Author's response to reviews

Title: Diagnostic performance of line-immunoassay based algorithms for incident HIV-1 infection

Authors:

Jorg Schupbach (schupbach.jorg@virology.uzh.ch)
Leslie R Bisset (leslie.bisset@swissonline.ch)
Martin D Gebhardt (martin.gebhardt@bag.admin.ch)
Stefan Regenass (Stephan.Regenass@usz.ch)
Philippe Burgisser (philippe.burgisser@bluewin.ch)
Meri Gorgievski (Meri.Gorgievski@ifik.unibe.ch)
Thomas Klimkait (Thomas.Klimkait@unibas.ch)
Corinne Andreutti (C.Andreutti@lasource.ch)
Gladys Martinetti (gladys.martinetti@ti.ch)
Christoph Niederhauser (christoph.niederhauser@bsd-be.ch)
Sabine Yerly (Sabine.Yerly@hcuge.ch)
Stefan Pfister (stefan.pfister@ifik.unibe.ch)
Detlev Schultze (detlev.schultze@zlmsg.ch)
Marcel Brandenberger (Marcel.Brandenberger@synlab.com)
Franziska Schoni-Affolter (Franziska.Schoni-Affolter@chuv.ch)
Alexandra U Scherrer (alexandra.scherrer@usz.ch)
Huldrych F Guenthard (huldrych.guenthard@usz.ch)
Swiss HIV Cohort Study SHCS (noname@noreply.ch)

Version: 2 Date: 23 February 2012

Author's response to reviews: see over
We wish to thank the two reviewers for their critical and unbiased assessment of our paper and for their suggestions!

Reviewer: Ming-Wei Lin

Major Compulsory Revisions:

1. It is not clear why the authors only used patients with incident infection for determination of the diagnostic sensitivity of the Inno-Lia algorithms, while used patients with older infection for determination of the diagnostic specificity of the algorithms. It will be necessary to calculate the information of sensitivity, positive predict value (PPV), and negative predict value (NPV) in both patient cohorts with incident and older infection, respectively.

Please consider the basic definitions of diagnostic performance: The diagnostic sensitivity of a test is the proportion of samples that have a positive test result among a sample population that should be positive. Therefore, in order to determine the diagnostic sensitivity of an algorithm for incident infection, one uses only patients with incident infection, in our case those infected for less than 12 months. Conversely, the diagnostic specificity is the proportion of samples testing negative among those that should test negative, and in order to determine the specificity of an incident infection algorithm, one has to test exclusively patients with older infection, i.e. > 12 months. We thus refuse to meet this request.

Dr. Ming-Wei Lin's further request that PPV and NPV be determined would make sense for tests that are used for individual diagnosis (e.g. answering the question whether a patient is infected by HIV). There is ample agreement, however, that STARHS, because of their low sensitivity and specificity, are not such tests and should only be used for population-based studies. This was clearly stated in the limitations (last sentence of the paper):

"Finally, as the diagnostic performance of the method, particularly the sensitivity of the algorithms, does not meet the standards for tests used for diagnostic purposes, the method is unsuitable for diagnosing incident or older infection in individual persons. The method should only be used in population-based studies."

Thus, as in our previous paper published in BMC Infectious Diseases, we refuse to add the PPV or NPV to the paper — it simply wouldn't make sense. However, since new molecular methods are now being developed which in the future may allow a more reliable differentiation of incident and older infection, we have added new text to the first paragraph of the introduction, plus 3 new refs, as follows:
The performance of STARHS, i.e. the sensitivity and specificity with which they recognize or exclude an incident infection in an individual patient is low and does not meet the standards required for tests used for diagnostic purposes. Therefore, STARHS should not be used for individual diagnosis. Recently, new procedures based on HIV genetic diversity, as detected by single-genome analysis, have been developed, which in the future may lead to more reliable results also enabling diagnosis of incident infection in individual patients [13-15].

2. Were the patients with older infection used for evaluation of specificity the same as those patients in the previous study published in 2011 BMC Infectious Disease? It should be clarified and addressed in the manuscript.

This point was addressed in Methods under the header "Patients with older infection", but only with a citation and thus perhaps not with sufficient clarity. We have thus modified this text as follows:

The SHCS patients, already investigated in a previous study [36], had been infected...

