PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Validation of prescribing appropriateness criteria for older Australians using the RAND/UCLA Appropriateness Method
AUTHORS	Basger, Benjamin ; Chen, Timothy; Moles, Rebekah

VERSION 1 - REVIEW

REVIEWER	Dr Tobias Dreischulte NHS Tayside c/o University of Dundee Dundee DD3 6TU Scotland, UK I declare that I have no competing interests.
REVIEW RETURNED	27-Jun-2012

GENERAL COMMENTS	**1. Introduction and research question:** • It is not clear, how 'adding recommendations for co-morbidity and the oldest old' would advance or improve the original set developed by the authors or other published criteria sets, such as START/STOPP. If the original criteria set was described in a bit more detail, ideally contrasting the set with other well established published sets, this may allow the reader to better put this research into perspective. • More clarity regarding the (primary) intended use of the criteria would help to better understand the potential relevance of the work to the Australian health care setting and beyond. Although the 'detection of DRPs as part of the Australian medication review process' is identified as the purpose of the original set, the discussion proposes multiple additional purposes of the new set under 'intended use'. Different purposes (e.g. prospective use in clinical decision support versus retrospective large scale application for performance management) pose different demand s on the resulting criteria set (e.g. with respect to the specificity /explicitness of criteria). The extent to which the resulting criteria set is suited to support the primary intended purpose could then serve as a focus in the discussion. • Related to the previous point, it would add clarity if the authors could indicate from the start, whether the intention is to develop explicit criteria (that would then be supplemented by clinical judgement in the review process) or implicit criteria.

	2. Inclusion criteria: It is not clear what is meant by the 'oldest old'. A brief description of the conditions and drugs covered by the original criteria set would be helpful. **3. Description of methods:** • Page 3, line 22-24: It is not clear what is meant by 'cross-reference'. This should be briefly explained and reference be made to the original criteria set (although this is done two sentences later). 'Most common' reasons should also be more clearly defined - was there a threshold for commonness? The literature review is an important part of the RAND appropriateness method. The relatively low number of criteria presented to the expert panel suggests considerable pre-selection of criteria. It would therefore be desirable if the literature review process and the pre-selection of criteria by the research team could be made more transparent. • Page 3, line 39: An international audience may be less familiar with the term 'hostel' in a health care context. • Page 3, lines 40-43: The context in which 'data collection' should be 'feasible' is not clear, i.e. prospective/retrospective, with or without medical notes etc.. Please, also see under research question above. • Page 3, lines 49: It is not clear who is to be 'comprehensively represented'. • Page 4, lines 22-26: A brief description of how the median and inter-percentile range was used to assess agreement at this stage (rather than in round 2) would be helpful. **4**. **Outcome measures** • For criteria without disagreement, it is not clear at which threshold (median appropriateness rating) were criteria retained/discarded? **5. Results** • Page 5, line 7: The sentence starts with 'Of criteria with disagreement....', but then provides the total number of criteria scored in the denominator in the brackets. This can be a source of confusion. Table 2: • Given that the set is focussed on conditions most commonly seen in the elderly, it is surprising that asthma features in the set

	6. Discussion • A comparison of strengths and limitation in comparison to already existing tools is missing and would be desirable. This is particularly relevant as it appears that some criteria for commonly used drugs that commonly cause harm in the elderly are missing from the set., such as high-risk use of NSAIDs in patients with risk factors for GI bleeding (without gastroprotection). • An expansion on how specifically the criteria set may facilitate the 'prioritisation of therapeutic goals' would be desirable. The specific work that will be required before the set can be implemented for the various proposed purposes should be specified. • The RAM has traditionally been used and tested in order to develop explicit assessment criteria and to establish thresholds for appropriateness in an average patient (e.g. at which threshold for MI risk would coronary intervention be warranted). For this purpose, experts are offered several potential thresholds (e.g. different thresholds of cardiovascular risk), which are then scored by each panellist for the extent to which performing the procedure at the proposed threshold is appropriate or not. It seems that the present study departs from the original method in that a group consensus was sought first (either by pre-selection (before round 1) or discussion (before round 2)) and an amended criterion was then scored again. A further difference is that panellists in this study were not asked to rate the appropriateness of performing or not performing a specific medical process in a certain clinical situation, but rather on the appropriateness of following the criterion. If the criterion states 'achievement of blood pressure as a appropriate' (as in criterion 1), the resulting criterion is judgement-based rather than explicit and there appears to be little room for disagreement. These (and any other) departures from the original RAM should be acknowledged and any implications discussed. **7. Article focus/Key messages/ Strengths and limitations** *Article focus* Page 1, line 15-16: The primary intended use of the criteria should be described *Strengths and limitations* • The claim that the criteria address established evidence based practice gaps is mentioned only marginally in the main text. If this is a key message, it should be demonstrated in the text using key examples • A comparison of strengths and limitation in comparison to already existing tools is missing and would be desirable (see under discussion above)

REVIEWER	Dr. Paul Gallagher PhD FRCPI Consultant Physician in Geriatric Medicine Cork University Hospital, Wilton, Cork, Ireland.
REVIEW RETURNED	05-Jul-2012

