WHO Technical Working Group on Statistical Approaches for Development, Validation and Use of HIV Incidence Assays

Place: Room D46031, D Building, WHO Headquarters, Geneva
Time: 22-24th April 2009

Meeting Report

Table of Contents

1. Background .................................................................................................................. 2

2. Meeting Objectives ..................................................................................................... 4
   2.1. Expected Meeting Outputs ..................................................................................... 4

3. Meeting Proceedings .................................................................................................. 4
   3.1.Bill and Melinda Gates Foundation .......................................................................... 4
   3.2.Overview of HIV Incidence Assays and Issues of Terminology ............................... 5
       3.2.1.Terms used to estimate incidence ................................................................... 7
       3.2.2.Other terms .................................................................................................... 8
       3.2.3.Uses of Assays to Detect Recent HIV Infection ............................................ 9
       3.3.Main Statistical Issues Related to Methods used to Estimate HIV Incidence ........ 10
           3.3.1.The McWalter/Welte Model ........................................................................ 10
       3.4.Reflections on the Incidence Window Period ..................................................... 13
           3.4.1.Better Methods are Required to Estimate the Incidence Window Period. .... 13
           3.4.2.Understanding the Variability of the Incidence Window Period.................. 13
           3.4.3.Estimation of Standard Error ...................................................................... 14
           3.4.4.Usefulness of Demographic and Health Surveys and other Cohort Studies 15
       3.5.Assay Validation Pathway ..................................................................................... 16
       3.6.Policy-Relevant Questions about Assays for Recent HIV Infection ..................... 19

4. Conclusions: ................................................................................................................ 19

5. Associated Technical Documents .............................................................................. 21

6. Bibliography ............................................................................................................... 21
1. Background

There have been considerable efforts made to develop HIV assays in order to detect recently-acquired HIV infection. Such assays would have three testing objectives: 1) estimation of the incidence of HIV infection at a population level; 2) detection of recent HIV infection at the individual level; and 3) for use in clinical intervention trial settings. Consensus is still required for the necessary statistical approaches required to optimize interpretation of results from existing or future HIV incidence assays to achieve these objectives.

Historically, assays were developed that provide a snapshot of the evolving immune response to HIV; individuals with an immature response are likely to have acquired HIV recently. To date, three assays have been commercialized, and several custom assays have been developed. When these assays were evaluated rigorously, it turned out that the imperfections in the assays were substantial. Used alone, these assays introduced large errors to incidence estimates—in some cases, overestimating incidence by 300%\(^1\). In particular, a number of people whose assay results indicated an immature immune response, had actually been infected for several years. Thus, this first generation of assays identified too many people with long-term infection as recent HIV acquirers, which came to be known as the “long-term specificity” problem. In 2005, UNAIDS released a statement suggesting that the most widely studied commercial assay, the BED assay should not be used for routine surveillance of HIV incidence.

Since 2006, a series of formal and informal meetings have taken place to discuss progress towards a validation protocol for existing and future HIV incidence assays. In 2008, WHO convened two meeting of experts (in Geneva and in Mexico City) to further the agenda. A draft guidance protocol for validation of HIV assays has been developed by the HIV Incidence Working Group. An outcome of the last meeting held in Mexico City was that a working group on statistical approaches be organised to identify the optimal statistical and epidemiologic approach to using HIV incidence assays. In addition, a grant was awarded by the Bill & Melinda Gates Foundation to Family Health International, Inc. (FHI) to review the market, technical, and epidemiologic challenges that were obstacles to wide available of an accurate HIV incidence assay.