Furthermore, we have added a footnote to the respective section of Table 1:

* same patients as investigated in [36]

3. A total of 26 algorithms (Algs) for incident HIV-1 infection were developed and employed in the study, however, the twenty-six algorithms are not completely independent. It may be problematic to apply this approach to evaluate the diagnostic performance of Inno-Lia antibody patterns simultaneously.

It is true that many of the algorithms are similar, and caution is thus necessary when statistically assessing differences in incident infection rates between groups. We have therefore introduced some cautionary remarks:

in Results / Application of Algs for estimating the annual incident infection rates (IIR)
For 2008 (cohort B), with each model, all 10 algorithms indicated an increase of IIR averaging 30.6%, which was highly significant (P < 0.0001; two-sided paired t-test). It should be considered, however, that some of the algorithms are similar and thus do not yield truly independent measurements.

In the Discussion / Advantages of using a combination of algorithms: Addition of the following sentence at the end of the paragraph:

It should also be considered, however, that the algorithms are not truly independent of each other, and this has to be kept in mind when performing statistical evaluations.

4. The authors suggested the use of the 10 best-performing algorithms in combination to minimize the impact of their individual limitations and the effects of sample bias and to monitor changes in IIR in a population over time. The cost-effectiveness should be considered and evaluated in the manuscript. Moreover, the diagnostic sensitivity remains far below the standards for tests used for diagnostic purposes. It would be advisable to combine other assays with high sensitivity.

In some countries, including ours, this test is routinely used (i) to confirm HIV infection and (ii) to differentiate between HIV-1 and HIV-2. This is important because HIV-2 requires other methods of viral load monitoring and antiretroviral treatment than HIV-1. The most frequently used VL tests, the Roche Cobas TaqMan and the Abbott Real Time PCR assay for HIV-1 do not recognize HIV-2. Regarding ART, HIV-2 is naturally resistant against NNRTIs. It is thus necessary to diagnose HIV-2 before starting with viral load monitoring or ART and that's why, in Switzerland, since 2006 we test all newly diagnosed patients with the Inno-Lia. The costs to the health insurances are about half as much as for a viral load test, and in view of the fact that we do 4 viral load tests per year for the rest of the life of a patient, these costs are insignificant. We believe that this initial expenditure is justified because the cases of HIV-2 viral load underestimation leading to false treatment decisions, false treatment leading to resistance development etc. will be avoided. The consequences of unrecognized HIV-2 infection would in the end be more expensive than these initial Inno-Lia, not to mention the damage done to the HIV-2 infected patients.

Please consider also that the additional evaluation of the Inno-Lia results by the algorithms comes at no extra costs. The population-based evaluation is done by the federal health authority, and it is an automated procedure which requires no more than a simple Excel sheet. Whether only one or 26 different algorithms are used makes no difference. In order
to make this clear to the reader we have added information in several places. This is re-
iteration of information provided in earlier papers, but you asked for it:

In Background, 3rd paragraph:

We have previously shown that a patient’s antibody reaction in a widely used commercial
line immunoassay, the Inno-Lia™ HIV I/II Score (Inno-Lia), provides information on the
duration of infection similar to that of a commercial enzyme immunoassay (EIA), the so-
called BED Incidence EIA [8, 16]. The Inno-Lia is a type of second-generation Western blot
(WB) that measures antibodies to different HIV antigens in a semi-quantitative way and is
used for confirming HIV infection and to differentiate between HIV-1 and HIV-2 [17, 18].
Timely diagnosis of HIV-2 is important, because this virus requires different tests for viral
load quantification than the widely used and FDA-approved tests from Roche, Abbott,
BioMérieux, or Bayer. Furthermore, HIV-2 treatment requires different antiretroviral drug
regimens, as the virus is naturally resistant to some frequently used drugs including the
whole class of non-nucleoside reverse transcriptase inhibitors (NNRTI) [19-22]. In some
countries, the Inno-Lia is thus used routinely at the time of diagnosis, and in Switzerland the
test has become a mandatory confirmatory test for HIV in 2006 [23].

In Background, end of 4th paragraph:

Of note, the utilization of the Inno-Lia results for population-based studies of HIV-1
incidence comes at no additional costs — no additional test is needed.

