Reviewer's report

Title: Graphical Presentation of Diagnostic Information

Version: 1 Date: 14 January 2008

Reviewer: Jennifer Clark Nelson

Reviewer's report:

I. General

The authors aim to 1) summarize the main types of graphical displays available for use in diagnostic accuracy studies and in reviews of accuracy studies and 2) review the use of such methods in the literature. Both topics are useful to consider and could potentially offer interesting contributions to the active dialogue in the literature on methods for assessing diagnostic tests. However, there are key gaps that need to be filled in and, in general, a fuller treatment of each topic is needed. The authors should consider the possibility dividing coverage of these subjects into two distinct manuscripts.

II. Major Compulsory Revisions

Part 1: Summarizing graphical displays

1. The authors devote considerable space to describing available methods for graphical description of diagnostic test information. However, their summary is not up-to-date. In particular, it lacks some of the most recent and potentially useful contributions:


Although the graphical methods proposed in these papers are too recent to have
been used in the primary accuracy studies that the authors review, a discussion of existing methods is not complete without this material. In particular, since a primary conclusion made by the authors is that "work is required to improve graphical displays" (p2, abstract) and some of these newer methods directly address the limitations of existing methods that the authors raise, these new methods should be described.

2. The paper would benefit from using terminology to define the two primary ways in which diagnostic tests are typically evaluated: their "risk predictiveness" (i.e. what may be most clinically relevant) and "performance as a classifier" (i.e. sensitivity, specificity, etc). The authors refer to these two aspects in several ways throughout the manuscript, but they do not clearly define these terms. Doing so at the outset would provide considerable clarity and enhance their description of existing methods.

3. A fuller summary and clearer evaluation of existing methods (per 1 and 2 above) would allow the authors to make more concrete recommendations in the second half of their paper, when they review the use of graphical displays. For example, instead of generally recommending that studies "include multiple types of graphics," the authors could explicitly say which types of graphics and which methods specifically provide the desired description. Also, the authors should use their figures as opportunities show the ideal graphical presentation for each method (or clarify that the current figures already provide the optimal, versus simply an example, display).

Part 2: Review of use of graphical displays

1. The authors should describe their methods for selecting articles from identified journals in more detail. For example, what criteria were used for inclusion of "primary diagnostic accuracy" studies? This is necessary to know how heterogeneous such studies might be with respect to the primary study aim or type of outcome, factors on which the authors' expectations about use or nonuse of a particular graphical tool should depend. For example, I don't have a good sense from the description of the review methods as to whether all included accuracy studies aimed to evaluate both "risk predictiveness" and "performance as a classifier" and would therefore be expected to use graphical tools for both purposes. Or are the authors arguing that all accuracy studies should include assessments of both aspects?

2. Why was the search for studies limited to 2004 (and 2003 for systematic reviews)? A review comparing several years might show interesting differences, optimistically improvement over time, in the use of graphical methods. Further, a review including more recent years would provide a more state-of-the-art assessment, an assessment that may well have changed in the most recent times since the biostatistical methods for accuracy studies are relatively young and rapidly evolving.

3. How were the 10 journals selected? The sample is heavily weighted by articles from a single journal (~50% in Clin Chem) which makes it less representative and
less generalizable. Why weren’t any major epidemiologic journals (such as Am J Epidemiol, Int J Epidemiol, Epidemiology, or J Clin Epidemiol) included in the review? These journals regularly contain accuracy studies (e.g. I found >10 possibly relevant articles in 2004 from selected epidemiology journals upon cursory search) and, given their relatively methodological emphasis, they might contain studies with relatively higher utilization of graphics, which would be important to note.

4. The authors need to improve their arguments (although I personally agree with them) that studies would importantly benefit from increasing their use of graphical methods. Why are current non-graphical methods (e.g. summaries of data in the text, tabular displays) insufficient? What do the graphical methods gain us beyond what can be described non-graphically? Perhaps use 1-2 specific studies included in the review as examples and discuss how they could have been improved with use of graphics. This question stood out to me because the authors spend a good deal of time discussing deficiencies in existing graphical methods, which, at first glance, seems in conflict with their message that these graphical tools should be used more. This apparent conflict could be resolved by a) Clearly identifying problems with not using any graphical method, and b) Making more concrete and specific recommendations about what combination of graphical methods is optimal.

III. Minor Essential Revisions

1. The following sentence is too strong: However, truly dichotomous tests rarely occur in practice (page 5). Dichotomous classifiers are common in administrative database studies, for example, where ICD-9 code indicators are often used as surrogates for true disease status (Wilchesky et al. Validation of diagnostic codes within medical services claims. J Clin Epidemiol 2004 Feb;57(2):131-41; Rosamond et al. Trends in the sensitivity, positive predictive value, false-positive rate, and comparability ratio of hospital discharge diagnosis codes for acute myocardial infarction in four US communities, 1987-2000. Am J Epidemiol. 2004 Dec 15;160(12):1137-46.) Consider replacing with something like the following: In this evaluation, we are concerned with continuous or categorical tests for which the majority of existing graphical methods apply. Studies with dichotomous outcomes should then be excluded from the review.

2. The authors should acknowledge that, in addition to empirical and parametric methods, a wide array of semi- and non-parametric methods exist for fitting ROC curves and refer key references in this literature.

3. The references cited that suggest that ROC plots are poorly understood by clinicians are very old and may not reflect current trends, especially given that a big increase in the use of ROC methods has occurred in recent decades. Better to cite more recent literature (e.g. Puhan et al. A randomized trial of ways to describe test accuracy: the effect on physicians’ post-test probability estimates. Ann Intern Med 2005 Aug 2;143(3):184-9.) or to refer to the current debate on ROC (e.g. Comments on Cook et al. Circulation. 2007 Feb 20;115(7):928-35).

What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

Level of interest: An article of importance in its field

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.