RADAR.

Comment #27:
- The authors should discuss why they did not use the reference-standard (either DSM-IV or 5) but the CAM for validation (it is not the same!).

Our reply:
Dr Spies is invited to read our responses to Comments #16 and #20

Comment #28:
- The authors did mention neither the testing for psychometric properties of the items, their development nor the evaluation of their cut-off. How did they manage that? Please, insert this into the main text. This should not be mentioned exclusively in the supplementary files.

Our reply:
Dr Spies is invited to read our responses to Comments #14 and #23

Comment #29:
- How sustainable was the “45-minutes-training-package” the authors mentioned?

Our reply:
The 45-minute training package went quite well. As indicated in Table 9, 98% or more of the nursing staff agreed with these statements:

- The knowledge provided in the RADAR training package was sufficient for me to feel competent using the tool.
- I have sufficient knowledge to be able to answer the RADAR items.