Author's response to reviews

Title: AQUILA: Assessment of QUality In Lower limb Arthroplasty: An expert Delphi consensus for total knee and total hip arthroplasty

Authors:

Bart G Pijls (b.g.c.w.pijls@lumc.nl)
Olaf M Dekkers (o.m.dekkers@lumc.nl)
Saskia Middeldorp (s.middeldorp@amc.uva.nl)
Edward R Valstar (e.r.valstar@lumc.nl)
Huub J.L. Van der Heide (h.j.l.van_der_heide@lumc.nl)
Enrike M.J. Van der Linden-Van der Zwaag (h.m.j.van_der_Linden@lumc.nl)
Rob G.H.H. Nelissen (r.g.h.nelissen@lumc.nl)

Version: 4 Date: 28 June 2011

Author's response to reviews: see over
Dear Editor,

We thank you and the reviewers for the very helpful work and for the opportunity to improve our paper entitled (MS ID 1574039662524224):

“AQUILA: Assessment of QUality In Lower limb Arthroplasty
An expert Delphi consensus for total knee and total hip arthroplasty”.

We have carefully considered all comments and we have changed the manuscript accordingly. The paper has been thoroughly revised and we feel that it is now a stronger manuscript. As one of the reviewers suggested, the Delphi was repeated and consensus was achieved. Additionally, we provided a more detailed description of the internal rounds and the manuscript has been copyedited by a professional copyediting service to improve the style of English.

We hope that you will now reconsider our revised manuscript for publication in BMC Musculoskeletal Disorders.

With kind regards

Bart Pijls
Olaf Dekkers
Saskia Middeldorp
Edward Valstar
Huub van der Heide
Enrike van der Linden – Van der Zwaag
Rob Nelissen
Answers and comments to the Remarks or Questions are in italic.

- Changes to the manuscript text are in blue underline.

<table>
<thead>
<tr>
<th>Index:</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviewer 1</td>
<td>3 - 4</td>
</tr>
<tr>
<td>Reviewer 2</td>
<td>4 - 9</td>
</tr>
<tr>
<td></td>
<td>Minor Compulsory Revisions</td>
</tr>
<tr>
<td></td>
<td>Minor Essential Revisions</td>
</tr>
<tr>
<td></td>
<td>Discretionary Revisions</td>
</tr>
<tr>
<td>Erratum</td>
<td>15 - 15</td>
</tr>
</tbody>
</table>
Reviewer 1:

General Comments

This is an interesting article and approach worth to be published, there are some minor issues to be considered.

We thank the reviewer for the positive comments and the interest in our work.

Comment

1. Page 11, last line; Maybe nevertheless is a better term than nonetheless (according to my assistant, a professional editor and translator)

Line 301

“Nonetheless” has been replaced by “Nevertheless”.

Comment

2. page 12 lines 253-258: I guess these statements are not covered by the literature or the study. The ideas are correct, but a more defensive wording would be better.

Indeed, this is an interpretation by the authors of the manuscript and not directly supported by either the study or the literature. A more defensive wording has therefore been chosen:

Lines 313-318

“However, some of the reporting and methodological quality items may also be useful for the appraisal of these types of lower limb arthroplasty studies, since the mechanisms of bias (e.g. selection bias and competing risks) are the same[2, 14]. On the other hand, the recommended minimal number of arthroplasties at baseline (100) may not be realistic for
TAA Studies. Some of the generalizibility items, especially regarding component positioning and post-operative functioning may also not be applicable to TAA studies.”

Reviewer 2:

General Comments
Thank you for the opportunity to review this interesting, well written paper. I too am intrigued by the capacity of a Dephi when used to measure quality in selective study designs.

We thank the reviewer for the positive feedback, the thorough evaluation of our paper and the helpful comments.

Major Compulsory Revisions

Comment
1. I feel the introduction needs further information to explain why a specific tool such as the AQUILA is needed when tools such as STROBE already exist. Please include this and indicate the limitations of the STROBE.

We agree with the reviewer that this important topic needs further explanation. There are two aspects important here:

1) STROBE is a general reporting guideline which was not developed specifically for TKA and THA studies. It therefore lacks details that are important for TKA and THA studies for example type of implant and surgical technique.
2) STROBE is a reporting guideline and it should not be used for the assessment of methodological quality (risk of bias). A recent publication by the STROBE-group clearly demonstrates that STROBE is frequently used inappropriately for the appraisal of methodological quality of observational studies, especially in systematic reviews and meta-analysis. A possible explanation may be the lack of tools for quality assessment. Therefore a tool is needed for the assessment of methodological quality for observational studies in TKA and THA. The purpose of the AQUILA is to develop such a tool for TKA and THA studies.

