REVIEWS

A critical analysis of the methods used to develop explicit clinical criteria for use in older people

JENNIFER MARRIOTT1, PAULINA STEHLIK2

1Faculty of Pharmacy and Pharmaceutical Sciences, Centre for Medication Use and Safety, Monash University (Parkville campus), 381 Royal Parade, Parkville, Victoria 3052, Australia
2Monash University, Parkville, Victoria 3052, Australia

Address correspondence to: J. Marriott. Tel: (+61) 3 9903 9533; Fax: (+61) 3 9903 9629. Email: jennifer.marriott@monash.edu

Abstract

Older people are the biggest users of medications and with the majority of the population ageing it is important to ensure that their medications are managed properly. Many have developed explicit criteria in order to assist in making appropriate drugs choices in the older population. This paper explores whether the methods used to develop the currently available explicit criteria for appropriate prescribing in older people are applied appropriately, and if not, whether this invalidates the criteria themselves. The widespread use of the Delphi technique to develop medical criteria indicates that the technique itself should be evaluated for its suitability in the development of criteria in older people before the criteria are themselves evaluated. A number of criteria have been reviewed and none fulfils the requirements for appropriate development. There is a need for new criteria, with transparent referencing of recommendations and rigorous final evaluation.

Keywords: clinical practice guideline, Delphi technique, criteria development, older people

Introduction

Medication use and misuse in older patients is an important issue in medicine due to an ageing population [1], high rates of medicine use in older people [2], an apparent lack of clinical evidence for safe, effective and appropriate medication use in older people [3] or a combination of these. Many have tried to formulate explicit clinical criteria to assist practitioners in making appropriate medication choices. Evidence-based criteria developed specifically for an older population is problematic as they are commonly excluded from well-designed clinical trials [3]. To overcome this, some criteria have been developed using the Delphi technique, a survey technique intended to find consensus among ‘experts’ [4–8], to assist with formulation of recommendations on suitable treatments in older people. In 2008, O'Mahony and Gallagher [9] published a commentary criticising some of the criteria available at the time for their lack of usability, inability to be generalised and incompleteness including omission of some known instances of inappropriate prescribing and no mention of under-prescribing, drug–drug interactions and duplicate drug therapy. Issues with currently available explicit criteria go much deeper and lie with the development methodology itself. Of interest is the lack consistency between recommendations despite similar development methods [10]. Therefore the question this paper attempts to address is—are the methods used to develop the currently available explicit criteria for appropriate prescribing in older people applied correctly, and if not, is their widespread use justified?

Guidelines, criteria and indicators

Quality assurance is important in health care as it minimises inappropriate prescribing. There is an increasing global demand for clinical guidelines to assist practitioners in delivering the best care for their patients according to the most up-to-date evidence [11], and to evaluate their current practices for potential improvement [12]. It is clear that guidelines, medical review criteria and prescribing indicators have different applications in clinical practice (Supplementary
data are available in Age and Ageing online, eBox 1), but their development is almost identical—in fact, medical review criteria, including prescribing criteria, (referred to simply as criteria) can be developed specifically to evaluate the appropriateness of medical care, or be derived from explicit clinical guidelines developed to improve the quality of care [13]. How a clinical guideline should be developed is described below; these same principles should be applied to the formulation of criteria.

Methods

Search strategy

Prescribing criteria were identified by searching Medline, EMBASE and International Pharmaceutical Abstracts databases using the following terms:

- prescribing appropriateness’ or ‘medication appropriateness’ or ‘inappropriate prescribing’ or ‘inappropriate medication’ AND ‘screening tool’ or ‘quality indicator’ AND ‘elderly’ or ‘aged’ or ‘geriatric’

Articles were considered if they were written in English, described the development and/or evaluation (validity or reliability) of explicit criteria or evaluated medication appropriateness using a specific explicit criteria tool (subsequent searches were performed to find articles describing the development of the criteria if the development was not found in the original search) regardless of publication date. Criteria developed for a specific setting were excluded (e.g. hospital specific, etc …) for consistency in evaluation and as they are less likely to be widely used. Criteria which were local modifications of identified criteria were not included as this was considered redundant; however, all three versions of the Beers criteria were included due to their widespread reference in the literature.

