Reviewer's report

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

Version: 2 Date: 1 June 2009

Reviewer: Karen Drukker

Reviewer's report:

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). 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).

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? 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.

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).

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 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?

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

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:

I declare that I have no competing interests