Author's response to reviews

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

Authors:

Leslie Shanks (Leslie.shanks@amsterdam.msf.org)
M. Ruby Siddiqui (ruby.siddiqui@london.msf.org)
Jarmila Kliescikova (jarmila_k@hotmail.com)
Neil Pearce (Neil.Pearce@lshtm.ac.uk)
Cono Ariti (cono.ariti@lshtm.ac.uk)
Libsework Muluneh (libse77m@gmail.com)
Erwan Pirou (erwan.pirou@amsterdam.msf.org)
Koert Ritmeijer (Koert.Ritmeijer@amsterdam.msf.org)
Johnson Masiga (jmasiga.im@gmail.com)
Almaz Abebe (almaz_abe1@yahoo.com)

Version: 2
Date: 22 October 2014

Author's response to reviews: see over
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: 4 August 2014
Reviewer: Bianca Bruzzone

Reviewer's report:
The study is an interesting and thorough analysis of the performance of different diagnostic algorithms for HIV testing in the setting of a developing Country with a >1 seroprevalence in the general population. The sample size was large enough to draw significant conclusions. The paper is well written with only minor imperfections. On the whole the authors’ efforts at offering an extensive and accurate analysis of several possible algorithms can deservedly add to the literature on the subject and become a useful help to individuate the best possible diagnostic algorithms in similar settings.

One negative aspect of the article is its length, mainly due to the meticulous description of all the different comparations, which makes the lecture long and tiring. However, making the work more concise could negatively impact on its exhaustiveness, so that we should be cautious on that recommendation. A way to help resolving this defect could be extrapolate all the theme of the impact of leishmania infections on HIV testing, maybe by submitting that part in another separate article, giving only a limit in this paper.

Thank you very much for your positive feedback on the manuscript and for the suggestion to extract the leishmaniasis objective from the paper. We removed this aspect of the work from this paper and are preparing it for publication separately. This does simplify the paper considerably and makes it easier to read without losing important detail. We have also done another edit to try and improve readability.

Finally, an up to date adjustment of bibliography would be welcome.

We have updated the reference list as suggested.

Level of interest: An article of importance in its field
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' below
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 Schuppach

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., test 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.

We thank the reviewer for his thoughtful comments and feedback. It is encouraging to see that the poor specificity of weakly reactive test lines is recognized. As we detail in our paper, evidence of the problem has been accumulating for some time. However test manufacturers continue to state that any positive reaction on a test line should be read as positive. One of the contributions we hope to make with this paper is to push forward this change.

On the issue of the third test to be used with a discordant result, we respectfully disagree with the reviewer. This is illustrated in our analysis by the fact that the tie-breaker algorithm performs poorly whether or not the third test is Unigold or Stat-pak. We maintain, and provide collaborating evidence from the literature in our paper, that a different type of test is needed to resolve discordant RDT results. Simply adding another RDT that is vulnerable to the same cross-reactivity as the other RDTs will not solve the problem.

Further while we do not dispute that the performance of the algorithm depends on the specificity of the tests included, the additional data we have now provided in the revision shows clearly that our results cannot be attributed to the choice of Unigold as the third test, as all tests perform with a specificity \( \geq 99\% \). Further, we would like to emphasize that RDTs have been developed as screening tests with high sensitivity, and are not manufactured as diagnostic tests. It is now well documented that the specificity of HIV RDTs can vary depending on different populations, locations and over time. Confirmation tests on the other hand, are more likely to be consistently accurate due in part, to their ability to test separately for different antigens. The use of confirmation tests is routine in well resourced settings where supplementary tests of a different nature than the screening tests are used routinely. We are proposing that the same standard be applied in resource limited settings.

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
We have considered the reviewer’s suggestion to change the primary objective of our study, however we believe our original primary objective remains the right choice. We do agree that the individual RDT performance should be reported and have revised the paper to include this data. These results show that all 3 RDTs (including Unigold) achieved 99% specificity, which is above the 98% standard proposed by WHO in choosing RDTs to include in the algorithm. Despite the good performance of the individual RDTs, the tie-breaker algorithm still does not meet the minimum criteria of WHO for the algorithm performance (PPV > 99%). The additional analysis therefore adds weight to our argument that it is the tie-breaker algorithm that is the problem not the choice of individual RDTs.

