Author's response to reviews

Title: Computer-aided assessment of diagnostic images for epidemiological research

Authors:

   Alison G Abraham (agump@jhsph.edu)
   Donald D Duncan (donald.duncan@bme.ogi.edu)
   Stephen J Gange (sgange@jhsph.edu)
   Sheila West (shwest@jhmi.edu)

Version: 3 Date: 17 July 2009

Author's response to reviews: see over
July 17, 2009

BMC Medical Research Methodology

Dear Editor:

We were happy to receive on June 8th a very helpful review of our original submission and appreciate the opportunity to resubmit a revised version of our manuscript entitled “Computer-aided assessment of diagnostic images for epidemiological research”. The manuscript has been refined based upon the insightful comments of the reviewers. Our responses to the points raised during the review are presented below (our responses in red) and we found, in preparing the response, a need to add content to the manuscript for additional clarity. We have detailed in the responses the location of changes that were made to the manuscript and hope this will assist the reviewers in assessing the revised version.

Thank you for providing us with the valuable review of this work and for this opportunity to share our results with the readers of BMC Medical Research Methodology.

.

Sincerely,

Alison Abraham, PhD (Corresponding Author)
Johns Hopkins Bloomberg School of Public Health
Department of Epidemiology
615 N. Wolfe Street, Room E-7640
Baltimore, MD 21205 USA
Phone: 410-502-9763
Email: agump@jhsph.edu
Comments for Authors:

Reviewer 1:
This paper will be interesting for readers, because the authors described this from the other standpoint. However, there are some points which statement are insufficient, so a referee hopes the authors revise this paper.

1. Why is there not 'Reference'?
   Much research of computer aided diagnosis (CAD) system or medical image analysis has been reported in the world. And, there are lots of papers of CAD evaluation.

   We agree with the reviewer that there is a rich body of literature on CAD systems and medical image analysis. While we could not provide a comprehensive review of either CAD systems or medical image analysis methodology literature, we have highlighted the clinical areas where CAD systems are in use and provided citations for the methods that were used for our computer-aided assessment algorithm. We chose methods that were available in prepackaged software or could be easily implemented to illustrate that such systems were accessible to most researchers. We apologize to the reviewer for the absence of citations. We submitted the manuscript as a .tex file and inadvertently specified the incorrect bibtex file in the latex code. We appreciate being informed of the missing information and have corrected the problem.

2. P.3 l.14 "The validity of...exists."
   A referee thinks that a cataract is not a disease which is measured quantitatively. Because, it is difficult for physician to assess a cataract. Could you add any comments for readers and a referee?

   The reviewer has made an astute observation that cortical cataract severity does not have a gold standard (ascertainment of the true cataract severity) and the assessment of the severity is only semi-quantitative. Much effort has gone towards standardizing assessment and quantifying severity through assessment methodologies developed at Johns Hopkins (Wilmer Eye Institute), Oxford, University of Wisconsin and Brigham and Women's Hospital among others. In such applications where no gold standard exists, our methods for assessing the performance of a diagnostic or measurement system using simulated images are the only option for obtaining an estimate of validity. As the reviewer points out, a typical evaluation of a cortical cataract severity assessment methodology would be to correlate with clinical endpoints; however clinical endpoints such as cataract surgery are dependent upon a number of individual and physician factors. We have added additional text to the introduction to highlight the nature of cortical cataract severity assessment (last paragraph of Introduction).

3. The statement of the algorithm is ambiguous and a referee does not understand well. The authors should revise arranging the algorithm.

   We appreciate the reviewer pointing out the lack of clarity in our discussion of the methods. We have revised the methods as suggested by the reviewer to try to clarify the development of the algorithm.
4. They should add five reviewers' information, which are kinds of occupation, experiences, age, etc..