THE STUDY	The statistical methods should be described in more detail. The likert scale used to rate appropriateness should be specified (It is vaguely presented as 1 = highly inappropriate and 9 = highly appropriate). The authors state that the median value with dispersion values (IPR and IPRAS) were calculated for each criterion. Which percentiles were included in the IPR? Were these quartiles? The authors state that the three-point region containing the median was used to determine agreement. What was the value of the median below which a criterion was judged not to be valid? A median of 5 would suggest that half of all panellists ratings were below 5 and half were above 5. What was the threshold above which a criterion was accepted? This requires clarification. A schematic representation of one example of median rating with IPR (boxplot) and IPRAS would be useful to illustrate the distribution of agreement amongst panellists according to the Likert scale. The authors should comment on whether or not the absence of three of the panellists from the second round influenced the rating scales in the second round. The references are accurate but the authors have failed to mention the latest publication of Beers' criteria (JAGS May 2012). These are the most frequently cited explicit criteria to date and have recently been revised and revalidated and should be referred to.
RESULTS & CONCLUSIONS	The statistical methods should be described in more detail. The likert scale used to rate appropriateness should be specified (It is vaguely presented as 1 = highly inappropriate and 9 = highly appropriate). The authors state that the median value with dispersion values (IPR and IPRAS) were calculated for each criterion. Which percentiles were included in the IPR? Were these quartiles? The authors state that the three-point region containing the median was used to determine agreement. What was the value of the median below which a criterion was judged not to be valid? A median of 5 would suggest that half of all panellists ratings were below 5 and half were above 5. What was the threshold above which a criterion was accepted? This requires clarification. A schematic representation of one example of median rating with IPR (boxplot) and IPRAS would be useful to illustrate the distribution of agreement amongst panellists according to the Likert scale.

	The authors should comment on whether or not the absence of three of the panellists from the second round influenced the rating scales in the second round. The references are accurate but the authors have failed to mention the latest publication of Beers' criteria (JAGS May 2012). These are the most frequently cited explicit criteria to date and have recently been revised and revalidated and should be referred to.
GENERAL COMMENTS	The authors are to be congratulated on (a) submitting a clearly presented and well written paper and (b) considering clinical situations where the decision to prescribe in a frail older patient relys on much more than just extrapulation of evidence from younger patient groups. Some minor points follow: The description of drug related problems is vague and should be more detailed. The relationship between prescribing appropriateness and drug related problems in the elderly should be expanded upon for the benefit of the reader who may not be familiar with this topic. The authors state that explicit criteria developed elsewhere may have little or no applicability to the Australian Healthcare environment - Is there evidence to support this statement? Have criteria such as Beers' Criteria, STOPP/START criteria been used in Australia? What are the difficulties with using explicit criteria developed elsewhere? In many respects, this study is an update of previously reported work without any discussion of the clinical application or development of the 2008 Australian indicators. The authors should comment on this. From a stylistic viewpoint some of the criteria require two or three readings as they are presented with single or double negatives in the sentence e.g. patient with risk factors for impaired renal function is not taking an NSAID, or , patient has no clinically significant drug interactions, or, Patient taking a PPI is NOT taking a medication that may cause dyspepsis unless prescribed for gastroprotection (this is unclear). These should be rephrased in a simpler way e.g. patient should not be prescribed two or more medications with potential to cause clinically significant drug interactions. Criterion 25 is unclear also. Patients with cardiac failure are often co-prescribed an ACE inhibitor or A2A with a diuretic. Falls are a major geriatric syndrome. There are many drugs which increase the risk of falls other than psychotropic drugs e.g. alpha blockers, vasodilators, drugs which increase the risk of orthostatic hypotension, those which predispose to hypoglycaemia. It seems odd that these were included by the panellists on face-to-face meeting. Was there an opportunity to add more criteria than those originally presented to the panel?

VERSION 1 – AUTHOR RESPONSE

REVIEWER ONE

1. Introduction and research question:
• It is not clear, how 'adding recommendations for co-morbidity and the oldest old' would advance or improve the original set developed by the authors or other published criteria sets, such as START/STOPP. If the original

criteria set was described in a bit more detail, ideally contrasting the set with other well established published sets, this may allow the reader to better put this research into perspective.

Recommendations for co-morbidity and the oldest old were missing from our original criteria, and are important issues in the medication management of older people which we felt should be included. In the DISCUSSION under subheadings “co-morbidity” and “the oldest old” we have added examples of; the treatment of coronary heart disease and the use of STOPP/START with reference to the Australian healthcare environment; and treatment of osteoporosis and the use of STOPP/START with reference to the Australian healthcare environment. Discussion has been added about our criteria set in the third paragraph of the INTRODUCTION, and in the first and 5th paragraphs of the DISCUSSION with respect to comparison with other indicators/criteria sets. Our criteria set is described in its entirety in table 2.

• More clarity regarding the (primary) intended use of the criteria would help to better understand the potential relevance of the work to the Australian health care setting and beyond. Although the 'detection of DRPs as part of the Australian medication review process' is identified as the purpose of the original set, the discussion proposes multiple additional purposes of the new set under 'intended use'. Different purposes (e.g. prospective use in clinical decision support versus retrospective large scale application for performance management) pose different demand s on the resulting criteria set (e.g. with respect to the specificity /explicitness of criteria). The extent to which the resulting criteria set is suited to support the primary intended purpose could then serve as a focus in the discussion.

We have added information regarding the intended uses of our criteria under the third dot point in key messages (1st page), in the 3rd paragraph of the INTRODUCTION, and the first paragraph of the DISCUSSION. Criteria such as ours have been used in a variety of settings, using a variety of study designs and in a variety of ways (3rd paragraph, INTRODUCTION). The possibility of international usefulness is discussed at the end of the 5th paragraph of the DISCUSSION.

• Related to the previous point, it would add clarity if the authors could indicate from the start, whether the intention is to develop explicit criteria (that would then be supplemented by clinical judgement in the review process) or implicit criteria.