Subsequent searches were conducted using the same search engines with the ‘search term’ AND ‘validity’ or ‘reliability’ (Table 1).

Quality assessment of identified criteria

Processes used in the formulation of clinical criteria

There are three main phases to criteria development; initially there is a need for a review of scientific literature and formulation of the criteria [11]. Once the criteria have been disseminated and implemented, a final step involves criteria evaluation [11]. In this review, we will examine these key sections to evaluate the appropriateness of the methods used to develop explicit criteria.

Literature review

Review of the current scientific evidence is the foundation for producing high quality and acceptable criteria and is preferred to the consensus of expert panels [11].

The evidence should be classified according its power, quality and relevance and therefore whether or not there is strong (or otherwise) evidence for a recommendation given in the criteria; it is generally recommended that this be presented within the criteria as an evidence table [11]. Notably there is no ‘gold standard’ classification system [14] which may cause confusion for readers. It is therefore important to indicate which system of classification was used.

Formulation of the criteria

The Delphi technique is used in the formulation of explicit criteria to address an area where there is a lack of agreement or where there is incomplete knowledge on the subject matter [11]. Because of the widespread reported use of the Delphi technique to develop criteria, the technique itself should be evaluated for its suitability for the development of criteria before these criteria are themselves evaluated.

An analysis of the Delphi technique

The Delphi technique does not have one set method and has been modified in a variety of ways to suit researcher’s aims and objectives [5]. There seems to be no agreement on some key aspects of the Delphi technique, even within each ‘version’ including the definition of an expert, number of experts used, the number of rounds and the consensus level [5]. The Delphi technique has some advantages over other qualitative research techniques. The most commonly mentioned advantage is the lack of discussion domination by any one panel member, which can occur during conventional face-to-face encounters such as focus groups or brain storming sessions [8]. By using anonymous mail survey participants are not swayed to change their opinion in the presence of a more dominant or ‘superior status’ panel member and can also change their views in subsequent rounds without ‘losing face’ [15]. Others argue that panel members may feel pressured to conform to the group’s view [6, 16] or that the anonymity of the Delphi technique may give the panel members a lack of accountability for their answers [16–18].

The Delphi technique can potentially result in the production of large quantities of difficult to analyse data [19]; how data are handled should therefore be decided beforehand. If every statement raised by the expert panel members is added for analysis in the subsequent round, the questionnaire could become overly long; however, generalising the statements into categories may result in statements which have considerably deviated from the original intention [5].

The expert must be defined by the researcher a priori and as there is no universal definition of an expert, this can prove difficult [5, 17]. There is no guidance on the ideal number of expert panel members required [5]. It has been suggested that the use of a heterogeneous group of experts helps to eliminate the potential for skewed results [4] and selecting panel members through methods other than acquaintance minimises bias [16]. Conflict of interest (COI)
also needs to be considered, not only for those conducting the Delphi study, but also participating ‘experts’. Many panel members do not disclose any COI despite its prevalence [20, 21].

The use of expert opinion in the formulation of criteria can be criticised as the findings may be overstated [5]. Using an expert potentially assumes that the opinion of an expert is equivalent to scientific evidence. Although there is a need for peer review, there should be a clear distinction between recommendations derived from a consensus on what scientific evidence means in practice and recommendations derived from personal opinion.

The number of rounds undertaken during development can be influenced by the amount of time it takes to reach consensus [5] although an increased number of rounds may lead to a decreased response rate from panel members [5, 6]. This, or any other reason for the lack of response, can create biased results [19] and alter the validity of the findings [15].

The meaning of consensus should also be established in advance as there is no universally accepted definition; often it is decided ad hoc [22]. Consensus can be anywhere from 51 to 100% agreement among panel members, others do not even set a target but suggest that consensus can be ‘implied’ from the results [16]. Ideally the panel members should come to a consensus, however, if there are time, practical or financial constraints, achieving consensus may not be possible. It is therefore important to ask: what is done with the statements for which no consensus is reached?

The ‘test-retest’ reliability or precision of an experiment or study design describes the extent that results are reproducible when the study is repeated under the same conditions [19, 23]. The Delphi technique has been criticised for its lack of proven reliability [24]. The results produced from the same initial survey vary depending on the experts chosen and are therefore not reproducible [7, 8].