To conclude on this point, our choice to focus on algorithm performance rather than individual RDT test performance measures was made due to the following factors:

1. We were interested in the broader policy implications of the algorithm choice, rather than the very limited issue of how an individual RDT performs. By maintaining our original objectives for the study, we are able to provide evidence for 3 major policy changes: that of abandoning the tie-breaker regimen which is currently in use in many countries, adding a confirmation test and interpreting a weakly reactive test line as indeterminate.

2. As described above, individual RDTs vary in their specificity depending on the population tested and other factors that are not yet well defined. RDTs continue to be manufactured as screening tests. The choice of the algorithm is critical to protect against false positive tests. This is reflected in the fact that WHO guidelines for diagnosing HIV give specific guidance on the choice of the algorithm in addition to standards for sensitivity and specificity of the individual RDTs composing the algorithm.

3. The focus on diagnostic algorithm rather than individual test performance is consistent with the literature. See for example, one of the earliest and most influential papers in this area Gray et al, published in the BMJ in 2007. Other examples are: Galiwango, 2013; Baveeva 2013; Lyamuya 2009; Kroidl 2012; Crucitti 2011; Zeh 2011; Granade 2005; Aboud, 2006; Styer, 2011; Meda, 1999.

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

The results section has been adjusted to include the individual RDT test performance characteristics as suggested.

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.

We have added the first table as requested. We have not included the second table
suggested, as the emphasis in our paper is not how individual RDTs perform under
different prevalence levels but rather on the diagnostic algorithm as explained above. We
have therefore chosen to keep the focus on what we believe is most relevant to both
policy makers and clinicians in resource limited settings. The third table suggested is in
the paper (now Table 4).

We recognize that there are a lot of numbers in the paper but believe it is important to
include confidence intervals with our results in order to accurately reflect the uncertainty
in our estimates and allow comparison with other work. To improve readability, we have
removed where possible numbers from the narrative and referred readers to the tables.

Finally we have included predictive values in our tables as in our experience this is
standard in diagnostic studies. We elected not to include likelihood ratios however, in
order to avoid over-complicating the presentation. We feel reporting PPV of the
algorithm is important because it provides additional information to that of specificity,
and allows the reader to compare the performance against the WHO standards for
algorithm PPV. It is perhaps important to note that there are no standards for algorithm
specificity in the WHO guidelines, only for PPV. Further, HIV RDTs generally have a
high specificity, and it is hard to differentiate between algorithms on the basis of small
differences in specificity alone. Perhaps unique to HIV are the very significant
consequences of getting the diagnosis wrong---both to the patient and in program costs.
As can be seen in our results, even a seemingly excellent specificity can result in an
unacceptably high number of individuals being falsely diagnosed. This reality is made
apparent when the PPV is reported.

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.

We have incorporated the data on individual test specificities in the discussion as
suggested but have kept the focus on the algorithm performance consistent with our study
objective.

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.

   We have changed the abstract to include the individual RDT specificity.

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

   We have addressed this by removing the VL analysis. This shortens the paper and reduces the complexity that came from having two separate analyses. We have also edited the paper to reduce length and improve clarity. In addition, if requested, we can suggest putting Table 1 and 2 as supplementary material.

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).

   Thank you for pointing out this error. It has been corrected.

Methods,
Sample size: every Nth should be defined: apparently every 10th

   We have made the requested change.

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?

   We can confirm that the tests used were able to recognize the different clades of HIV-1 commonly seen in Ethiopia. There were no HIV-2 positive samples. We have added this information to the manuscript.

Discussion:
p.9. 4th Paragraph: You cannot "identify" false-positive results, use "yield" instead.
We have made the change suggested.

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).

Thank you. We have adjusted this sentence.

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".

We have removed this sentence.

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.