In the second paragraph of the INTRODUCTION, we have explained that the Australian Home Medicines Review program provides the sophistication lacking in explicit measures such as our criteria list. We had already explained (in Key messages) that our criteria were designed to be used with other Australian medication review processes. At the beginning of the 3rd paragraph of the INTRODUCTION we have added that the majority of our criteria were explicit.

2. Inclusion criteria:
It is not clear what is meant by the ‘oldest old’. A brief description of the conditions and drugs covered by the original criteria set would be helpful.

In the 2nd paragraph of the INTRODUCTION we have added the words “generally regarded as centenarians”. We have added a paragraph in the DISCUSSION under the subheading “oldest old” explaining the difficulties with the treatment of these patients.

3. Description of methods:

• Page 3, line 22-24:
It is not clear what is meant by ‘cross-reference’. This should be briefly explained and reference be made to the original criteria set (although this is done two sentences later). ‘Most common’ reasons should also be more clearly defined - was there a threshold for commonness? The literature review is an important part of the RAND appropriateness method. The relatively low number of criteria presented to the expert panel suggests considerable pre-selection of criteria. It would therefore be desirable if the literature review process and the pre-selection of criteria by the research team could be made more transparent.

We have amended the words “cross-reference”, and added “forty” to describe the threshold for the most common reasons for older Australians to seek or receive healthcare.

Under METHODS, we have added a paragraph under the subheading RAND/UCLA appropriateness method to explain our use of this method with our criteria set. Pre-selection of criteria was stated in the ABSTRACT (under objective) and in the 3rd and 5th paragraphs of the INTRODUCTION.

• Page 3, line 39:
An international audience may be less familiar with the term ‘hostel’ in a health care context.

We have amended “hostel” to “care home”

• Page 3, lines 40-43:
The context in which ‘data collection’ should be ‘feasible’ is not clear, i.e. prospective/retrospective, with or without medical notes etc.. Please, also see under research question above.

We have amended the words “feasibility of data collection” to “potential accessibility of data from the patient, their medical notes or their health care professional(s)” (under METHODS, criteria development)

• Page 3, lines 49:
It is not clear who is to be ‘comprehensively represented’.

We have amended the words “To ensure comprehensive representation” to “We recruited a multidisciplinary group...” (under METHODS, validation of criteria – participants)

• Page 4, lines 22-26:
A brief description of how the median and inter-percentile range was used to assess agreement at this stage (rather than in round 2) would be helpful.

We have added a description of how median and interpercentile range were used to assess agreement under METHODS, RAND/UCLA Appropriateness Method round one

4. Outcome measures
• For criteria without disagreement, it is not clear at which threshold (median appropriateness rating) were criteria retained/discarded?

We have added a description of the way in which agreement and disagreement were reached in METHODS, RAND/UCLA Appropriateness Method round two

5. Results
• Page 5, line 7:
The sentence starts with ‘Of criteria with disagreement....’, but then provides the total number of criteria scored in the denominator in the brackets. This can be a source of confusion.

We agree that clarity could be improved. In paragraph one of RESULTS we have re-worded the text and removed all denominators. We have also done this wherever percentages have been used.

Table 2:
• Given that the set is focussed on conditions most commonly seen in the elderly, it is surprising that asthma features in the set

Asthma was the 32nd of forty problems most commonly managed by Australian general practitioners for people of 65 years or over, according to BEACH (Bettering the Evaluation and Care of Health) data, as discussed and referenced in METHODS under the subheading criteria development.

6. Discussion
• A comparison of strengths and limitation in comparison to already existing tools is missing and would be desirable. This is particularly relevant as it appears that some criteria for commonly used drugs that commonly cause harm in the elderly are missing from the set., such as high-risk use of NSAIDs in patients with risk factors for GI bleeding (without gastroprotection).

We have added a discussion of the comparison of strengths and limitations with other tools, beginning at the end of the first paragraph of the DISCUSSION, and continuing under the subheading prescribing appropriateness tools in Australia. With respect to NSAIDs, issues with their use actually occur in five of our criteria (Table 2, criteria 7, 13, 24, 25 and 31).

• An expansion on how specifically the criteria set may facilitate the ‘prioritisation of therapeutic goals’ would be desirable. The specific work that will be required before the set can be implemented for the various proposed purposes should be specified.

We have added an explanatory example regarding prioritization of therapeutic goals in the DISCUSSION under the subheading Co-morbidity. We have added the information required for implementation of our criteria to the first paragraph of the DISCUSSION.

• The RAM has traditionally been used and tested in order to develop explicit assessment criteria and to establish thresholds for appropriateness in an average patient (e.g. at which threshold for MI risk would coronary intervention be warranted). For this purpose, experts are offered several potential thresholds (e.g. different thresholds of cardiovascular risk), which are then scored by each panellist for the extent to which performing the procedure at the proposed threshold is appropriate or not.
It seems that the present study departs from the original method in that a group consensus was sought first (either by pre-selection (before round 1) or discussion (before round 2)) and an amended criterion was then scored again. A further difference is that panellists in this study were not asked to rate the appropriateness of performing or not performing a specific medical process in

a certain clinical situation, but rather on the appropriateness of following the criterion. If the criterion states 'achievement of blood pressure as a appropriate' (as in criterion 1), the resulting criterion is judgement-based rather than explicit and there appears to be little room for disagreement. These (and any other) departures from the original RAM should be acknowledged and any implications discussed.