The Delphi technique is far from a scientific method, with a number of inconsistencies and areas with the potential for bias. More importantly, the information gained from a Delphi technique is not fact. Where data are lacking it is merely a consensus among chosen experts about a given practice at a specific point in time, and does not replace scientific evidence. Jones and Hunter [25] point out the danger of using a consensus of expert opinions in place of science:

The existence of a consensus does not mean that the “correct” answer has been found – there is the danger of deriving collective ignorance rather than wisdom [25].

### Evaluation

Generally, evaluation of criteria refers to evaluation of their ‘validity’—does the method used measure whatever it is
intended to measure [23]? As there are a number of types of validity (Supplementary data are available in Age and Ageing online, eBox 2), whether or not results obtained using the Delphi technique are ‘valid’ depends on which type of validity is being considered. The Delphi technique, by definition, assesses face and content validity [23]. However, it is a consensus technique and therefore does not ensure that the ‘right answer’ has been found—it does not measure internal validity of a proposed criteria or prescribing tool. Unless the findings of the consensus can be validated by scientific methods, such as well designed prospective studies, knowing whether or not the ‘right’ answer has been reached cannot be established [25]. This final part of any criteria formulation, evaluation of whether or not the criteria have achieved their primary goal [11, 14, 26–28], is commonly omitted.

Search results

The MEDLINE search yielded 17 results, of which 2 were appropriate to the review. One was a comparative overview [29] of explicit criteria published in 2010; nine criteria [30–38] were identified from this article. One was the development of Australian indicators [35] also identified in the comparative overview. One other criteria [39] was identified from the references cited by one the Australian indicators. To expand this search, a second search was performed with the following terms: ‘inappropriate medications’ or ‘inappropriate prescribing’ or ‘medication appropriateness’ AND ‘elderly’ or ‘aged’ AND ‘evaluation’ or ‘tool’ or ‘criteria’; this new search yielded 60 results. From this only one additional criterion [40] was identified. Nine [36, 41–48] articles were deemed useful regarding ‘validity’ and ‘reliability’.

The EMBASE search yielded 32 results of which 18 were deemed useful. One identified article was the RASP criteria [49]; however, this was a modified STOPP-list criteria and was therefore not included in the review. Seventeen additional articles described medication appropriateness evaluation with prescribing criteria but used already identified tools. Two [50, 51] articles were deemed useful regarding ‘validity’ and ‘reliability’.

The International Pharmaceutical Abstracts search yielded 282 results. Two new criteria [52, 53] were identified that were not identified in previous searches. No articles were deemed useful regarding ‘validity’ and ‘reliability’. In total, 14 criteria (Table 2) were identified for quality assessment in this article, and 11 articles (Table 1) were identified regarding validity and reliability.

Discussion

Quality assessment

Literature review

None of the criteria identified by the authors for this critique present their evidence graded according to strength of evidence and presented in an evidence table (Table 2). The STOPP/START tool [38] references their recommendations, but provides no strength of evidence. The NORTEP [30] has some recommendations referenced but they not are categorised according to strength of evidence. The PRISUS List [40] give an example of how the information in the preliminary list was presented to their expert panel, which did list references used; however, this was just an example and not the full list, and no strength of evidence was assigned, nor were any of the final recommendations referenced. All of the other identified criteria (Table 2) merely mention that a literature review was conducted, but give no supporting references [31, 32, 33, 34, 37, 53].

Three criteria give ‘severity ratings’ or ‘significance rating’ for the likelihood the listed drug is going to cause a clinically significant problem [32–34]; however, these ratings are based on ‘expert consensus’ rather than scientific evidence and should be viewed as opinion rather than fact. Beers et al. [31] argued, in contrast to current recommendations, that consensus methods ‘have several advantages over literature review’; it is noteworthy that they base their initial recommendations on a literature review and then use consensus to conclude which statements from the literature are applicable. None of these recommendations are referenced, nor were the surveys presented to their expert panel. The Australian Prescribing Indicators tool [35] have referenced all of their recommendations, but on closer analysis of the references themselves, they are other explicit criteria for prescribing in older people (many of which have not yet been ‘validated’), test books, indicators for quality prescribing and minimal primary literature. The Zhan Criteria [52] used the Beers criteria [32] in order to formulate their own list, forgoing any literature search whatsoever; the IPET tool [36] did the same using McLeod’s criteria [34].