Although progress has been made toward a final protocol for calibrating and validating incidence assays (Methodologic Guidance for Validation of Existing and Future HIV Incidence Assays, v5.0, October 2008), several key issues remain unresolved. Among those are:

\(^1\) Personal communication, Dr Bill Rodriguez April 2009
The statistical approach to using incidence assay results to estimate incidence in various populations has been discussed widely in the literature, but a consensus approach has not been defined and endorsed by a representative group of experts;

The population to be tested under each testing objective may have different characteristics that affect incidence test results. The approach to measuring incidence in different populations, especially those infected with diverse viral subtypes, has not been established;

There is a need to agree on the specifications of HIV incidence assays and to establish a standardized approach to validate their performance. These approaches include determination of the cut-off value for recent infection and agreement on the characteristics of the specimen panels for these assessments i.e. sample sizes, comparability with "gold standard";

There is a need to establish a consensus on the precision and margin of error that can be accepted when measuring HIV incidence at a population level;

Terms used by epidemiologists, public health officials, clinicians and lab experts are sometimes conflicting, and produce confusion. There is a need to clarify and standardize terms that are using a test for recent infection in the HIV incidence estimation process;

The level of HIV prevalence, incidence and trends may affect the variance coefficient of HIV incidence assays, and therefore sample size calculations need to be addressed;

There is a need to identify specimen panels of recent seroconverters (from vaccine trials, cohort studies, etc) that can be used in calibration and validation studies of HIV incidence assays.

Among all the issues mentioned above, in the Mexico City meeting (held in August 2008) it was agreed that one of the priorities was to understand and reach consensus on the statistical approach to incidence testing. Therefore, a small working group meeting was called in Geneva for a limited number of technical experts on this field.

The meeting was jointly organized by HSS/EHT/DLT and HTM/HIV/SIR. The main purpose of this meeting was to review statistical approaches that may be used to validate HIV incidence assays and estimate HIV incidence in populations.
2. Meeting Objectives

1. To determine the assay requirements for each of the testing objectives of HIV incidence assays
2. To determine the statistical methods required for each step of the assay development pathway
3. To determine statistical approaches for field validation of HIV incidence assays for each testing objective

2.1. Expected Meeting Outputs

1. Agreement on the pathway for development of HIV incidence assays
2. Agreement on statistical requirements for:
   a) assay development
   b) field evaluations
3. A draft document of statistical approaches for HIV incidence assays

3. Meeting Proceedings

The meeting was relatively informal with a mixture of short technical presentations, followed by lengthy discussion for each session.

3.1. Bill and Melinda Gates Foundation

The Bill and Melinda Gates Foundation (BMGF) has a strong interest in the wide availability of assays that can identify recent HIV infection accurately. BMGF supports numerous projects whose goal is to prevent HIV infection and reduce the incidence of new infections at a population level. Many of these projects necessarily measure HIV incidence as part of their self-assessment activities. Given the pitfalls of using existing assays to estimate incidence, there is clear need for global guidance on the use of data taken from HIV incidence assays. There is also a need to provide guidance for developers of new, more accurate incidence assays. Of particular interest to BMGF and its grantees are questions of what sample sizes are required to assess the impact of an HIV prevention intervention with reasonable confidence; what HIV incidence assays (and/or HIV incidence testing algorithms) can be recommended; what HIV incidence assays have been validated satisfactorily; what approach should be used to eliminate the error associated with poor long-term specificity; and how much should be budgeted for activities to estimate HIV incidence. In brief, BMGF expects that, in coordination with FHI, the WHO HIV incidence working group will provide clear guidance on how to estimate HIV incidence, along with information on the precision of this measurement, and the range of certitude or incertitude about the results.
3.2. Overview of HIV Incidence Assays and Issues of Terminology

While there are laboratory techniques for recent infection that detect RNA and other viral antigens, the bulk of effort around incidence testing has focused on serological testing. Other methods do exist but were not the primary focus of this particular meeting.

The key terms that are currently used by users and developers of HIV incidence assays are presented in figure 1 and illustrate the confusion that incorrect terminology presents in the field. Some terms are incorrectly inter-changed, while other terms related to a clinical rather than biological (or laboratory-based) definition.

![Figure 1 - HIV incidence terminology](image)

An overview of the basic principles of serological testing for recent HIV infections was given and comparison of the currently available assays made.
A number of possible confounding factors that explain the different results by different assays. Those are:

- Some individuals may never become reactive on currently available incidence tests (or non-reactive in the case of the IgG3 assay), i.e. some individuals would never be identified as recently infected. It is important to understand how large this proportion of people would be in any given population that is used to deduce an estimate of HIV incidence.