We have added an explanation of how our criteria and others have been applied using the RAND process in METHODS under the subheading RAND/UCLA appropriateness method. We have provided examples of others who have presented criteria like ours for evaluation by the RAND method. As discussed in METHODS, our panellists had the opportunity (after round two discussion) to amend criterion that were presented to them, and score them again. The resulting criteria, whether implicit or explicit, were the result of panellists amending and rating the criterion. We believe we have followed the RAND method meticulously.

7. Article focus/Key messages/ Strengths and limitations Article focus Page 1, line 15-16: The primary intended use of the criteria should be described

We have added the primary intended uses of our criteria under the third dot point in key messages, page 1. We have supported this intention in the text, in the third paragraph of the INTRODUCTION

Strengths and limitations
• The claim that the criteria address established evidence based practice gaps is mentioned only marginally in the main text. If this is a key message, it should be demonstrated in the text using key examples

We have added a discussion of evidence-practice gaps (with respect to type 2 diabetes, cardiovascular disease, vaccination, asthma and pain) at the end of the first paragraph in the DISCUSSION, and further under the subheading the oldest old (osteoporosis).

• A comparison of strengths and limitation in comparison to already existing tools is missing and would be desirable (see under discussion above)

We have discussed aspects of the Medication Appropriateness Index (MAI), Beers criteria (including the latest version), the Improving Prescribing in the Elderly (IPET) tool, and the STOPP/START criteria in the DISCUSSION (1st paragraph, under prescribing appropriateness tools in Australia, and further in the DISCUSSION under co-morbidity, and under the oldest old.

REVIEWER TWO

The statistical methods should be described in more detail.

We have added discussion regarding statistical methods in the METHOD under the subheading RAND/UCLA Appropriateness Method round one, and under RAND/UCLA Appropriateness Method round two. There is discussion regarding the rating scale, levels for which there is agreement and disagreement, and details regarding the interpercentile range. We have added a table (table 1), which gives an example of the application of the RAND method for one of our criteria (criteria one) from round one. It lists the voting pattern of panelists, the interperecentile method calculations, and interpretation of the results.

The likert scale used to rate appropriateness should be specified (It is vaguely presented as 1 = highly inappropriate and 9 = highly appropriate).

We have expanded on the description of the 9-point scale used in METHOD, under RAND/UCLA appropriateness Method round one.

The authors state that the median value with dispersion values (IPR and IPRAS) were calculated for each criterion. Which percentiles were included in the IPR? Were these quartiles? The authors state that the three-point region containing the median was used to determine agreement. What was the value of the median below which a criterion was judged not to be valid? A median of 5 would suggest that half of all panellists ratings were below 5 and half were above 5. What was the threshold above which a criterion was accepted? This requires clarification.

We have added information regarding the percentiles used, and values of the median for which agreement or disagreement was reached in METHOD, under RAND/UCLA appropriateness method round one. Application of this information is demonstrated in Table one for one criteria.

A schematic representation of one example of median rating with IPR (boxplot) and IPRAS would be useful to illustrate the distribution of agreement amongst panellists according to the Likert scale.

We understand and agree that the text does not clearly convey the actual voting process, the way in which the IPRAS calculations were applied, and how this together with the median value was interpreted. We felt that a table showing this would add more clarity than a boxplot. We have therefore added a table (table 1) and renumbered the other tables.

The authors should comment on whether or not the absence of three of the panellists from the second round influenced the rating scales in the second round.

We have added to the discussion under METHOD, RAND/UCLA Appropriateness Method round one and round two by explaining that an adjustment was made in agreement and disagreement criteria for a fifteen member panel (round one) becoming a twelve member panel (round two), according to the RAND method.

The references are accurate but the authors have failed to mention the latest publication of Beers' criteria (JAGS May 2012). These are the most frequently cited explicit criteria to date and have recently been revised and revalidated and should be referred to.

We have added to the DISCUSSION a discussion about the Beers 2003 criteria and the Beers 2012 criteria, and its relevance to Australia.

The validation process is well described but, as the authors state in their limitations, these criteria require further development, particularly with respect to the relationship between the potentially inappropriate prescribing practices and real clinical outcomes e.g. adverse drug events, morbidity, mortality, hospitalization, cost, falls, mortality, quality of life etc. The criteria need to be validated as an intervention with respect to these outcomes. The construct validity and reliability / reproducibility / usability of these criteria in different clinical settings has not been determined. The authors have not presented any such data from the previous iteration of these

criteria published in Drugs Aging 2008.

It is true that our criteria need to be validated as an intervention with respect to outcomes (we stated this under strengths and weaknesses), but so do all other criteria, including Beers and STOPP/START. None appear to show a decrease in rehospitalisation or death, health care costs, or reduced use of health resources. (Drugs Aging 2010; 27(12): 947-957, Int Med J 2010; 40: 7-18, Clin Interv Aging 2010; 5: 75-87, Drugs Aging 2012; 29(6): 437-452).

We have discussed the fact that criteria like ours have been shown to have various types of validity, feasibility and quality assessment capabilities - at the end of the first paragraph of the DISCUSSION. Finally, it is untrue that we have not presented any data from our 2008 criteria. They have been applied to a cohort of older Australians. We found a high incidence of undertreatment, and use of inappropriate medicines (IJPP 2012; 20(3): 172-182). This is now stated in the 3rd paragraph of the INTRODUCTION, and did appear (with reference to further developmental work) in the 2nd paragraph of the unrevised manuscript under strengths and weaknesses.

The authors should consider expanding on the role of such criteria in everyday clinical practice in light of existing evidence.

