Reviewer's report

Title: Can HIV incidence testing be used for evaluating HIV intervention programs?

Version: 1 Date: 18 March 2010

Reviewer: Peter H Kilmarx

Reviewer's report:

Dr. Peter Kilmarx reviewed the paper together Drs. S. Michele Owen and Debra Hanson. We have combined our comments here, with apologies for any redundancy or inconsistencies.

The paper describes a novel approach of evaluating intervention programs using cross-sectional incidence assays. In particular the approach utilizes the concept of Incidence rate ratios to show differences between control and intervention groups.

HIV incidence testing algorithms were implemented and results were compared to those from longitudinal follow-up data to evaluate a male circumcision intervention. Samples collected from participants of the randomized controlled trial conducted in South Africa were used toward this objective. BED and avidity assays were performed.

In general approach appears valid and the data would be useful to others interested in using cross-sectional incidence (“recency”) assays for measuring intervention success.

One of the main uses of HIV incidence assays is to monitor and evaluate treatment and prevention programs. Evaluation of incidence estimates is paramount to understanding how well this method performs in the assessment of an intervention. The authors’ provide the proper framework by describing the challenges of incidence testing, i.e. misclassifications and the mean recency period (time between seroconversion and when an incidence assay cutoff value, selected to distinguish between recent and longstanding infection, is reached). Though it is important that the authors define the parameters of the incidence algorithms, describe the limitations, and compare advantages and disadvantages of both methods, the paper provides little motivation for comparisons at multiple cutoff values and window periods. If estimates of incidence are accurate, then cross-sectional incidence testing may be useful in evaluating interventions, provided the design, timing, sample size, and perhaps prevalence in the target population are accurately and appropriately defined. These are among the issues to address in this study. A comparison has been made to efficacy estimates using longitudinal analysis where confidence limits are extremely wide.

Major Compulsory Revisions
1. In describing the avidity assay it is not clear how the assay is performed. Specifically, it should be stated that each sample is loaded into 2 wells of the ELISA plate and the Avidity Index was calculated as the ratio of the OD values. OD of treated specimen/OD of untreated specimen x 100 = AI

2. An avidity index should never be reported as a value greater than 100% as this is biologically not relevant. A number of greater than 100% can be obtained but this occurs usually at the extremes of the serologic assay. The suggestion would be to report any value greater than 100% as 100%

3. It is not clear from the methods how samples that would be classified as negative based on the Bio-Rad Kit negative cutoff are handled for the AI calculation. For example, assume that the cutoff of the assay for negative vs positive is .280 and you obtain values of - for treated of .150 and untreated .275. The calculated AI would be 54.5 and would likely indicate a long term infection when in reality it would be a recent infection because the sample was negative in the assay and an AI could not be calculated. How this type of situation was handled in regards to the data presented should be addressed. It is important because it could influence the analysis done looking at individual assay performance results with time. Page 7. If samples were included that were in fact negative it would appear the slope over time would decrease when in fact it does not. This could also affect the number of false recent seroconverters calculated for the Avidity assay.

4. To assess the ability to use incidence testing from cross-sectional data to evaluate an intervention does require that the authors demonstrate the accuracy of the assay. Sensitivity and specificity were calculated. In addition, FPR estimates were computed from longitudinal data. Though the derivation provided in the annex is useful, it was not necessary to provide estimates for crude and adjusted incidence rate ratios. The paper would be better served if it focused the attention on the adjusted rate ratios computed from the two assays and the dual combination using the cutoff values that result in an estimated window period which compares as precisely as possible to the follow-up duration used in the original manuscript of 21 months. The presentation of estimates at multiple cutoff values, window periods, and with or without adjustment detracts from presentation of the methods and results related to the well-posed objective of comparing the two methods to evaluate the MC intervention as a way of determining whether cross-sectional data and incidence testing can be used for this purpose.

5. Based upon the PLOS Medicine manuscript published by Auvert et al in 2005, there were 3059 HIV-, 69 recent HIV+, and 146 prevalent HIV+ participants, representing the total 3274 that were randomized to the MC intervention trial. A subset of these totals was used in this study from subjects who completed the 21 month follow-up as identified in the annex. Comparisons of prevention effect should be made based upon this subset in which the efficacy was 62% instead of 60%. The original paper reports that “When considering only those participants who completed their M21 visit, the RR was 0.38 (0.22-0.67)”.