The results show there is a degree of between reviewer variability that contributes to the noise in cortical cataract severity assessment. Techniques such as periodic training can minimize the variability between reviewers. Though we do not explore the effect of our CAD system on the between reviewer variability in this manuscript, we hypothesize that a CAD system would decrease the inter-reviewer variability by providing consistent assessment information to the reviewer. For the purposes of this study, we did not explore the individual factors that contribute to differences in severity scores between reviewers. We feel that is beyond the scope of the current work but is an interesting question that could be explored. Though there have been studies that examined the components of variability contributing to the overall assessment variability, to our knowledge differences in personal characteristics of the reviewers as a source of noise has not been explored.

Reviewer 2:
This is a revised version of a paper on computer-aided diagnosis (CAD) for cortical cataract detection and classification using simulated retinal images. While the authors' statement that CAD is a relatively new development may be true for this specific research area, it has been around for a couple of decades in other research areas such as breast imaging. Hence, the first sentence of the 'Background' section is misleading: statistically sound methods have been developed for performance assessment, such as FROC analysis for detection and ROC analysis for classification (including multi-case-multi-reader ROC analysis for use in reader observer studies as performed here).

We appreciate the reviewer pointing out that the literature in the field of computer-aided assessment does have a longer history as applied to breast imaging and breast lesion assessment. Our introductory statements were intended to convey the under-usage of CAD in aiding in clinical assessment of diagnostic imagery. Diagnostic images are highly prevalent in most realms of medicine and statistical procedures for assessing the presence of abnormalities or lesions are not well developed. We have clarified our statements in the Background section to convey our intended message

While the authors' use of simulated images to obtain severity scores is interesting, it would be much more worthwhile to investigate the impact of these scores using ROC analysis and in terms of clinically relevant terms such as sensitivity, specificity, and positive predictive value. For example, in breast CAD one assigns a BIRADS category (much like the severity score used here) but clinically it is more important to assess whether or not the clinical follow-up (biopsy versus no biopsy of a lesion in this example) is impacted by the correct/incorrect assignment of the BIRADS category. In other words, how does the score translate into under- and over-diagnosed disease? The cost of a missed disease case is generally higher than that of unnecessary follow-up of a normal case, so sensitivity and specificity optimization involves trade-offs (which is why ROC analysis is so useful).
We are in agreement with the reviewer that methods such as ROC and FROC provide important information regarding observer performance. These methods provide the statistical inference on the performance of an observer. Measures of sensitivity and specificity are ultimately the most valuable in assessing the promise of a new method or the ability of an observer to correctly classify a lesion. However these methods and measures are dependent upon knowing the truth. A gold standard from biopsy, in the case of a breast lesion, is necessary for calculating the sensitivity and specificity or estimating the probability associated with a hypothesis of assessment modalities being identical. In many cases - and cataract severity is one example – a gold standard is unavailable or too difficult to collect. In this study we attempted to surmount this fundamental problem by providing our own gold standard in the form of a simulated image with known severity. However, we do not mean to suggest that this method should be used in any method assessment scenario where a true gold standard can be easily obtained and we have appropriately qualified our statements to explicitly inform the reader that our method is only intended for the intractable problem of no gold standard (first paragraph of Background).

The purpose of a CAD system is to help radiologists in their interpretation of medical images. The authors seem to have performed (limited) performance analysis of the stand-alone performance of the CAD system and the stand-alone performance of 5 readers. The first step in CAD performance analysis usually involves the stand-alone performance of the CAD system. If satisfactory, then the *impact* of the CAD system on the performance of a set of readers is assessed, i.e., the performance of the readers with and without the use of the CAD system is what is of interest. Is the introduction of CAD going to result in much more correctly diagnosed disease cases? But perhaps at the cost of more false alarms?

The measure we used to assess performance was bias, which was calculated as the difference between the assigned severity score and the true severity score. We feel this is an appropriate measure of performance for a continuous outcome such as a cataract severity score. However, in most CAD applications, the outcome would be presence or absence of a lesion. For such outcomes providing metrics based on binary outcomes (e.g. sensitivity, specificity) would be appropriate. The bias across levels of true severity and overall was calculated for both the CAD system and for each of the five reviewers. While our CAD system was designed to work in conjunction with a reviewer, who is an indispensable component of the assessment method for this application, our simulation images were designed such that the reviewer provided minimal input for this study. Thus the stand-alone performance of the system is observed in the estimates of bias. We acknowledge that our metric of performance assessment is non-standard for the field of observer performance methodology. We have also done an ROC analysis by choosing a severity score of 3.0 out of 16.0 to classify individuals as having cataract or no cataract, thus converting our continuous score to a binary outcome. Using this binary outcome, the area under the curve was greater than 0.98 for all graders and the CAD algorithm, which didn’t provide any discrimination in performance for the two methods. We have added a statement to the manuscript to that effect (first paragraph of Results).