- Specificity of STARHS assays i.e. misclassification of individuals as recently infected when they are not, may be affected by:
  - HIV-1 clade (incidence window periods differ in all assays)
  - AIDS
  - Low CD4+ T cell (Hargrove et al., 2008)
  - Anti-retroviral therapy (Cimerman et al., 2007; Killian et al., 2006)
  - Elite suppressors (Fisher et al., 2007; Hatano et al., 2009)
  - Total IgG concentration? (specifically with the BED assay– no firm evidence)

- Factors may be additive
- Case-based exclusions to counter misclassification
- Apply correction factors (ε) when estimating Incidence using assays with poor long-term specificity

---

### Figure 2 - Overview of existing serological assays to detect recent HIV infection

<table>
<thead>
<tr>
<th>Assay</th>
<th>Source</th>
<th>Principle</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calypte BED EIA</td>
<td>Commercial</td>
<td>Proportion of total antibody that is HIV-specific</td>
<td>Parekh BS, et al., AIDS Res Hum Retroviruses 2002; 18.</td>
</tr>
<tr>
<td>IDE-V3 EIA</td>
<td>In-house</td>
<td>Relationship of reactivity with selected HIV antigens</td>
<td>Barin F, et al., J Clin Micro 2005; 43</td>
</tr>
<tr>
<td>IgG3 isotype</td>
<td>In-house</td>
<td>Indirect EIA for IgG3 antibodies to p24 Ag</td>
<td>Wilson KM, et al. AIDS 2004; 18</td>
</tr>
<tr>
<td>InnoLIA HIV</td>
<td>Modified commercial</td>
<td>Relationship of reactivity with various HIV antigens</td>
<td>Schöpbach J, et al J Clin Micro 2005; 43</td>
</tr>
</tbody>
</table>
The transferability of certain in-house assays (or commercialized methods with extra treatment step i.e. avidity index) is under investigation. There is data to suggest that the performance of the same assay performed in two different sites is not always completely concordant, with reported discordances of 12.5% (for comparison of IDE-V3 performed in London and in Tours for n= 1733) to 19.1% (for Avidity-adapted AxSym performed in Italy and in London). These preliminary data suggest that assay results may not be reproducible, an issue that requires further study. This does not necessarily mean that these assays are not useful; if a single, centralized laboratory (i.e. the NRL or a regional laboratory) always performs all of the testing for recent infection, irrespective of the type of assay, then there would there be the possibility to eliminate variability and increase consistency of results.

The following definitions were debated by the group and proposed for review of the wider WHO HIV Incidence Assay Working Group at the meeting in Cape Town (16-17 Jul 2009).

### 3.2.1. Terms used to estimate incidence

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recent Infection</strong></td>
<td>A combination of specific laboratory and/or other methods that is intended to classify individuals as positive or negative for the state of ‘recent infection’, for the purposes of estimating HIV incidence. The test would usually only be performed on specimens from people who have been classified as HIV positive by a screening test of some kind.</td>
</tr>
<tr>
<td><strong>Recent Testing Algorithm (RITA)</strong></td>
<td>An ‘ideal’ RITA is one which has the property that all individuals transition from the RITA positive to the RITA negative state, and remain RITA negative until death.</td>
</tr>
<tr>
<td><strong>Recent infection</strong></td>
<td>A transient period soon after HIV Exposure. The rate at which the susceptible population enters this transient state is the incidence of HIV infection. Its duration varies across individuals and depends on the method used for detection.</td>
</tr>
<tr>
<td><strong>Incidence</strong></td>
<td>The incidence of a disease is the rate at which new cases occur in a population during a specified period of time. The natural units for expressing incidence are “probability of an infection occurring, per person, per unit time”.</td>
</tr>
<tr>
<td><strong>Annual risk of infection</strong></td>
<td>An alternative way of expressing incidence – the probability that an individual will become infected if subject to a constant incidence for a period of one year.</td>
</tr>
<tr>
<td><strong>Assay for recent HIV infection</strong></td>
<td>A laboratory method or assay (used either alone or as part of a testing algorithm) to classify HIV infection as recent, for the purposes of estimating incidence.</td>
</tr>
</tbody>
</table>
| **RITA non-progressors**                  | Individuals that perpetually stay RITA-positive. It appears that for most proposed recent infection tests, there are subpopulations of individuals
who never develop the RITA-negative biomarkers. Some time after infection, members of this anomalous subpopulation may be regarded as being ‘falsely’ RITA-positive.