In the Methods section, a short new paragraph was introduced:

**Collection of Inno-Lia results by means of HIV notifications**

Since September 2007, all newly diagnosed HIV patients are notified to the SFOPH by means
of an electronic, Microsoft Excel® based form. Anonymized personal identification and all
available diagnostic data of each patient including the detailed Inno-Lia results were entered
into this form by the 11 HIV notification labs or the SNCR and forwarded by e-mail to the
SFOPH. At the SFOPH, these data were transferred into a database subsequently used for
evaluation by the different incident infection algorithms.

5. It is not clear if Cohorts A, B, C, D were came from the same cohorts? If they were not the
same cohorts, it will be inappropriate to use paired t test to make comparisons.

The cohorts A, B, C, and D are indeed different, as each represents patients newly diagnosed
during different time periods. Nevertheless, the use of paired t test is appropriate because an unpaired test would not take into account the fact that the four results of a given algorithm are linked: a datapoint in cohort A has to be compared only with the datapoints in cohorts B, C and D that belong to the same algorithm. This can only be achieved with a paired test.

No changes made.

6. What is Fisher's r to z test? It is quite confused why the authors evaluate correlation by Fisher's r to z test.

Fisher's r to z transformation is used in order to transform the distribution of the Pearson’s correlation coefficient r, which is not normally distributed, into a variable z', which is normally distributed:

\[ z' = 0.5 \times \ln(1+r) - \ln(1-r) \]

This transformation leads to an improved assessment of the significance of a correlation. Fisher’s r to z is the standard parametric test for correlation contained in the StatView package.

Sorry for the small imprecision. I have modified the text as follows:

Means were compared by paired or unpaired t-tests, as specified under Results, and correlation was assessed by Pearson's test using Fisher's r to z transformation.

Minor Essential Revisions:

1. The references should be modified to have a consistent format.

This is just a formatting issue. We have tried to avoid it this time

Quality of written English: Not suitable for publication unless extensively edited.
Please consider that the coauthors including Leslie R. Bisset, who is a native speaker of English, already corrected and amended the paper where they deemed it necessary. While we cannot exclude that there were still some isolated typing errors, we think that this comment is not justified in such severity. Nevertheless, we have gone over the paper again and have tried to improve it further and hope that we can meet your high standards this time.
Reviewer: Maurizio Zazzi

Reviewer’s report:

Major Compulsory Revisions

None

Minor Essential Revisions

1. This study extends and in some respect complements two previous studies by the same group. The first (Schüpbach et al., PLoS Med 2007) reported the use of the INNO-LIA assay as a system to discriminate recent (<12 months) vs. chronic infection. The second (Schüpbach et al., BMC Infect Dis 2011) showed that the specificity of the system was generally unaffected by HIV-1 subtype, viral load, CD4 counts and disease stage yet specificity decreased when testing patients under successful HAART and with increasing patient age. The novelty of the present study appears to be the selection of a validation dataset which I understand has no overlap with the one used for training the system in the first study (Schüpbach et al., PloS Med 2007). This dataset is characterized by having more defined information about the estimated date of infection. However, both this and the previous study indicate that the INNO-LIA system is particularly suitable to detect a ‘recent’ infection when the infection occurred in the early phase of recent infection. It would be advisable to highlight the new knowledge conveyed by this paper in the discussion section.

Indeed, the dataset used here has no overlap with that in the PLoS Med paper, and we now mention this in the last paragraph of the Introduction:
The second study, presented here, now addresses the diagnostic sensitivity of the Inno-Lia algorithms in a new cohort of patients infected for less than one year.

A further modification is in the 1st paragraph of the Discussion:

The purpose of the present study was to further assess the diagnostic performance of Inno-Lia based algorithms for incident HIV-1 infection and to see whether previous estimates for sensitivity and specificity would be confirmed in a new, much larger dataset of newly diagnosed patients with no overlap to our first study [16]. In a recent study of patients known to have been infected for at least one year, we have demonstrated a high diagnostic specificity of the algorithms, which was not affected by the HIV-1 clade or the disease stage [36].

2. Page 7. The authors say that they conducted two studies to address whether any virus or patient related factors can impact the accuracy of the system. Then, they refer to only one of these two studies. The reader may miss what the second study showed in this context.

see the first part of the preceding response

3. Page 9. The inclusion of “Documented signs of ARS no more than 90 days before diagnosis”, in the absence of documented seroconversion (definition #3), as a proof of recent infection among the HIV notification subset appears to be much less accurate than the other criteria. As cited by the authors themselves, the paper by Hecht et al. (AIDS 2002) indicated that fever and rash are the most significant predictors of primary HIV infection. Did the authors consider the combination of these two signs to classify recent infection by definition #3 in their dataset? In general, the algorithms were more specific than sensitive. Also, a large proportion of reference recent infection cases were contributed by the notification registry. How does sensitivity change if definition #3 is removed from
the criteria to classify recent infections? It is unclear to me whether adjusted sensitivity according to model 3 of table 3 addresses this issue.