The following has therefore been added to the introduction:

Lines 64-68

“Although STROBE is available as a guideline for reporting in observational studies it lacks details that are important for TKA and THA such as details on type of implant and surgical technique. Additionally, the STROBE-group has recently emphasized that STROBE is a reporting guideline and that it should not be misused for the appraisal of methodological quality[9].”

Comment

2. Having performed a few Delphi studies myself, I understand the impact of an "internal round" on the final external product. In essence, it appears that the items were created in the internal rounds and were only voted for in the external rounds. This is not a typical Delphi and has limitations in that the creativity of the items and the selection of the items is not the responsibility of the expert team. Please clarify this in the paper.

We thank the reviewer for pointing this out. We share the reviewers’ opinion that the internal round may greatly determine the outcome of the Delphi survey. It may even be considered the foundation on which the rest of the Delphi is built. For this reason we took great care to further explain the actions prior to and during the internal rounds, see the blue text below.
Furthermore here are two examples of remarks from external experts to illustrate the importance of the external rounds (which were not merely voting rounds):

1) “The quality appraisal items do not sufficiently distinguish between 'risk of bias' (the main issue) and 'quality of reporting' (a different issue).” As a consequence the distinction between reporting quality and methodological quality was made.

2) “This item certainly is relevant, but i question the use of fixed cut-off points. Instead, the amount of follow-up should be related to the n failures (Murray 1997 JBJS Br).” As a consequence the FU quotient was added to methodological quality items (item 3).

In summary, the external rounds are not merely voting rounds, but they are an actively creative and essential part of the evolution of the AQUILA.

We hope this explanation takes away the reviewers’ concerns.

The following sections have been added to the paper:

Methods; Design of Delphi

Lines 90-110

“During the conceptual phase we determined that the checklists should require quality items (internal validity) and generalizability items (external validity) specific for TKA and TKA. Furthermore the quality items should include items for the appraisal of selection bias, confounding by indication and competing events[2, 14]. Additionally, the checklists had to be easy to use, be able to be completed in an acceptable amount of time and had to allow for the possibility that items be scored as "unknown" in cases with insufficient information.

A master list of relevant items was created as a pre-checklist to allow external experts to assess the face validity and to further develop the final checklist through a Delphi method
in an efficient fashion with the desire to optimize the construct validity. This kind of approach is common for consensus development through a Delphi[15-17]. The master list was generated from items of a recent systematic review of the literature and from the Equator Network website (http://www.equator-network.org/)[18, 19]. The authors of the manuscript, the internal working group, achieved consensus after evaluating and revising this master list in two internal rounds. The actions of the internal working group consisted of the rephrasing of selected items, so that these items met the requirements described above. Since item generation for the master list is an important initial step that may determine the course of the Delphi, we ensured that the members of the internal working group covered all fields (TKA, THA and epidemiology) of the Delphi, that no items were discarded during the internal rounds and that the master list was as comprehensive as possible. Additional aims of the internal rounds were completion of the master list and further testing and fine tuning of the web-based Delphi survey form.”

Comment

3. We need more information about your experts. Consider the lock and key approach to describing your experts and provide profession, background, etc. Were there functional experts such as PT's, were there biomedical engineers, surgeons?

More information on the experts has been added in the Results:

Lines 210-214

“These 37 external experts form the basis of this Delphi and had a mean experience of 16 years (range 3 to 30 years; S D7.5), see table I for the area of expertise. The professional background of the experts was as follows: 30 orthopaedic surgeons or residents, 5 epidemiologists, 1 biomedical engineer and 1 physical therapist. The mean number of publications for all expert was 80 (range 2 to 445).”

Please note that additional information is available in table I
Comment

4. You need a limitations section and I’d include the low response rate (this is lower than most Delphis) and the way in which information was provided.