Not to be confused with disease non-progressors.

**RITA regressors**

Individuals that transition from RITA-positive to RITA-negative as expected, but then revert to being RITA-positive at some later time, after which they may be regarded as being ‘falsely’ RITA-positive.

**RITA false positive rate (FPR)**

The fraction of non-recent infections that are falsely classified as RITA positives as a result of RITA non-progression and RITA regression. Under specific assumptions, this may be an intrinsic property of the RITA, but in general, it is a property like the Positive Predictive Value of a diagnostic test, which depends on a combination of factors like assay characteristics and the population on which the test is used, and hence may vary from place to place and from time to time.

**RITA specificity**

The fraction of non-recent infections that are classified RITA-negative.

\[
\text{RITA specificity} = 1 - \text{RITA FPR}
\]

**Mean window period for incidence**

The mean duration of the state of recency, which critically depends on the specific RITA being used. In the non-ideal case, where there are either RITA non-progressors or RITA-regressors, or both, this must be carefully defined, and attempts to measure it at the population level, in ‘calibration studies’, must take the precise definition into account. It must essentially be defined as the mean time spent in the first sojourn in the RITA-positive state, for individuals who in fact progress to the RITA-negative state.

**Assay threshold**

A critical value, chosen for an assay which yields a quantitative result, to produce a categorical value for use in a RITA. For example, normalised optical density below a critical value can be used to turn the BED assay into a stand-alone RITA.

### 3.2.2. Other terms

**Acute/primary infection**

The period immediately after HIV infection when the individual is highly infectious and may or may not show clinical symptoms. This is often clinically important at the individual level. The individual may display RNA reactivity/positivity, with or without appearance of serological markers, i.e. antibodies to HIV. To eliminate confusion, this term should not be used when referring to recent infection for the purpose of determining estimates of incidence.

**Established infection**

Infection that lasts for more than the mean incidence window period. It is greater than approximately 6 months post HIV infection, and may also include long-standing infection. This definition is used generally to contrast it with acute or primary infection. To eliminate confusion, this term should not be used when referring to estimates of incidence.

**Eclipse period**

The period of time after HIV acquisition when no viral or immunological markers are detectable. The appearance of a viral marker, e.g., RNA, DNA, or p24 antigen, signals the end of the eclipse period.
Diagnostic window period: The period of time post HIV infection when no serological markers are detectable. It is used most often to identify a population of individuals who are newly infected, but undiagnosed by standard HIV tests. The appearance of a serological marker, i.e. antibody, signals the end of the diagnostic window period. The diagnostic window period is dependent of the assay used to measure serologic markers of HIV infection, and particularly on the analytical and seroconversion sensitivity of that assay. The eclipse period and diagnostic window period may overlap, with the diagnostic window period being longer.

Seroconversion: The appearance of serological markers (marking the end of the diagnostic window period), is dependent on the diagnostic assay used.

3.2.3. Uses of Assays to Detect Recent HIV Infection
There are three distinct uses of assays to detect recent HIV infection, as detailed in Table 1 below. In particular, epidemiologists are interested in monitoring estimates of HIV incidence, over time or between different population groups, in order to impart such valuable information to policy-makers. Groups running or funding clinical trials of candidate interventions such as therapeutic vaccines, microbicides, etc would like to know if the intervention has an impact or not. A clinical trial can only be effective, if the study end-point is reached quickly and the sample sizes are appropriate and thus it is most cost-effective to place intervention trials in populations with high incidence.

