Author's response to reviews

Title: Development of explicit criteria for cataract extraction by phacoemulsification

Authors:

Jose M Quintana (jmquinta@hgda.osakidetza.net)
Antonio Escobar (aescobar@hbas.osakidetza.net)
Inmaculada Arostegui (meparmai@lg.ehu.es)
Jesus Martinez-Tapias (jesus.martinez.tapias.sspa@juntadeandalucia.es)
Eduardo Aguayo (eaguayo@fundacionhvn.org)
Emilio Perea-Milla (eperea@hcs.es)
Juan R Lacalle (jralex@us.es)
Marisa Bare (mbare@cspt.es)
Gemma Navarro (gnavarro@cspt.es)
Elena Andradas (elena.andradas@salud.madrid.org)
Juan A Blasco (elena.andradas@salud.madrid.org)
Txomin Alberdi (txotere@euskalnet.net)
Sandra de Fernando (sandradefer@hotmail.com)
Fabiola Eder (feder@chdo.osakidetza.net)
Jose M Beguiristain (jbegiristain@ej-gv.es)
Belen Elizalde (belizalde@ej-gv.es)
Idoia Garai (plsa1bi-san@ej-gv.es)
Joseba Perez de Arriba (utps2vi-san@ej-gv.es)
Jose I Pijoan (jipijoan@hcru.osakidetza.net)
Felipe Aizpuru (faizpuru@htxa.osakidetza.net)
Amaia Bilbao (abilbao@bioef.org)
Nerea Gonzalez (uinves2@hgda.osakidetza.net)
Iratxe Lafuente (uinves1@hgda.osakidetza.net)
Urko Aguirre (uinves5@hgda.osakidetza.net)
Susana Chic (schic@fw.fhcalahorra.com)

Version: 3 Date: 10 October 2005

Author's response to reviews: see over
Response to Reviewers #1

Major Compulsory Revisions.

1.- The Abstract and the Discussion has been revised and the mentioned references removed.

2.- We have included the RAND definition of uncertain and agreement /indeterminate / disagreement literally as the authors of the RAND method have exposed in several publications and their manual (see. The RAND/UCLA Appropriateness Method User's Manual. Kathryn Fitch, Steven J. Bernstein, Maria Dolores Aguilar, Bernard Burnand, Juan Ramon LaCalle, Pablo Lazaro, Mirjam van het Loo, Joseph McDonnell, John Paul Vader, James P. Kahain. ISBN: 0-8330-2918-5. MR-1269-DG-XII/RE, © 2001). The appropriateness algorithm only needs to have defined the disagreement and the three appropriateness categories (Appropriate-Uncertain or Inappropriate). The other two agreement categories (agreement or indeterminate) are only included to understand how we derived those groups on Tables 1 and 2, but have nothing to do with the appropriateness definition (see pages 56-59 of the User's Manual).

Nevertheless, we have included some changes in the text so that it can be more understandable.

3.- No inclusion of some conditions do not lead automatically to no intervention. This would only lead to the impossibility of evaluate with these explicit criteria the appropriateness or not of those conditions. On those cases, as usual, the best clinical judgment would be the criteria, because our explicit criteria wouldn't be applicable.

On the other hand, the very RAND method developers have recommended the inclusion of a manageable number of scenarios, which, in our experience should be less than 1500 scenarios. Inclusion of more chapters (diagnosis) would lead to the development of an enormous number of scenarios for very rare situations in clinical practice, which would conditioning the panel work. On one side, as explained in the discussion, if we include more conditions (and a large number of scenarios) we’d be requiring the panelist to score many scenarios with which the panelists probably have no experience with them (and no evidence on the medical literature). On the other side, a large number of scenarios lead to panelist fatigue and may bias their scoring or increase the number of panelist who refuses to complete them. In summary, as RAND developers recommend in their previously mentioned Manual: “The list of indications should be ...manageable, so that panelists can rate all the indications in a reasonable length of time”.

“If the list turns out to be very long, consideration might be given to reducing its length by concentrating on those indications that represent a substantial number of patients (if this is known) or by narrowing the goals of the study to a smaller group of patients. ”

Among our goals was to create explicit criteria to be used on the most prevalent circumstances. Rare circumstances would require a one by one evaluation.

Nevertheless, we agree with the reviewer suggestion that those considered uncertain can form the basis of a multi-center clinical trial so that more evidence can be found
to support clinical decision making on those cases. Our results already show a 30% of uncertain scenarios. But having more diagnostic chapters with a high percentage of uncertain scenarios, due to the extremely low prevalence and lack of experience of panelists, seems to us low cost-effectiveness for the panel work.

4.- We include in the Appendix 1 the definition of the four categories of visual function, which includes which ADLS we include in each category. Nevertheless, we have expanded the categories c and d to make them clearer as follows:
   c) Recreational difficulties. Visual difficulty with activities that do not affect patient autonomy (watching TV, reading, sew, playing).
   d) Difficulty with activities of daily living. Activities that affect patient autonomy (as basic ADL such as bathing or eating, or instrumental ADL such as shopping and light household chores), or those that affect their work, which includes glare when driving if this affects the livelihood of the patient.