A limitations section has been added to the discussion, see de paragraph below. In the manuscript we also discuss the limitations of the potential of non-responder bias and the low participation rate:

Lines 276-283

“We should also note some limitations. As mentioned above, although consensus was achieved on the relevance of the generalizability items, the preference for the mode value (e.g. 5 years) was mostly moderate and even low for some items. The latter should therefore be interpreted with some caution. Furthermore, the application of a pre-checklist may have dampened the creativity of the external experts. However, this approach has been successfully used in the development of other checklists[15-17]. The possibility that the results were affected by non-responder bias should also be considered. “

AND

Lines 293-294

“The participation rate was 44/272 (16%). This is towards the lower end of participation rates commonly achieved in this type of survey.[20, 30]”

Comment

5A. You need to further describe the actions of the internal work team during the external phase. Other than just tabulating the results, what else was done.

The following description of the actions of the internal work team has been added to the Methods; Design of Delphi:
“During the external rounds of the Delphi survey the internal working group analyzed and discussed the external experts’ answers after each round, modified the list of items accordingly and rephrased, merged and clarified individual items to optimize their clarity and conciseness.”

Furthermore a more detailed description of the modifications after each round can be found in the Results. See manuscript Result sections concerning reporting quality and methodological quality and generalizability for more details.

Comment

5B Did the 2nd round participants get the information from the 1st round? How was this processed?

Indeed the participants got the information from the 1st round. The following was added to “Methods; Design of Delphi” to clarify this issue:

“During the second and third round the experts received a newly created checklist which was modified according to the results of the preceding round. Each item of the newly created checklist was presented with a summary of the groups’ response to allow the experts to change their answer in view of the groups’ response[13].“
Minor Essential Revisions

Comment

1. Your Delphi references are secondary. Typically, the Delphi is a 3 round (it can also be four) but it generally exploits the abilities of the experts.

We thank the reviewer for this remark. It persuaded us to perform an additional third round, after which there was consensus on the generalizability. The manuscript has been changed accordingly and the most important changes to the manuscript are as follows:

Results section; Generalizability paragraph:

Lines 250-254
“Twenty-two items, related to the comparison of revision rates between studies, were agreed upon by the third external round. These items comprised domains of patient demographics, component positioning, post-operative functioning and regional influences. The final list of these generalizability items can be found in Table 3 and details about the procedure are available in Appendix 2.”

Comment

2. There seems to be a large delay between rounds 1 and 2 (external). Please explain why.

We hope the following Table may clarify this matter. The complete AQUILA initiative took place from July 2009 (beginning of 1st internal round) to June 2011 (end of last external round). See the table (not included in the manuscript) below for details:

<table>
<thead>
<tr>
<th>Round</th>
<th>Start</th>
<th>End</th>
<th>Duration (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Internal</td>
<td>July 2009</td>
<td>August 2009</td>
<td>2</td>
</tr>
<tr>
<td>2nd Internal</td>
<td>September 2009</td>
<td>October 2009</td>
<td>2</td>
</tr>
</tbody>
</table>
The duration of the external rounds consists of the time needed for the external experts to fill out the Delphi survey taking into account holidays and conferences. During the 1st external round one reminder was sent and during the 2nd and 3rd external rounds four reminders were sent to non-responders.

The time between the external rounds was necessary for the internal work team to analyze and discuss the results, to modify the list of items accordingly and to rephrase, merge and clarify individual items to optimize their clarity and conciseness.

The third external round was initiated on the reviewers’ advice, which explains the gap between the second external round and the third external round.

Comment

3. Who created the 3 domains of Reporting quality, method quality and generalizability and did this influence the Delphi participants too much or pigeon hole their responses?

Initially the Delphi consisted of two domains as determined by the internal work team: quality (internal validity) and generalizability (external validity).

During the first external round it was suggested by an external expert to subdivide quality into reporting quality and methodological quality:

“The quality appraisal items do not sufficiently distinguish between 'risk of bias' (the main issue) and 'quality of reporting' (a different issue).”

Therefore these three domains are the result of the Delphi process rather than a potential source of bias to the Delphi process.
“At the beginning of round one the Delphi consisted of two domains as determined by the internal working group: quality (internal validity) and generalizability (external validity). After round one a clear distinction between reporting items and methodological items was made, as suggested by one of the external experts. The quality items were therefore allocated to either the reporting quality checklist or methodological quality checklist.”