The lack of transparency on the literature used calls into question the validity of most of these criteria—are they based on sound evidence or simply informed opinion, and is this acceptable?

Formulation of the criteria

Of the identified criteria (Table 2), the reason for using the Delphi technique is not stated, with three exceptions [36, 38, 39], that state they used this technique to test the face and content validity of their recommendations, and one [30] which states that the Delphi technique is a validation method. Others mentioned that the Delphi technique is a consensus method [32, 33, 37, 53], but none justified the need for it. One research group [40] justifies their use of the Delphi technique because it was done for the PIM [potentially inappropriate medications] lists that were generated in other countries—perhaps an argument from popularity. The Beers criteria [31], as mentioned above, suggest that a consensus is more useful in formulation of criteria than a literature review alone because of the different results arising from the literature and narrow application nature of stringent trials.
Table 2. Summary of the methodology used in identified explicit clinical criteria for appropriate prescribing in the elderly

<table>
<thead>
<tr>
<th>Criteria name</th>
<th>Literature review sources</th>
<th>Sources graded?</th>
<th>Were new recommendations allowed by panel members?</th>
<th>Were accompanying references required?</th>
<th>Were the recommendations referenced?</th>
<th>Number of Delphi rounds?</th>
<th>Number of experts in expert panel?</th>
<th>Heterogeneous group of experts in expert panel? (number of professional types?)</th>
<th>Consensus level?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beers criteria 1991 [31]</td>
<td>Published guidelines</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>2</td>
<td>13</td>
<td>Yes (6)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>For round 2, a mean score of the Likert scale&lt;sup&gt;a&lt;/sup&gt; had to be above 4.5 with a 95&lt;sup&gt;b&lt;/sup&gt; CI&lt;sup&gt;c&lt;/sup&gt; to have no limit to dose, duration or frequency of therapy</td>
<td></td>
</tr>
<tr>
<td>Beers criteria 1997 [32]</td>
<td>Textbooks Literature from 1991 Beers criteria [31]</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>2</td>
<td>6</td>
<td>Yes (2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mean score of the Likert scale&lt;sup&gt;a&lt;/sup&gt; had to be below 3 with a 90&lt;sup&gt;b&lt;/sup&gt; CI&lt;sup&gt;c&lt;/sup&gt; to be included in the criteria, a mean score of 3 meant they were resubmitted to the expert panel</td>
<td></td>
</tr>
<tr>
<td>Beers criteria 2003 [33]</td>
<td>Primary literature? (unclear) Guidelines for used of medications in elderly persons Review articles Opinion articles Controlled trials</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>2 + additional round asking about severity of the potential medication problem (not a true Delphi technique as consensus was not required)</td>
<td>12</td>
<td>Yes (5)</td>
<td>Mean score of the Likert scale 1 had to be below 3 with a 95&lt;sup&gt;b&lt;/sup&gt; CI&lt;sup&gt;c&lt;/sup&gt; to be included in the criteria, a mean score of 3 meant they were resubmitted to the expert panel</td>
</tr>
<tr>
<td>McLeod’s criteria [34]</td>
<td>Beers criteria 1991 [31]</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>2</td>
<td>32</td>
<td>Yes (4)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Used a 4 point scale (1 = not significant; 4 = highly significant) to state agreement/disagreement with proposed alternatives from round 1; consensus was not required</td>
<td></td>
</tr>
</tbody>
</table>