This information is now contained in our manuscript, which discusses the reasons for DRPs to occur, the growing importance of comorbidity and the oldest old, the many ways in which criteria like ours have been used (INTRODUCTION), the availability of information to apply our criteria, the many identified Australian evidence-practice gaps they address, the fact that they are not just a medicines to avoid list but also address other issues such as undertreatment and the need for monitoring, and the fact that they may address Australian healthcare needs better than other tools. We have given examples of most of these issues.

The authors correctly state that existing criteria may not be applicable to Australia. The converse is also true. These criteria may not be applicable outside of Australia. This should be discussed in detail. What are the difficulties which limit transferability and generalizability?

We have discussed possible international usefulness at the end of the 4th paragraph of the DISCUSSION under prescribing appropriateness tools in Australia. We have discussed generalizability in the 3rd paragraph of the INTRODUCTION.

The description of drug related problems is vague and should be more detailed. The relationship between prescribing appropriateness and drug related problems in the elderly should be expanded upon for the benefit of the reader who may not be familiar with this topic.

We have added an explanation about how DRPs can occur - in the first paragraph of the INTRODUCTION. We have stated that appropriateness of prescribing has been assessed by explicit or implicit measures in an effort to identify and reduce DRPs (1st paragraph under the subheading Prescribing appropriateness tools in Australia, in DISCUSSION.

The authors state that explicit criteria developed elsewhere may have little or no applicability to the Australian Healthcare environment - Is there evidence to support this statement? Have criteria such as Beers' Criteria, STOPP/START criteria been used in Australia? What are the difficulties with using explicit criteria developed elsewhere?

We have added a discussion on applicability to the Australian healthcare environment, use of other criteria in Australia, and the difficulties of doing this in the DISCUSSION, under the subheadings prescribing appropriateness tools in Australia, co-morbidity, and the oldest old.

In many respects, this study is an update of previously reported work without any discussion of the clinical application or development of the 2008 Australian indicators. The authors should comment on this.

We agree that this study is a development of previous work. We have explained in the INTRODUCTION that, unlike our 2008 criteria, other prescribing criteria or tools have combined evidence with expert opinion to provide face validity (2nd last paragraph), and now we wish to do the same, using the RAND/UCLA appropriateness method. Our 2008 criteria have been applied to a cohort of older Australians. We found a high incidence of undertreatment, and use of inappropriate medicines (IJPP 2012; 20(3): 172-182). This is now stated in the 3rd paragraph of the INTRODUCTION, and in the 2nd paragraph under strengths and weaknesses.

From a stylistic viewpoint some of the criteria require two or three readings as they are presented with single or double negatives in the sentence e.g. patient with risk factors for impaired renal function is not taking an NSAID, or , patient has no clinically significant drug interactions, or, Patient taking a PPI is NOT taking a medication that may cause dyspepsis unless prescribed for gastroprotection (this is unclear). These should be rephrased in a simpler way e.g. patient should not be prescribed two or more medications with potential to cause clinically significant drug interactions.

The style of the final validated criteria was determined by an expert multidisciplinary panel. Panellists had the option of deleting, updating, creating or amending the wording of all criteria. Vigorous discussion occurred for all criteria during round two. The style of the criteria represents the result of the application of the RAND/UCLA appropriateness method. To change the style would mean invalidating the process we have been through. We therefore feel that we cannot rephrase the criteria at this stage.

Criterion 25 is unclear also. Patients with cardiac failure are often co-prescribed an ACE inhibitor or A2A with a diuretic.

The reviewer may have misread this criterion. Criteria 25 states that a patient should not be receiving a combination of three drugs (1) ACE inhibitor or angiotensin 2 antagonist (2) diuretic, and (3) NSAID. This combination increases the likelihood of renal impairment, and has been called the “triple whammy” (Br J Clin Pharmacol 2005; 59(2): 239-243, Australian Adverse Drug Reactions Bulletin 25(2): 2006).

Falls are a major geriatric syndrome. There are many drugs which increase the risk of falls other than psychotropic drugs e.g. alpha blockers, vasodilators, drugs which increase the risk of orthostatic hypotension, those which predispose to hypoglycaemia. It seems odd that these were included by the panellists on face-to-face meeting. Was there an opportunity to add more criteria than those originally presented to the panel?

Panellists had the option of deleting, updating, creating or amending the wording of all criteria. Vigorous discussion occurred for all criteria during round two. Panellists voted for the addition of two new criteria. They appear as the last two criteria in table 2 noted as “new, and as criteria 5 and 20 in

the final list of validated criteria in table 3, arranged in disease state order.

Thank you again for the opportunity to revise this manuscript.
Sincerely,
Ben Basger (corresponding author)