Since the authors did have readers available, it is not clear to me why the important issue of the impact of CAD on the readers' performance was not
assessed.

We are in agreement with the reviewer’s insight into the importance of the impact of the system on a reviewer’s performance. This study, however, was intended to provide a proof of concept of the use of computer-aided technology in the assessment of outcome measures from images for clinical research. The simulation study provided images with minimal signal noise from which we could evaluate the bias and variability associated with the assessment methodology that arose purely from the task of identifying and summing areas of opacity. This ideal setting does not provide an appropriate context for assessing the impact of the use on a reviewer since reviewer input is not required for images with no noise. This is evidenced by the zero variability of the CAD system when assessing the same image multiple times. However, this important question deserves study and we have continued this work using real image data to provide measures of impact. We appreciate the reviewer’s comment on the importance of this measure and hope that soon we may provide that information. We have provided an explanation for the absence of that analysis in the discussion (paragraph two of Discussion).

While the CAD system itself is briefly explained, there seems to be no explanation on exactly how the system was trained (how many cases were used, were real images or simulated images used, were the testing cases described here truly independent from the cases used to train the CAD system etc).

We would like to thank the reviewer for pointing out this lack of detail in the manuscript. The system was developed using real images from a cohort study of eye disease. Parameters values were determined empirically based on performance. While methods exist based on ROC analysis for aiding in the optimizing of parameter values to maximize performance, they were not easily implemented for this application. We have updated the manuscript with this information (end of first paragraph of Methods).

The authors state that severity scores lie between 0 and 16, but the x-axis in Figure 3 ends at a score of 12. Why?

The theoretical range of scoring based on the Wilmer cortical cataract lens grading methodology is 0 – meaning no identifiable opacities anywhere on the lens – to 16 – meaning the entire viewable lens area is covered with opacities. However, when we created the simulated lens images, we drew from a trivariate normal distribution to determine location and intensity of the opacity with parameter values set to determine the spread of the opacity. While we had some control over the area of coverage by setting the number of opacity clusters that were placed on the image, within the restricted ranges of mild, medium and high severity, the final area of coverage was determined by chance. Thus we had no images that had more than 11.8/16.(74%) of the area covered by opacity. We have noted this in the legend of Figure 3.

However, for evaluating performance, images near the boundaries of the scoring range are easier to grade. Thus we were not concerned by the absence of images representing the extremes given that performance differences were more likely to be noted in a narrow range of scores.

The bias of the CAD system seems to increase dramatically for higher severity
scores, which seems undesirable. Why does this happen and what would the impact be on clinical decision making?

As the reviewer notes, the CAD system fails at higher severities due to a failure of the standardization algorithm to adjust for background noise (due to heterogenous lighting conditions) as more of the area is covered by opacity. This is certainly undesirable and we discussed alternative methods for standardizing in the manuscript (see first paragraph of the Discussion). However, given that cataractous lenses are typically removed once they reach a moderate degree of severity – and thus images with high severity are rare in clinical research datasets - the implications of such a failure are less critical for this particular application. One goal of the research was to show that a CAD system for assessing images can be developed using standard statistical and image processing methods and thus we did not explore more complicated methods for improving the standardization. For a system that would be adopted for use in clinical research and most certainly for clinical practice, a further exploration of better methods would be necessary.

Many of the literature citations are '?' (presumably because bibTex needed to be run again).

We apologize to the reviewer for the absence of citations. We submitted the manuscript as a .tex file and inadvertently specified the incorrect bibtex file in the latex code. We appreciate being informed of the missing information and have corrected the problem for resubmission.