5.- Digits at Table 1 had been reversed

6.- On relation to use of the Kappa statistic, we misunderstood the reviewers suggestion in the first review ("Statistical analysis of kappa scores to assess agreement/disagreement would be helpful") and performed a Kappa statistics of the results of the CART analysis with the panel original scores. That is why the Kappa result was so high.
Now, we understand that the reviewer was referring to kappa between panelist in the panel results, therefore we have deleted the previous information from Table 5. In contrast, we have calculated Intraclass Correlation Coefficient (ICC) between the 12 panelist at 2nd round, which we think is more appropriate than the 66 kappa scores resulted by combining the 12 panelists scores. Nevertheless, we have also calculated the 66 kappa scores and we included a graphical display of them for the reviewer, although we do not included this information in the manuscript because we do not consider it relevant to the purpose of the paper.

![Graph showing Kappa score distribution](image-url)
On the other hand, the misclassification error has been defined as the probability of misclassifying an scenario, more specifically the number of misclassified scenarios divided by the total number of scenarios (see ref Zhang H, Singer B. Recursive Partitioning in the Health Sciences. New York: Springer-Verlag; 1999.). Misclassification error is commonly used as a measure of goodness of fit of statistical classification methods, as well as a way to compare different methods of classification (see Zhang H, Singer B. Recursive Partitioning in the Health Sciences. New York: Springer-Verlag; 1999.). It is defined in page 9 of the manuscript as the number of misclassified scenarios divided by the total number of scenarios. For instance, for simple cataract, it is calculated as \((2+1+3)/153 = 0.039\), which is stated in the second paragraph of the page #13 (numbers are extracted from Table 5). Again, we used the misclassification error to compare the results of the CART analysis against the original panel scores (as a gold standard).

The 0.9% disagreement of Table 1 and from the Results section comes from dividing the 7 scenarios where there was disagreement at 2\(^{nd}\) round by the total number of scenarios (765), \(7/765=0.9\%\).

**Minor Essential Revisions.**

1.- As suggested, we have changed the language with regard to surgical complexity adding the word “anticipated”.

2.- We corrected the grammatical error.

3.- Not they did not received the conformity score. This is an analysis we have performed after the panel’s work and specifically for this article.

4.- We changed the word “important” to “large”.

5.- Change in page 26\#7 .

We have underlined the changes introduced, which included the changes performed after the 1\(^{st}\) revision.
Response to Reviewers #2

General.

1.- We have clarified in the methods section which methods were used to assess validity and which ones to assess reliability. Basically we are performing a face and content validity since the selection of the variables of our explicit criteria match those included by other researchers and the results of our general linear and logistic regression models identified as more important those variables identified by others and by experts in the field. Nevertheless, additional information about the validity of the explicit criteria could be presented once we perform a field study, as we did in a previous appropriateness study of cholecystectomy (Quintana JM, Cabriada, J Arostegui, I Oribe, V Perdigó L, Varona M, Bilbao. A Health-related quality of life and appropriateness of cholecystectomy. Ann Surg. 2005 Jan;241(1):110-8.) but this is outside the goals of this manuscript. The main goal of the manuscript was to present the explicit criteria, not to apply them.

Tables 1 and 2 are showing the original results of the panel summarized by agreement and appropriateness. Table 3 is showing differences from first to second round. Reliability of the panel results should be preferable evaluated by a test-retest analysis or a comparison between two different panels. None of these methods is available in our study, we have just performed one panel of experts and it evaluates the list of indications just once.

However, reliability can be measured somehow by the intra-class correlation coefficient between the panelists in the first and second round, we have included this calculation in the paper. (ICC round 1: 0.5871 y ICC round 2: 0.6934).

Finally, Table 5 shows the comparison between the original result from the panel of experts and the results from the CART analysis, which is shown in Figure 3.

2.- A list of 765 indications was created based on literature review and the second round meeting based on the panelists opinion, as RAND appropriateness methodology suggests to do it. The suggestion of the reviewer (to explore the potential relationship among these indications to determined if the list can be shortened) only can be done after the 2nd round of the panel work, but the list of scenarios is what the panelists want to be. Now, the main goal of this paper is to show and summarize the original results from the panel of experts. CART analysis is used to shorten the list, as other authors did (see Wietlisbach V, Vader JP, Porchet F et al. Statistical approaches in the development of clinical practice guidelines from expert panels. Med Care. 1999; 37: 785-797.), although statistical models are a way to summarize the original results and the risk of misclassification must be taken into account.

3.- As we explain in the Methods section the panelists were selected through the medical society (Spanish Ophthalmologic Society in this case) and then selected by the ophthalmologist of our research team: “Third, we compiled a national panel of ophthalmologists recognized in the field, the names of whom were provided by their respective medical societies and members of our research team.
4.- After the first revision, weighted kappa was calculated and included in Table 5 to measure agreement between CART and panel results. The reason is that we misunderstood a suggestion made by the other reviewer who suggested us to do “Statistical analysis of kappa scores to assess agreement/disagreement would be helpful”. Now, we know that she was referring to test agreement between panelist in the panel results, not between CART and panelists scores. Therefore we have deleted this information from Table 5. Now, we have calculated ICC between the panelists, which we think is more appropriate than the 66 kappa scores resulted by combining the 12 panelists scores. In any case, the gold standard in the comparison panelists versus CART, was the panelists scores, and then Kappa is not the appropriate test for it.