VERSION 2 – REVIEW

REVIEWER	Dr Tobias Dreischulte NHS Tayside c/o University of Dundee Kirsty Semple Way Dundee DD2 4BF Scotland, UK
REVIEW RETURNED	20-Jul-2012
GENERAL COMMENTS	[T1]: Suggest a different wording because under-treatment is a cause of a DRP, ratherthan the DRP causing undertreatment [T2]: It should be made clear that this is further development of previous work Tthe second aim of adding contextual recommendations to optimise their use should be added [T3]: Suggest a different wording because under-treatment is a cause of a DRP, rather than the DRP causing undertreatment [T4]: Do you define DRP as including adverse events or are adverse events a possible consequence of DRPs? [T5]: Why is under-treatment not listed here? [T6]: For patients or for professionals? [T7]: Suggest to make clear how this is different from 'electronic health record clinical decision support' which can also be based on assessment criteria (e.g. Beers criteria have been incorporated in software) [T8]: I suggest to move this paragraph to directly before the last paragraph of the introduction (study aims), because it highlights limitations of the previously developed and other criteria sets, which presumably this study aims to address. [T9]: The criteria set has not been introduced at this point [T10]: It is not clear what this sentence adds at this point, when the rationale for the further development of the criteria set should be argued [T11]: An extensive list of intended purposes of the previously developed criteria set are described here, which I find distracting. The authors should consider to concentrate on applications that provided the rationale for this update/further development of the criteria set developed in 2008 [T12]: The relevance of different study designs in this context is not clear to me [T13]: I suggest to add'and further develop' [T14]: I suggest that 'supplement existing criteria with recommendations... and adding new criteria where necessary would add clarity [T15]: I suggest to use the word'identified' instead [T16]: It is not clear whether the criteria development assumed that such data would or would not be accessible [T17]: I suggest to move this paragraph to the discussion [T18]: Reference missing. If this was sufficient to define (dis)agreement, what is the relevance of the IPR and IPRAS? Were they used to overrule a verdicat (in case of agreement/disagreement) made by the median method? [T19]: This should say criteria

	[T20]: with disagreement? If so, this should be added [T21]: This does not add up to 41, presumably because the number of criteria accepted after amendment is missing [T22]: Although this illustrate the analysis process, I think the text now sufficiently explains the process [T23]: Although this illustrate the analysis process, the relevance of this remains unclear. [T24]: 1 criterion, 2 criteria [T25]: See previous comments. Which method was the decisive one to determine agreement/disagreement. I think this table is too detailed and the criteria get lost in it. In my opinion, the process of deciding agreement/disagreement should be clearly described in the text and the table should only state the criteria and whether there was or was no disagreement. Alternatively this could be moved to an appendix/supplementary material [T26]: I do not understand the point this sentence is trying to make. If it is it to argue that criteria in the set address established evidence practice gaps, then this should be made clearer [T27]: Why co-morbidity rather than morbidity [T28]: How is this relevant to the use of ACE inhibitors or ARBs in these patients? Do the developed criteria state when arthritis or respiratory disease should be prioritised over CHD? If not, can the claim that the criteria help practitioners to prioritise treatments be maintained? Can any criteria developed by a panel actually claim to do this? Or is this something that needs to be decided case by case, based on clinical experience and a discussion between the health care provider and the patient? [T29]: Specifying the relevant criteria would be helpful [T30]: I suggest to discuss the strengths/limitations of the methodology before moving on to discuss the strengths/limitations of the resulting set

REVIEWER	Dr. Paul Gallagher PhD FRCPI Consultant Physician in Geriatric Medicine Cork University Hospital Wilton Cork Ireland
REVIEW RETURNED	08-Aug-2012

THE STUDY	Minor error in reference section - reference number 21 and 110 are the same.
GENERAL COMMENTS	The paper is much improved. The comments of both reviewers appear to have been adequately addressed by the authors.

VERSION 2 – AUTHOR RESPONSE

Comment [T1]: Suggest a different wording because under-treatment is a cause of a DRP, ratherthan the DRP causing undertreatment

We have amended the wording to explain that there may be two outcomes from DRPs – drug treatment goal not reached, and/or drug treatment causes undesirable effect(s)

Comment [T2]:
It should be made clear that this is further development of previous work Tthe second aim of adding contextual recommendations to optimise their use should be addedd

T2 – We have added the words “previously published”. We have added contextual recommendations to optimise their use

Comment [T3]: Suggest a different wording because under-treatment is a cause of a DRP, rather than the DRP causing undertreatment

T3 - We have amended the wording to explain that there may be two outcomes from DRPs – drug treatment goal not reached, and/or drug treatment causes undesirable effect(s)

Comment [T4]: Do you define DRP as including adverse events or are adverse events a possible consequence of DRPs?

T4 – We use the definition of DRPs most often quoted in the literature – “A drug-related problem is an event or circumstance involving drug therapy that actually or potentially interferes with desired health outcomes”. Accordingly, adverse drug events are one of many types of DRPs which may result in either the drug treatment goal not reached and/or drug treatment causing undesirable effect(s), as per response to T1 and T3.

The reviewer asks if we define DRP as including adverse events, or are adverse events a possible consequence of DRPs? DRPs include adverse drug reactions (ADRs), adverse drug events (ADEs) and medication errors (MEs). ADRs always result in harm and are not preventable, whereas MEs may or may not result in harm and are all preventable. ADRs, ADEs and MEs are all DRPs. MEs that cause harm are classed as ADEs.[2-5]

Comment [T5]: Why is under-treatment not listed here?

T5 – We are grateful for this comment and acknowledge that under-treatment is a cause of DRPs. We have amended the text to reflect this change, adding under-treatment to the other causes

Comment [T6]: For patients or for professionals?

T6 – We have amended the text and added “health care professional directed”

Comment [T7]: Suggest to make clear how this is different from ‘electronic health record clinical decision support’ which can also be based on assessment criteria (e.g. Beers criteria have been incorporated in software)

T7 – We have amended the text to reflect the difference between electronic health record clinical decision support and medication assessment criteria

Comment [T8]: I suggest to move this paragraph to directly before the last paragraph of the introduction (study aims), because it highlights limitations of the previously developed and other criteria sets, which presumably this study aims to address.