6. The mean follow-up duration of subjects used in estimation of the intervention
effect from the conventional longitudinal analysis was 21 months. Though it would be ideal to choose cutoff values that best distinguish recent from long-standing infection, for comparative purposes and to avoid use of the window period, cutoff values should be chosen to equalize follow-up and the window periods. This can be provided as the rationale for choice of cutoff values that result in an estimated window period of 21 months. Alternately, methods could be employed to annualize incidence or standardize to a common time period; or strong assumptions about the constancy of incidence and incidence ratios could be made.

7. To verify how accurately these assays predict incidence, they must be validated against an independent measure of incidence. Two or more assays for recent infection can be compared to each other, though limited by the fact that none may be considered gold standards. The ultimate validation of these assays requires testing of cross-sectional specimens obtained from cohorts for which there is conventional measurement of observed incidence. Agreement between two incidence assays using the Kappa statistic, in which neither assay can be considered a gold standard, is not informative. Agreement between each assay and true incidence is preferable. Compare the test results for each incidence assay and for the combined assay to the known status of those specimens. Report the Kappa statistic for agreement in these comparisons.

8. Provide one table of the results for BED, AI, and BED/AI. Show the cutoff value that results in a window period of 21 months, sensitivity and specificity, Kappa, each of the parameters use in the corrected IRR (NTR, N+, #2, and N-), and the corrected IRR with bootstrap confidence limits.

9. The fact that variance of the bioassay measurements increased over time suggests that a data transformation may be required if parametric methods are used to explore the relationship of these measures with time since seroconversion. How is this relevant to the methods used in this manuscript?

10. Ideally, the window period estimates and the false recent parameter should be calculated from independent data that are representative of the study population. If the same data were used to calculate the correction as was used for the estimation of incidence ratios, the estimates are biased towards a better outcome. This limitation should be discussed.

Minor Essential Revisions:

1. Page 2 In the abstract the complete name of the Bio-Rad assay should be used as there are multiple ELA's produced by Bio-Rad.

2. In results page 6- it is suggested that the results reported for the incidence assay be recent and not recent and not negative and positive. Negative and positive are confusing terms in that Negative indicates long-term and positive indicates recent infection.

3. In the discussion on page 10/11 it is assumed that increasing prevalence effects misclassification the same way in BED and the Avidity assay described. There are not sufficient data reported for this Avidity assay to make this generalization.
4. Page 11- It should be noted that the avidity assay used in this study is not that same as that described in reference 13 so this could account for the only moderate agreement between BED and the Avidity test described in the manuscript.

5. The abstract does not stand on its own. Provide the numbers of specimens collected at last follow-up and how they were used. Focus on the best estimates for the protective effect, instead of presenting crude and adjusted rates. State the best algorithm, estimates for the parameters, and provide the best estimate possible of efficacy. “Calculation of the effect was independent of the value of a window period.” This sentence should be included in the methods section.

6. Describe which subjects have the additional 596 follow-up observations.

7. The term window period or recency period should be defined up front. Prefer the wording “estimated” instead of “crude” window period.

8. Refer to the annex in the text for explanation of numbers of specimens and how they were used, and for methods related to the window period.

9. Page 5, the proportion of false long-term seroconverters should have been calculated from known recent seroconverters, whereas false recent seroconverters are calculated from known long-term seroconverters. Since the false long-term rate is not adjusted for in the IRR algorithm, this method and result is of less interest.

10. Are both methods subject to bias related to generalizability? Is there a cohort effect?

11. If incidence testing algorithms, avidity vs. BED vs. AI/BED, are to be compared with respect to their ability to accurately estimate the prevention effect, this should be stated in the introduction as an objective.

12. Comment on the comparability of the window period to the follow-up period as it relates to the comparability of the estimates of efficacy between the two methods.

13. It should be made clear in the methods (as is mentioned in the discussion) that HIV-infected individuals were also included in the cohort.

Discretionary Revisions
1. Vernacular such as “censored” on page 4 of the methods such be written for understanding by a broader audience.

2. Can the avidity index reach values higher than the maximum 100% (page 4 methods)?

3. It is not necessary to provide the crude IRR formula in the main text.

**Level of interest:** An article of outstanding merit and interest 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 and my co-reviewers have no competing interests.