Unfortunately, the notifications do not contain detailed information on the type of symptoms of PHI. Physicians are offered a link to a list of the symptoms and then, back in the notification form, have the choice of ticking YES or NO. In case of YES, they should also provide the date when symptoms started. Thus, we cannot answer your first question except that definition #3 includes the cases of fever and rash.

The second question you ask is indeed answered by model 3. Here, all cases ruled incident due to the presence of PHI symptoms alone (i.e., no combination with a previous negative test) were removed from the analysis. In order to make this clearer, we have modified the respective text under Results p. 14, 2nd paragraph:

In the third model, we excluded all cases judged incident because of reported signs or symptoms of ARS (incidence definition 3 for the HIV notifications in Methods) and considered only the notifications with a previous negative HIV test. Model 3 thus did not select for symptomatic acute infections.

4. Page 12. The inclusion of the “94 samples with a less precise EDI” in the sub-analysis of the performance of the algorithms after stratification by quarter seems incorrect. These samples have indeed uncertainty right in the parameter used for stratification. I suggest to do the analysis without these cases or confirm that they do not bias the results.

If we remove these patients from among the notifications, i.e. restrict the evaluation to the patients who had either a negative test up to 12 months or a date of PHI up to 3 months prior to diagnosis, the notification patients will contribute zero cases to quarters 3 and 4, and the only patients remaining are those 15 of the ZPHI in whom the sensitivity in quarters 3 and 4 was zero, as stated on p13. If we assume a sensitivity of zero for these quarters, the adjusted sensitivities S1, S2 and S3 are as shown in the Excel tables below.
Please note, however, that if we totally exclude the 94 samples, we could no longer estimate how many infections are diagnosed in the later quarters of the first year of infection; this cannot be zero. Thus, instead of omitting the 94 cases from tables 3 and 4 we have added text at different places of the manuscript:

p14, end of 1st paragraph:

If the sensitivity in quarters 3 and 4 was set to zero, as suggested by the results of the 15 ZPHI patients, mean $S_1$ was reduced to 22.7%.

p14, end of 3rd paragraph:

If, again, the sensitivity in quarters 3 and 4 was set to zero, the mean $S_2$ was reduced to 47.9%.
A reduction to a mean of 32.8% was obtained when assuming a zero sensitivity for quarters 3 and 4.

When sensitivities calculated on the assumption of a zero detection rate in quarters 3 and 4 were used, the average IIR of the four cohorts increased to 0.63, 0.82, 0.77 and 0.64 when using $S_1$, to 0.26, 0.34, 0.32 and 0.26 when using $S_2$ and to 0.40, 0.52, 0.49 and 0.41 when using $S_3$. In all these instances, the relative differences between the four cohorts were as in Table 4. Thus, the relative changes between different annual cohorts were independent of the absolute value of the diagnostic sensitivity.

5. Discussion. The authors make it clear that the different algorithms provide (slightly) different results and are confident that using multiple algorithm has the advantage to minimize the impact of the inadequacy of any individual algorithm. However, it remains elusive to me how one should use all the algorithms to analyse a dataset in practice. Is there a preferential summary measurement derived from the individual measurements provided by the different algorithms?

Discretionary Revisions

This comment relates to comment #4 of Dr. Ming Wei Lin. Please see our response and the introduced modifications there.

6. Discussion. Although not uniformly supported by international and national guidelines, the use of molecular diagnostics has been replacing second-line serology as a confirmatory test for HIV diagnosis. It should be emphasized that a confirmatory antibody assay such as INNO-LIA has the added value of detecting recent infection, an information not provided by HIV DNA/RNA assays. Cost differences between the two approaches should probably also be mentioned.
This comment also relates to comments of Dr. Ming Wei Lin. Please refer to the modifications made under points 4. Regarding cost differences, this is a technical/methods paper. We have just added a sentence under Conclusions:

Our method is particularly attractive for countries in which the Inno-Lia is already used as a confirmatory assay and for differentiation between HIV-1 and HIV-2.