T8 – We agree that this paragraph would be better placed before the last paragraph of the introduction, and have moved it

Comment [T9]: The criteria set has not been introduced at this point

T9 – We have amended the text to introduce the criteria set. This now occurs in the second paragraph of the introduction

Comment [T10]: It is not clear what this sentence adds at this point, when the rationale for the further development of the criteria set should be argued

T10 – We have deleted this sentence

Comment [T11]: An extensive list of intended purposes of the previously developed criteria set are described here, which I find distracting. The authors should consider to concentrate on applications that provided the rationale for this update/further development of the criteria set developed in 2008

T11 – Details of the possible applications of our criteria list (both within Australia and outside Australia) were requested by reviewer 2, and added during revision one. We have considerably reduced this description and moved it to the last paragraph of the discussion for improved flow. We believe that the range of potential applications of our current “further developed” criteria set do not differ from the range of potential applications from our previously published criteria set

Comment [T12]: The relevance of different study designs in this context is not clear to me

T12 – We have deleted a description of the different study designs

Comment [T13]: I suggest to add‘and further develop’

T13 – We have replaced the word “update” with the words “further develop”

Comment [T14]: I suggest that ‘supplement existing criteria with recommendations... and adding new criteria where necessary would add clarity

T14 – We have amended the text as requested

Comment [T15]: I suggest to use the word‘identified’ instead

T15 – We have replaced the word “found” with the word “identified”

Comment [T16]: It is not clear whether the criteria development assumed that such data would or would not be accessible

T16 – The criteria were developed with regard to the likely accessibility of data, and this has been added to the text

Comment [T17]: I suggest to move this paragraph to the discussion

T17 – We have moved this paragraph to the discussion, and added “Rationale for the use of the” to the paragraph heading “RAND/UCLA appropriateness method”

Comment [T18]: Reference missing. If this was sufficient to define (dis)agreement, what is the relevance of the IPR and IPRAS? Were they used to overrule a verdicat (in case of agreement/disagreement) made by the median method?

T18 – We have added the missing reference after “...fifteen member panel”.

T18 - To explain how we used the interpercentile method, we have inserted the following explanation into the METHODS paragraph headed “RAND/UCLA Appropriateness Method round 2”;

“Because the median method provided a clear visual interpretation of the ratings for each criterion, we used this method during the face-to-face meeting. Our aim was to ensure that by the end of the meeting, there was agreement between the median method and the interpercentile method for all accepted criteria”.

We have accordingly moved information about the interpercentile range into the paragraph headed “RAND/UCLA Appropriateness Method round 2”, as it was not used in round 1

Comment [T19]: This should say criteria

T19 – We have amended “criterion” to “criteria”

Comment [T20]: with disagreement? If so, this should be added

T20 – We have added “... irrespective of whether there was agreement or disagreement” to clarify this point

Comment [T21]: This does not add up to 41, presumably because the number of criteria accepted after amendment is missing

T21 – We have amended the text so that the indicators add up to 41. There were 17 criteria with agreement that were amended and retained, 10 with disagreement that were retained, 14 for which there was no change, and two new criteria. (17+10+14+2 = 43). Two of the criteria for which there was disagreement but retained ALSO HAD NO CHANGE, which then adds up to 41.

Comment [T22]: Although this illustrate the analysis process, I think the text now sufficiently explains the process

T22 – At revision one, reviewer 2 stated... “A schematic representation of one example of median rating with IPR (boxplot) and IPRAS would be useful to illustrate the distribution of agreement amongst panellists according to the likert scale”.

We understood the difficulty the reader may have in visualising the RAND/UCLA process, and were grateful for this suggestion, which we felt would add greater clarity to our manuscript. At revision one, Reviewer 1 had stated as an overall comment “I think that this is an important piece of work, but requires some clarification”. In addition, we found the description of statistical methods in the RAND/UCLA Appropriateness Method User’s Manual far from clear, and needed to solicit advice from others who had used this method in publications in the international literature for advice on its application.

We feel that the table we added for revision one (Table 1) describes in the clearest way the mechanics of applying this appropriateness method, and respectfully request that it be retained.

Comment [T23]: Although this illustrate the analysis process, the relevance of this remains unclear.

T23 – This comment refers to T22 and has been addressed

Comment [T24]: 1 criterion, 2 criteria

T24 – We have amended “criteria” to “criterion”

Comment [T25]: See previous comments. Which method was the decisive one to determine agreement/disagreement. I think this table is too detailed and the criteria get lost in it. In my opinion, the process of deciding agreement/disagreement should be clearly described in the text and the table should only state the criteria and whether there was or was no disagreement. Alternatively this could be moved to an appendix/supplementary material

T25 – We have addressed the way in which we used the two methods above (T18). We would agree with any editorial decision made with respect to moving Table 2 to an appendix or as supplementary material if it was decided that doing this would not detract from the readers understanding of the manuscript.

Comment [T26]: I do not understand the point this sentence is trying to make. If it is it to argue that criteria in the set address established evidence practice gaps, then this should be made clearer

T26 – We have amended the text to make the point clearer...

“Evidence-practice gaps in Australia have been identified in other areas besides diabetes and cardiovascular disease, such as in asthma, pain and vaccination status.[6-10]The existence of these gaps formed part of the developmental process for these criteria”.

Comment [T27]: Why co-morbidity rather than morbidity

T27 – We have amended “co-morbidity” to “morbidity”

Comment [T28]: How is this relevant to the use of ACE inhibitors or ARBs in these patients? Do the developed criteria state when arthritis or respiratory disease should be prioritised over CHD? If not, can the claim that the criteria help practitioners to prioritise treatments be maintained? Can any criteria developed by a panel actually claim to do this? Or is this something that needs to be decided case by case, based on clinical experience and a discussion between the health care provider and the patient?

