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: 3
Date: 2 December 2014

Author's response to reviews: see over
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: 2 Date: 20 November 2014
Reviewer: Jörg Schubach
Reviewer's report:
The revised manuscript is still difficult to read, and remains unclear in many aspects.
One major difficulty arises from using the term "tiebreaker". While WHO uses this term for the 3rd decisive test after discrepant results of two initial tests always performed in parallel, the "tiebreaker strategy" implemented in Ethiopia apparently is a sequential testing approach in which only samples reactive in the first test undergo further testing. Why this should be fundamentally different from what you call a sequential approach (with a confirmatory test at the end) remains unclear. What you actually compare are variations of sequential testing algorithms – at least that is what you describe under Methods. Your confirmatory test has no other role than that of a tiebreaker. Or does Table 3 actually show results of initial parallel testing? I really do not know what you mean! And this goes throughout the paper.

Our testing strategy is serial for all algorithms including the tiebreaker as described in the first paragraph of the paper, and later in the methodology. We have now also included explicit mention that the tiebreaker is 3 tests in series in the abstract.

We are not aware of any definition that states a tiebreaker algorithm is always parallel testing despite reviewing again the relevant WHO guidelines. The common strategy recommended by WHO in resource limited settings is serial testing rather than parallel (as per the older Strategy I,II,III and Figure 6 & 7 in the 2012 guidelines). Therefore in most cases, a tiebreaker is referred to in the context of serial testing in WHO documentation. We did find a reference to the use of tiebreakers in the 2002 guidelines for test evaluation\(^1\), and here both scenarios of tiebreaker with parallel and serial testing are explicitly referenced. Our approach has therefore been to define clearly in the background and methodology section and now also the abstract, what is the definition used for tiebreaker in our study, which by our reading, appears to be consistent with that in most common usage.

Table 2. You should also add the OIC to this table (and perhaps also WB!).

We have added the performance characteristics of OIC to the table.

Table 3. I don't understand what "Discordant" means when you use a

\(^1\) Department of Health and Human Services/Centre for Disease Control and Prevention. Guidelines for Appropriate Evaluations of HIV Testing Technologies in Africa. 2002
confirmatory test at the end of the testing sequence. Was the end result correct or not? What was discordant? Was it not possible to clarify with the OIC? I also think that Table 3 is unnecessarily complicated. What interests me is how the different algorithms perform in resolving the 203 positive and the 225 negative results, as defined by the gold standard. One doesn't know when looking at the table in its current state. Why not make things clear and understandable?

We apologize for this misunderstanding. Discordant refers to the RDT algorithm results, not the confirmatory algorithm results. It is only samples with two positive RDT results that get a confirmation test. Discordant RDT results do not generally receive a confirmatory test (either OIC or WB) because the discordant result could be due to early seroconversion, which a serological confirmation test such as WB or OIC will not resolve. As well, it is not common practice to resolve discordants with a serological confirmation test because in the setting of cross-reactivity, the WB or OIC will be indeterminate. Therefore clients with discordant RDT results are asked to return for follow up testing---in line with WHO guidance. This is discussed on page 41-2 of the latest WHO guidelines:

Using the testing result of A3 as a tie-breaker may not always resolve HIV status. Possible reasons for the discrepant result (A1+; A2−) could be poorer than expected specificity of A1 or early seroconversion detected by A1 but not by A2—i.e. A1 has better seroconversion sensitivity than A2.

We have included the discordant RDT results in the table, so that the reader can see how each of the 203 positive and 225 negative samples were resolved by gold standard. We have added a footnote to the table to explain this.

Table 4. I do not understand why, with a specificity of 100%, the PPV can be less than 100%. A specificity of 100% means that you do not have any false-positive samples. Thus, as per definition, PPV = TP / (TP+FP), PPV will equal TP/TP, i.e. 1 = 100%.

This reflects the fact that the results are rounded to one decimal point. For example, in the first line of Table 4, the specificity result is 99.96% which rounds to 100.0. This is consistent with Table 3, where it can be seen that there is a single false positive result. We considered keeping multiple decimal points in the results, but decided against it to improve clarity. As the reviewer has pointed out in the first review, the volume of figures is quite large and we wanted to improve readability. We do include the CI, so that it is clear this is an estimate with a degree of uncertainty.

Table 5. As long as the issue of Table 4 is not resolved I do not trust Table 5.

As all the false positives included at least one weak positive, this table shows that both

---

specificity and PPV are 100%.

Table 6. Overlap with Table 5: if the multitest algorithms with strong positive signals have good PPV and NPV, the individual tests must also have good PPV and NPV.

We agree this is the case when the results of the individual tests are concordant. However the relationship between test specificity and algorithm specificity changes if the results are discordant. The tiebreaker algorithm is a mix of concordant (test1+ & test2+) and discordant results (test1+, test2 -, test3+) resulting in an algorithm that can perform poorly despite good individual test specificities.

To explain further, one can think of a 2 test algorithm conducted in series. If the algorithm states that all Test 1 positives are re-tested with Test 2, the net specificity of the algorithm increases with a consequent decrease in sensitivity. However when the algorithm is to give the second test to those that are negative on the initial screen, then sensitivity improves (you find more positives), but specificity decreases (you have more false positives).

The univariate analysis is of little value. You cannot draw any conclusions from it You should do a multivariate analysis (logistic regression) to sort things out or otherwise leave the table away.

We accept that it is only possible to draw inferences and no conclusions from this table as we state in the paper. We have not done multivariate analysis because the overall numbers are too small for the model to be stable.

We would argue to keep this table due to the scarcity of information published on this topic. We are aware of very few papers that test associations between false positives and other factors----including those that report only univariate analysis. It is challenging to have large enough sample sizes to do this kind of analysis, therefore we feel this information can help direct future research into this topic.