Continued
### Table 2. Continued

<table>
<thead>
<tr>
<th>Criteria name</th>
<th>Literature review sources</th>
<th>Sources graded?</th>
<th>Were new recommendations allowed by panel members?</th>
<th>Were accompanying references required?</th>
<th>Were the recommendations referenced?</th>
<th>Number of Delphi rounds?</th>
<th>Number of experts in expert panel?</th>
<th>Heterogeneous group of experts in expert panel? (number of professional types?)</th>
<th>Consensus level?</th>
</tr>
</thead>
<tbody>
<tr>
<td>STOPP/START [38]</td>
<td>British National Formulary</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes (most)</td>
<td>2</td>
<td>18</td>
<td>Yes (5)</td>
<td>Mean score of the Likert scale had to be below 3 with a 95% CI to be included in the criteria, a mean score of 3 meant they were resubmitted to the expert panel</td>
</tr>
<tr>
<td>NORGEP [30]</td>
<td>Text books of geriatric pharmacotherapy, Primary literature, Reviews including Cochrane reviews, Expert consensus or consensus statements, Published guidelines</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes (most)</td>
<td>3</td>
<td>57 first round</td>
<td>Yes (3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50 s round</td>
<td>NOTE: one group includes ‘random group of GP Norwegian specialists’</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>47 completed all three rounds, hence only data from these panellists was included</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method used to develop explicit clinical criteria for use in older people</td>
<td>Indicators tool [35]&lt;sup&gt;f&lt;/sup&gt;</td>
<td>International literature</td>
<td>Australian consensus documents</td>
<td>Clinical practice guidelines for medication use in the elderly</td>
<td>Other prescribing criteria</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhan [52] French Consensus panel list [37]</td>
<td>None</td>
<td>N/A</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>2</td>
<td>7</td>
<td>Yes (3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prescribing criteria and guidelines for prescribing in the elderly</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>2</td>
<td>15</td>
<td>Yes (5)</td>
<td></td>
</tr>
<tr>
<td>PRISCUS list [40]</td>
<td>Prescribing criteria developed by other countries</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes in list provided to panel members</td>
<td>2</td>
<td>38</td>
<td>Yes (8)</td>
<td></td>
</tr>
<tr>
<td>Preventable drug-related morbidity in older adults [39]</td>
<td>Primary literature</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>2</td>
<td>7</td>
<td>Yes (2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reviews including Cochrane reviews</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>2</td>
<td>9</td>
<td>Yes (2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Peer-reviewed medical and pharmacy articles</td>
<td>No&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>2</td>
<td>7</td>
<td>Yes (2)</td>
<td></td>
</tr>
<tr>
<td>Clinically important drug–disease interactions and their prevalence in older adults [53]</td>
<td>Primary literature</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>2</td>
<td>9</td>
<td>Yes (2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Referenced texts</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>2</td>
<td>9</td>
<td>Yes (2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reviews</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>2</td>
<td>9</td>
<td>Yes (2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Textbooks</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>2</td>
<td>9</td>
<td>Yes (2)</td>
<td></td>
</tr>
<tr>
<td>Improving prescribing in the elderly tool (IPET) [36]&lt;sup&gt;f&lt;/sup&gt;</td>
<td>None</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>No</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Likert scale: 1, strongly agree; 5, strongly disagree; 3, equivocal (STOPP/START included 0, unable to offer an opinion).

<sup>b</sup>Concise inclusion criteria for references to be used was listed.

<sup>c</sup>CI, confidence interval.

<sup>d</sup>QR, inter-quartile range.

<sup>e</sup>Visual analogue scale: a 100 mm scale where 0, highly irrelevant, 100, highly relevant.

<sup>f</sup>IPET and the Australian prescribing indicators tool did not use the Delphi technique to develop their tool.

<sup>g</sup>Likert scale: 1, definitely not serious; 5, definitely serious.
There are a few issues that appear in the identified literature in the application of the Delphi technique. The Beers criteria 1997 [32] and 2003 [33] used face-to-face meetings as their second round to arrive at a consensus, which removed any anonymity. Management of data such as additional recommendations proposed by the expert panel was handled differently—some added all the statements gathered from the expert panel [30, 32, 33], others did not specifically mention, or were ambiguous about, how they dealt with the data collected [30, 32, 34]. The definition of consensus also greatly varied among the identified criteria and whether consensus was reached during all studies is not entirely clear (Table 2). Some reached a full consensus by the conclusion of the study [30, 37–39], some stated that a consensus was not reached and seem to exclude the statements altogether [52], while others listed them separately so it was clear that a consensus could not be reached [31, 34, 40]. Also, many of the identified criteria did not give a COI statement [31, 32, 34, 36, 39, 52, 53], some made a COI statement on behalf of the researchers [30, 33, 35, 37, 40], none gave a COI statement on behalf of the experts involved.