Table 1 - Potential Uses and Characteristics of Assays to Detect Recent HIV Infection(s)

<table>
<thead>
<tr>
<th>Assay Purpose</th>
<th>Tracking HIV Infection: Surveillance</th>
<th>Monitoring the impact of interventions as part of clinical trial</th>
<th>Selection of cohort for intervention clinical trial</th>
<th>Individual Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Population</td>
<td>General Population</td>
<td>Clinical trial cohort</td>
<td>Potential clinical trial cohort</td>
<td>Individual</td>
</tr>
<tr>
<td>Rationale for Use</td>
<td>To measure incidence at a population level or in specific risk groups</td>
<td>To measure change in incidence over time with and without an intervention</td>
<td>To measure incidence for selection of study sites/cohorts for intervention studies</td>
<td>Partner notification &amp; tracing</td>
</tr>
<tr>
<td>Target Population</td>
<td>General population High risk group ANC, etc</td>
<td>High risk groups</td>
<td>High risk groups</td>
<td>General population High risk groups</td>
</tr>
<tr>
<td>Sample size for baseline measurement</td>
<td>5,000 - 10,000 TBD</td>
<td>1000’s TBD</td>
<td>1,200 - 2,400 TBD</td>
<td>TBD</td>
</tr>
<tr>
<td>Sample size to monitor trends over time</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
</tr>
<tr>
<td>Expected Turn Around Time of Laboratory Testing</td>
<td>1 month</td>
<td>1 month</td>
<td>1 month/ depends on</td>
<td>A few days</td>
</tr>
<tr>
<td>Specimen receipt</td>
<td>Batched</td>
<td>Batched</td>
<td>Batched</td>
<td>Immediate</td>
</tr>
<tr>
<td>Study duration</td>
<td>6 months</td>
<td>With interim analysis</td>
<td>12-18 months, with interim analysis</td>
<td>Immediate</td>
</tr>
<tr>
<td>Desirable specimen type/assay format</td>
<td>Serum/plasma, DBS,</td>
<td>Serum/plasma, DBS</td>
<td>Serum/plasma</td>
<td>Fingerprick WB, serum, plasma</td>
</tr>
</tbody>
</table>

Key: TBD denotes still to be decided by statistical power and sample size calculations, it will depend on two parameters: the accuracy of the HIV incidence assay and the levels of incidence and prevalence.
3.3. Main Statistical Issues Related to Methods used to Estimate HIV Incidence

Reliable estimation of HIV incidence from the prevalence of recent HIV infection remains elusive because there are fundamental problems of expressing historical incidence in terms of instantaneous population state variables; substantially imperfect long term specificity; and a lack of locally determined calibration parameters.

A new solution has been proposed to analyse the relationship between HIV incidence and the prevalence of an experimentally defined state of 'recent infection'. Within any population, there will exist a subgroup (fraction) of individuals that will be defined by a laboratory method (or testing algorithm) as recently infected for all time i.e. assay non-progressors, even if they are long-term infected. The model is described below (McWalter and Welte, in press).

3.3.1. The McWalter/Welte Model

In the perfect and simplest world, we could expect a population of susceptible individuals (i.e. the healthy population) to become infected with HIV and progress to the 'infected below the threshold' population ($P_u$) i.e. RITA positive. There members of this population will remain for a period of time, with each member's time reflecting a draw from a probability distribution $f_B$, before moving to the 'infected above the threshold' population ($P_o$) i.e. RITA negative. The waiting time to go over the threshold is the incidence window period and is assay-specific.

$$H(t) \xrightarrow{I(t)} P_u \xrightarrow{f_B} P_o$$

- $H(t)$ = healthy population (susceptibles)
- $I(t)$ = incidence
- $P_u$ = population under threshold
- $P_o$ = population over threshold
- $f_B$ = waiting time for an individual to pass from under to over the threshold

With probability $P_{NP}$, some individuals fail to progress over the threshold defining recency. There is an accumulation of non-progressing long-infected individuals in the under-threshold category.

Figure 3 - Basic model of disease progression ($f_B$ is a density)
denoted as NP. It is necessary to know what this will mean for determination of $E[\pi_U]$ and $W$, and hence $I_w$. Progressing individuals enter the under-threshold category denoted as $P_U$ and will transition to the over-threshold category denoted as $P_O$, with the same dynamics as before. If one assumes, the same survival function for assay progressors and assay non-progressors, then one can introduce two categories $NP_R$ and $NP_L$ for recent-infected and long-infected non-progressors. Non-progressors will transition from $NP_R$ to $NP_L$ with waiting times drawn from $f_R|A = f_{PU}|A$.