Comment

4. Were there any opportunities for open ended questions?

Indeed, there was ample opportunity for open ended questions. We added the following to the Methods; Design and handling of the E-survey:

“The survey consisted of a mixture of multiple-choice and open questions and included text boxes for remarks in order to take full advantage of the knowledge of the expert panel and to ensure creativity of the items. Furthermore all the multiple-choice questions in the first external round had the “other” option with a free text field, so that no restrictions were placed on the answers of the experts. Additionally, opportunity was given to the experts to add items, to modify wording of items and to give explanations and reasons for their answers. Text boxes for remarks ensured that experts could make additions, suggestions and remarks in an unrestricted manner.”

Comment

5A. There are major weaknesses to the functional measures accepted. Why these? The Harris hip and Knee Society Scores are both weak compared with modern evidence.

5B The point differential is compared against what? Please provide further information.
Response 5A:

We thank the reviewer for this remark. We think this issue has been resolved in the third round that was conducted on the editors’ and reviewers’ advice: the generalizability checklist has been completed with KOOS, HOOS, WOMAC and Oxford score, as suggested by one of the external experts during round two. The following text has been added to the results section of generalizability

Lines 243-254

“After round 1 the following items were dropped from the checklist, because less than two thirds of the external experts found them relevant: Hospital for Special Surgery Score (TKA), Merle D’Aubigné Score (THA) and Range of Motion (THA). After the second round the following items were added to the checklists, as suggested by one of the experts: KOOS (TKA), WOMAC (TKA), Oxford Knee Score (TKA), HOOS (THA), WOMAC (THA) and Oxford Hip Score (THA). All these six items were considered relevant in the third round and thus remained in the final checklist. Twenty-two items, related to the comparison of revision rates between studies, were agreed upon by the third external round. These items comprised domains of patient demographics, component positioning, post-operative functioning and regional influences. The final list of these generalizability items can be found in Table 3 and details about the procedure are available in Appendix 2.”

Response 5B:

We further clarified this in the Methods; Generalizability and in the Discussion:

Methods, Generalizability paragraph:

Lines 180-184

“In order to identify relevant items, the experts were asked to select items that are important for case series and cohorts with aseptic loosening in TKA and THA. When an
item was chosen they were then asked to specify the extent of the allowable difference, for each relevant factor, that would be acceptable when comparing different studies in terms of generalizability.”

Discussion:
Lines 303-309

“It is not uncommon that studies of the same type of TKA or THA report rather different revision rates[31]. What factors have caused this difference? Are dissimilarities in patient demographics the cause, or component positioning, or post-operative functioning or perhaps regional influences (including skill and experience of the surgeon)? The generalizability checklist provides a tool to help address this issue. For example: if the difference in mean age between two study populations is larger than 5 years, age is considered an important factor according to the results of the AQUILA.”

Discretionary Revisions

Comment

1. You mention STARD and CONSORT and these are for different designs. I would remove these. STROBE is the only pertinent quality scale. Please mention STROBE in the introduction.

STARD and CONSORT have been removed from the discussion:

Lines 258-264

“The STROBE guidelines are already available for use in reporting original patient research in TKA and THA. The AQUILA checklist now adds to these guidelines, as a treatment specific extension of STROBE, addressing items that are specific for TKA and THA in observational studies. Additionally, the AQUILA checklist addresses both methodological quality and generalizability, while STROBE is strictly a reporting guideline[9].”
Additionally, STROBE has been mentioned in the introduction, see also comment 1 of Major compulsory revisions:

Lines 64-68

“Although STROBE is available as a guideline for reporting in observational studies it lacks details that are important for TKA and THA such as details on type of implant and surgical technique. Additionally, the STROBE-group has recently emphasized that STROBE is a reporting guideline and that it should not be misused for the appraisal of methodological quality[9].”

Erratum

We realize that term “agreement” has been used inappropriately in Table 3. For this reason we have replaced it with “mode” and changed Table 3 accordingly. An additional paragraph was added to the statistics section to clarify this matter:

Lines 187-195

“For an item to be included in the final checklists it must have been selected by at least two thirds of the experts[25]. For generalizibility items the mode was determined, which is the value that was chosen most frequently (e.g. 5 years). The preference for the mode value was calculated by dividing the number of experts who chose the mode value by the total number of experts who considered the generalizability item relevant (NMode/ NTotal ) The preference was considered high in case 80% or more of the experts chose the mode value. The preference was considered moderate in case 67% to 80% of the experts chose the mode value and the preference was considered low in case fewer than 67% of the experts chose the same value.”