Reviewer's report

Title: Evaluation of HIV Testing Algorithms in Ethiopia: The role of the tie-breaker algorithm and weakly reacting test lines in contributing to a high rate of false positive HIV diagnoses

Version: 1
Date: 18 August 2014
Reviewer: Jörg Schüpbach

Reviewer's report:

General comments:
The paper analyses the performance of various HIV testing algorithms used in Ethiopia. HIV screening there is based on rapid tests. In case of a reactive result in rapid test A, a second rapid test, B, is performed. If this test is also reactive, the sample is ruled HIV-positive. If the result is discordant, i.e., tst B non-reactive, a third test is performed (C). The authors now compare two possible strategies for test C. The regular “tiebraker” strategy uses a third rapid test (Unigold), while the alternative strategy uses a rapid confirmatory assay (Immunocomb), which assesses antibodies to p24, p31, gp41 and gp120 (for HIV-1) or respectively gp36 (for HIV-2).

Using these two strategies in two separate testing sites and by screening a total of 2622 individuals, the authors have put together a study sample that consisted of 428 plasma specimens. Among these, there were 203 HIV-positives (representing all positives among the screened total; HIV prevalence = 7.7%) and 225 HIV-negatives (every 10th negative selected). These 428 samples were re-tested in a laboratory; each sample was systematically tested by all of the tests involved in the screening, namely, KHB, StatPak, Unigold, and OIC. Western blot was used for gold standard, and "PCR" for resolving the numerous WB-indeterminate results.

The data generated in this study would have enabled the authors to determine the primary parameters of test performance for each of the tests, namely their diagnostic sensitivity and specificity. These are the most important test characteristics, while PPV and NPV depend not only on sensitivity and specificity, but also on HIV prevalence. PPV and NPV describe how a test of given sensitivity and specificity will perform when the prevalence of HIV is varied. If the prevalence is high, a test rendered highly sensitive at the expense of its specificity will give more correct results than a test of high specificity at the expense of sensitivity, and vice versa. If the prevalence is constant, as it is the case in this paper (namely 7.7%), the PPV and NPV do not add much additional information compared to a test's sensitivity and specificity.

Unfortunately, the sensitivity and specificity of each test are not presented in this report; otherwise the authors would have realized that the Unigold has an inferior specificity and is thus bound to generate false-positives. The question is
therefore not so much whether a tie-braker strategy is inferior to a confirmatory step; it is more simply the question of which test should be selected as the third test. Whether this is another rapid test or a multi-line confirmatory test does not matter per se, as long as their performance, in particular their specificity, is the same. Clearly, the low-specific Unigold should not be used in this function. Given the high diagnostic sensitivity of today’s rapid tests, samples that had a weakly-reactive result in test A and a negative result in test B are highly unlikely to be truly HIV-positive, i.e. they represent a sample of a low HIV prevalence. In this situation, only tests of the highest specificity yield reliable results.

As further topics, the paper addresses the meaning of low-intensity reactions and the question of whether Leishmaniasis promotes false-positive HIV test results.

Following the formal list of points to be addressed by the reviewer:

1. Is the question posed by the authors well defined?
Based on my considerations described above I think that the authors should reconsider their strategy of test evaluation and put their primary focus on individual test performance; performance of testing algorithms follows directly from test performance. I consider this as Major Compulsory Revision

2. Are the methods appropriate and well described?
Should be modified accordingly. I consider this a Major Compulsory Revision

3. Are the data sound?
The primary data should be a table with test performances: sensitivity, specificity, confidence intervals. A second table could then describe how the tests perform under different HIV prevalence: low, e.g. 1%, the 7.7% of Ethiopia, 10%. Show the PPV and the NPV; giving the CI here is not that important — tables should be kept simple. A further table could then list those testing algorithms that provide the highest overall sensitivity and specificity, at minimal expenses and confront the two algorithms currently used in Ethiopia with them. I also consider this as a Major Compulsory Revision

4. Does the manuscript adhere to the relevant standards for reporting and data deposition?
Yes

5. Are the discussion and conclusions well balanced and adequately supported by the data?
Not really; they should be adapted to the suggested new structure of the paper.

6. Are limitations of the work clearly stated?
Yes, but they should also be adapted to the suggested changes.

7. Do the authors clearly acknowledge any work upon which they are building, both published and unpublished?
Yes
8. Do the title and abstract accurately convey what has been found?
Not really should also be adapted to the suggested changes.

9. Is the writing acceptable?
Yes, but the paper is unnecessary long and complicated, contains too many numbers and is very difficult to read.

Specific points (minor, but also considered essential):
Background, p.3, bottom paragraph: the antigens of HIV assessed by the OIC are, p24, p31, gp41, gp120, and gp36 (not p40, not p36).

Methods,
Sample size: every Nth should be defined: apparently every 10th
DNA-PCR: Method should be indicated (reference); it would also be important to know whether the DNA-PCR procedure used recognizes the different clades of HIV-1 present in Ethiopia (C, D etc.). Also, were there no HIV-2 positive samples, as based on the OIC? Regarding the reference laboratory in South Africa: who, what method, sensitivity to subtype D?

Discussion:
p.9. 4th Paragraph: You cannot "identify" false-positive results, use "yield" instead.
p. 10, Description of false positives, 2nd paragraph. Antibody to p24 is not the earliest, but one of the earliest antibodies (usually, anti-gp41 is a bit earlier).

A few lines down: antibodies to gp41 are very specific — in a line immunoassay they had a specificity of 100% (see Schupbach et al, BMC Infectious Dis 2011, 2012 and PLoS One 2013). I therefore do not believe that in Eastern DRC 12 of 24 false reactions on WB were "anti-gp41"; either this was reaction to a cellular protein migrating to the same position in the gel (not present in line immunoassays) or these were truly early antibody-positives missed by the "gold standard".

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

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**
I declare that I have no competing interests.