Where:

- $H(t) =$ healthy population (susceptibles)
- $NP_L =$ non-progressors long-term
- $NP_R =$ non-progressors recent
- $P_{NP} =$ probability of non-progression
- $NP =$ non-progressor i.e. falsely identified as recently infected
- $NP_R =$ non-progressor, truly recent
- $NP_L =$ non-progressor, truly long-term
- $P_U =$ proportion of individuals under the threshold i.e. progressors
- $P_O =$ proportion of individuals over the threshold i.e. progressors
- $P_{U'} =$ proportion of individuals that regress back to apparent recency
- $E[\pi_U]$ = waiting time under threshold
- $W =$ statistical weight
- $I_w =$ HIV incidence point estimate

Figure 4 - Model that includes a subpopulation of non-progressing individuals, circled in red are the individuals who are truly recently infected.
This model will yield a formula for determining incidence, even when there is a substantial $P_{np}$ i.e. proportion of apparent non-progressors that have actually been infected for a long time. The ideal method for incidence estimate would use $P_u$ and $NP_R$, but the individuals in $NP_R$ cannot be distinguished from those in $NP_L$. However, if one can assume that both progressing and non-progressing individuals have similar survival times post-infection. Then, separate the non-progressors into two categories: $NP_R$ (the recently infected) and $NP_L$ (the long infected). The $f_B$ (mean waiting time in state of recency) is assigned to $NP_R$ and the simplified equation below is elaborated.

\[ I_w = \frac{U_0 - e O_0}{H(t_0) E[\pi_U]} \]

Where:
- $I_w$ = HIV incidence point estimate
- $U_0$ = population under threshold at time 0
- $O_0$ = population over threshold at time 0
- $H(t)$ = healthy population (susceptibles)
- $e = P_{NP}$
- $E[\pi_U]$ = expected time for a progressor to pass from under to over the assay threshold

The described model gives the equation for an entire population but often incidence estimates are generated using a sample of the population of interest. This may lead to a certain amount of uncertainty/bias in the resultant estimates. The performance of a BED-like assay (i.e. with $P_{NP}$=5%) can be graphically represented as a function of the sample size required against the incidence in the population and this was most useful for the discussions. Computation and graphs of the probability of correctly inferring an incidence decrease are still under review and can not be made publically available at this stage (McWalter and Welte have a developed methods/software for plotting incidence and sample size for assays of certain $P_{NP}$). In order to use an assay with poor $P_{NP}$ e.g. of 5%, enormous sample sizes must be used and in high prevalence and incidence settings i.e. where prevalence 20% and $I=5\%$ for sufficient power in incidence estimates. For low incidence settings, an even bigger sample size i.e. 30,000 is required and this may be unrealistic to expect for situations other than well-organized and supported national health surveys. It is critical to remembering that a high incidence setting does not necessarily confer a high prevalence population and vice versa.

Still to do:

a) More complex protocols/calibration (but not estimators) to handle violation of equal survival for P and NP, and non-regression assumptions
b) Non Normal posteriors for near zero $P_{NP}$

Still to do:

c) Estimation of bias when assumptions are violated to approximately known extent
d) Take uncertainty estimates seriously (no point estimates for BED-like assays)
3.4. Reflections on the Incidence Window Period

3.4.1. Better Methods are Required to Estimate the Incidence Window Period
There was thought given to the current models and methods for estimation of the mean incidence window period distribution. An assay with a long incidence window period tail is undesirable as the mean window period may be small but the maximum window period may be too big.

![Diagrammatic representation of the distribution of incidence window periods.](image)

Figure 5 - Diagrammatic representation of the distribution of incidence window periods.

3.4.2. Understanding the Variability of the Incidence Window Period
The incidence window period is the time spent under the threshold i.e. persisting in the state of recency. It is marked by the assignment of the threshold through assay calibration and thus is assay specific and so different assays for recent infection have different window period which makes comparison of assays rather difficult. Some mathematic modeling has taken place using a pre-existing cohort of which there are multiple biological specimens with known bleed dates and an approximate date of seroconversion. In particular, the response of the specimen to the IDU biomarker was plotted as a optical densities (OD) against time. Different estimates were obtained by varying the assay threshold and the parametrization (linear or non-linear relation), and by including or excluding the non-progressors identified in the cohort. Non-progressors may result from elite controllers or by virtue of the assay itself. The different distributions of window
periods were used to calculate population incidence. The variation of assay threshold showed no impact on incidence but the inclusion of non-progressors had a major impact on final estimates.