**Evaluation**

The majority of the identified criteria have not undergone any formal evaluation. The authors of the STOPP/START tool [38] have, however, shown that the STOPP/START criteria have good inter-rater reliability among 9 physicians from 6 European countries [48] and among 10 pharmacists from both hospital and community settings [46]. O’Mahony et al. [51] conducted a prospective randomised controlled trial using the STOPP/START criteria to evaluate the medications of 200 hospitalised aged patients and make subsequent recommendations with respect to medication changes, while the other 200 received usual hospital care. The outcome of this trial was a significant improvement in the medication appropriateness index (MAI) and assessment of underutilisation of medication scores in the STOPP/START arm. As this trial did not measure patient outcomes and the ability of the MAI to predict adverse outcomes has not been fully established [50], it cannot be conclusively stated that the STOPP/START criteria have internal validity.

The Beers criteria have been used to analyse medication use in number of health care settings to evaluate the relationship between adverse drug reactions (and associated outcomes such as hospitalisation, morbidity, death, etc...) and potentially inappropriate medication (PIM) use; so far only a correlation between the two can be established [42, 54, 55]. Notably, explicit criteria are commonly used for retrospective evaluation of the prevalence of PIMs in a variety of settings [29]; however, none of the criteria (to date) has been validated to show that they are capable of identifying all PIMs. Therefore the prevalence of PIMs in a given situation measured using explicit criteria can be skewed depending on which criteria are used. This has been demonstrated by Steinman et al. [56] who compared the types of PIMs identified by the Beers criteria, Zhan criteria and their own expert pharmacist-physician panel—they concluded that there was a wide variety of PIMs identified by all three measures indicating that no one measure was capable of identifying all PIMs. These types of results were already demonstrated in another study by Steinman et al. [57] where the group compared the number of PIMs identified using the Beers criteria, MAI and polypharmacy (≥9 medications as a marker of PIM).

**Conclusion**

There are a number of issues with the methods used in the development of criteria for prescribing in older people. The main issues concern the presentation of results from the Delphi technique with no distinction between consensus on the ‘scientific literature’ and ‘expert opinion’, a lack of COI reporting on behalf of the expert panel members, and a lack of final evaluation of the published explicit criteria. This analysis of the methods used in development suggests that widespread use of most criteria may not be justified.

**Key points**

- Clinical criteria development needs to be reviewed to ensure the validity of information presented.
- The Delphi technique is used in the development of most clinical criteria but the findings are not presented appropriately.
- There is a need for a new criteria, with a transparent methodology, well referenced recommendations, an evidence table, and subsequent evaluation.

**Conflicts of interest**

None declared.

**Supplementary data**

Supplementary data mentioned in the text is available to subscribers in Age and Ageing online.

**References**

Methods used to develop explicit clinical criteria for use in older people

A review of vertebroplasty for osteoporotic and malignant vertebral compression fractures

ALEXANDRA MONTAGU1, ARCHIE SPEIRS2, JAMES BALDOCK1, JAMES CORBETT1, MARGOT GOSNEY1,3

1Department of Elderly Care, Royal Berkshire Hospital, Royal Berkshire NHS Foundation Trust, London Road, Reading RG1 5AN, UK
2Department of Radiology, Royal Berkshire Hospital, Reading, UK
3Department of Clinical Health Sciences, University of Reading, Building L046, London Road, Reading RG1 5AQ, UK

Address correspondence to: A. Montagu. Tel: (+44) 118 3226720; Fax: (+44) 118 3226544. Email: alexandramontagu@yahoo.com

Abstract

Vertebral compression fractures are a common clinical problem and the incidence of them will increase with the ageing population. Traditionally management has been conservative; however, there has been a growing trend towards vertebroplasty as an alternative therapy in patients with persisting severe pain. NICE produced guidance in 2003 recommending the procedure after 4 weeks of conservative management. Recent high-quality studies have been contradictory and there is currently a debate surrounding the role of the procedure with no agreement in the literature. We examine the evidence in both osteoporotic and malignant vertebral compression fractures; we also describe the benefits and side effects, alternative treatment options and the cost of the procedure. Finally, we recommend when vertebroplasty is most appropriately used based on the best available evidence.

Keywords: vertebroplasty, osteoporosis, malignant, compression fractures, elderly