T28 – We strongly agree that the purpose of any medication assessment criteria is to raise awareness of medication issues, not necessarily to resolve them. We made this point as clearly as we could in the conclusion. We have amended the text to reflect this...

“...may mean that medicines such as these are never commenced. While we wished to further develop our original criteria set by identifying problems such as these, the ultimate decision regarding medicine use should always be made on a case by case basis, based on clinical experience and a discussion between the health care professional and the patient”.

Comment [T29]: Specifying the relevant criteria would be helpful

T29 – We have specified the relevant criteria

Comment [T30]: I suggest to discuss the strengths/limitations of the methodology before moving on to discuss the strengths/limitations of the resulting set

T30 – We agree that explanation of the strengths and limitations of the RAND/UCLA appropriateness method should be moved from the later parts of the discussion to the METHODS section, and have done this.

We now address further comments from your returned “letter to the editor”;

The authors explain the range of potential applications, but the relevance of the present work has only been argued for the medication review/decision support application. It would add clarity if the authors either concentrated on this application in the introduction or also explained the relevance of the present study (rather than the previously developed set) to other mentioned applications. Please, also see more detailed comments in the annotated manuscript.

We have addressed these comments under T2, T8, T10, T11, T12 and T14 above. We believe that the range of potential applications of our current “further developed” criteria set do not differ from the range of potential applications from our previously published criteria set.

In our article summary under article focus we have stated; “The aim of this study was to further develop and validate a previously published list of prescribing appropriateness criteria for use in older people to improve the quality of the Australian medication review process, and for use in quality assessment and education”. In KEY MESSAGES in the Article summary we stated; “Use of these criteria, together with other Australian medication review processes, may assist in improving patient care in a variety of settings by efficiently identifying DRPs to common medical conditions and commonly used medicines. They may also contribute to the medication management knowledge of health care professionals through education programs and by use in daily practice, and for the evaluation of the quality of pharmaceutical care in older people”. We have also stated at the end of the introduction; “The aim of this study was to further develop our list of criteria, supplementing it with recommendations for co-morbidity and the oldest old where possible, and adding new criteria where necessary through expert consensus”.

I understand that the majority of the criteria in the previous set was explicit, but some of the rewording of criteria in the RAND process has turned previously explicit criteria into implicit criteria. For example, in criterion 1, adding the term ‘appropriate to them’ makes the criterion reliant on the judgement of the assessor and therefore implicit, because it no longer assesses treatment against a pre-specified target that applies to an ‘average’ patient.

We agree that the number of implicit criteria has increased from 3/48 (previously published criteria set) to 6/41 (current further developed criteria set) due to the RAND process. The majority of our criteria are therefore still explicit.

It should be clearly stated which of the following the application of criteria would require: medical notes only, access to health care professionals, access to patient or their carers

In METHODS we stated “Major considerations in their development were likely accessibility of data from the patient, their medical notes and/or their health care professional(s)”. Application requires the patient and either their medical notes or their health care professional, or both. We believe this has been clearly stated.

The statistics are now described in much detail but remain ambiguous.
Please, see annotated manuscript for details

We have addressed this comment – see T18 above

The numbers still do not seem to add up. Please, see annotated manuscript for details

We have addressed this comment – see T21 above

I can't see that NSAIDs are mentioned in criterion 31. The other criteria address NSAID use in patients at risk of cardiovascular or renal complications. The fact that NSAID use in patients with GI risk factors (GI events from NSAIDs including aspirin are among the most common reasons for preventable drug-related hospital admissions, see Howard RL, Avery AJ, Slavenburg S, Royal S, Pipe G, Lucassen P, Pirmohamed M. Which drugs cause preventable admissions to hospital? A systematic review. British Journal of Clinical Pharmacology2006;63(2):136-47.) do not feature in the criteria set suggests a general limitation of the common disease/common drug approach taken here. It suggests that the set may benefit from adding criteria that target common causes of common preventable adverse events.

NSAIDs are included in criterion 31, referred to as part of the list of medicines that cause dyspepsia in table 4. Dyspeptic medicines may cause a variety of upper GI problems such as gastro-oesophageal reflux and peptic ulcer disease.

I have difficulties to follow the rationale of subjecting a criterion, such as criterion 1 ('Patient taking an antihypertensive is at the target blood pressure appropriate for them') to a formal process that asks panellists to rate the appropriateness of following the criterion: The appropriateness of treatment is already implied by the criterion. It would not be logical to consider following 'appropriate' treatment as anything else but 'appropriate'. Even in those examples mentioned in METHODs, the criteria that were developed and rated remained explicit. Panellists only added further explicit (!) specifications. Therefore, although Likert scales and statistical methods are as per RAND protocol, I suggest (as a minimum) that this departure from the RAM is significant enough to refer to the method applied here as a 'modified' RAM.

While it may seem obvious that the appropriateness of treatment is already implied in criterion one, and that it would not be logical to do anything else, nevertheless less than half of those patients with hypertension have it adequately controlled.[11] A criterion on hypertension is included in our criteria set because hypertension is the most frequently managed problem by Australian general practitioners, and anti-hypertensive's the most frequently managed medicines.[12] The fact that sub-optimal treatment represents a major evidence-practice gap[6]in Australia appears to indicate that logic has little to do with appropriateness of treatment. Our criteria were designed to act as a flag for sub-optimal treatment.

The strength of the RAND/UCLA process was that criteria 1was accepted, but the wording was amended to more appropriately reflect optimal treatment. In fact, panellists amended the wording of two thirds (27) of our criteria with this aim. We do not believe that the process we followed differed from the RAND/UCLA process, which has been used just as we did by others.[13 ,14]