All currently published methods to estimate HIV incidence will require that:

- survival is weakly correlated to assay progression;
- there is a good (local) estimate of long term specificity.

On the other hand, the inference based on \( H(t), P_U, P_O, E[\pi_U] \) and \( P_{NP} \) requires that only long term specificity for relative incidence and the window period estimate for absolute incidence.

Bharat Parekh from CDC highlighted ongoing work on the BED assay incidence window period (or “duration of recency”, [DoR]) analysis and distribution of the DoR using about 3000 longitudinal specimens from >700 seroconverting subjects. The specimens represent several different cohorts/subtypes and collaborations with different groups. A non-parametric survival analysis was used to calculate the DoR. Due to ongoing analysis, the numbers are not final and could change due to selection criteria (e.g. interval between last negative and first positive) and the method employed. However, the following points summarize the major findings:

- The overall Mean DoR for all subtypes/specimens is about 190 days. This is longer than previously estimated (~155 days), possibly due to more subjects/specimens/subtypes and a non-parametric survival analysis model used for determination of the mean DoR. This approach does not exclude people due to weak response (i.e. those not crossing the threshold).
- Distribution of duration of recency is somewhat asymmetric with a tail on the right; about 64% of subjects had DoR less than 190 days. The remaining 36% have DoR longer than 190 days.
- Therefore, the mean DoR should be used for population incidence estimates only and cannot be used for individual diagnosis or to determine sensitivity and specificity of the incidence assays.
- It took about 430 days for 95% of the people to cross the threshold cutoff.
- About 5% of individuals remained below the threshold cutoff for the duration of observations (~3 years).
- A comprehensive manuscript is in development to describe these results.

3.4.3. Estimation of Standard Error

It has been proposed to use a test of equivalence to validate assays for recent HIV infection. A population is split in half and one subset is subjected to candidate assay for recent HIV infection and the other subset forms a longitudinal cohort for the directly observed incidence estimate method. This could be used to assess whether the two estimates are statistically equivalent but the same patients should not be used in calculating both estimates. Some feasibility problems
arise as this validation would be setting-specific and therefore would need to be done in each region, in different populations, at different levels of HIV incidence, etc. A careful analysis of the additional benefit to the costs should be undertaken as only well-resourced and supported countries may be able to undertake this study. For example, for a country that only tests 1,000-5,000 specimens per year for incidence, then the sample sizes of 10,000 or more required to adequately power the test of equivalence for reasonable equivalence margins would seem insurmountable.

3.4.4. Usefulness of Demographic and Health Surveys and other Cohort Studies

Demographic and Health Surveys (DHS) or national population representative surveys often include an element of HIV testing and testing for other biological markers. Those surveys have taken place in more than 30 countries and several countries will undertake a second and even third national population survey with HIV testing in the coming years. Thus there exists an opportunity to understand the sample sizes of the various DHS and determine if HIV incidence estimates can be generated from biological specimens collected during these studies, it is known that the sample sizes of DHS range from 10,000 to 50,000 households.

See Table 2 for presentation of a review of national population based surveys and DHS undertaken in 2006.

In many cases, blood specimens are taken by dried blood spot (DBS) and other surveys collect a whole blood specimen. In some recent surveys, a CD4 T+ cell count was performed and there has been discussions about use of chromatography or other laboratory methods to check for ART use as in some countries interviewers are unable to ask the question whether or not participants have taken ARV (ethics committees would not allow this question). There are several DSS (Demographic Surveillance Sites) in Africa that have been conducting regular cross-sectional surveillance studies and keep records of known seroconverting individuals.

Other pre-existing or ongoing cross-sectional studies used for surveillance purposes include MSM, IDUs, FSWs but the small sample sizes are much smaller i.e. 100-500 and often they collect a DBS for HIV testing.
Table 2: Completed population based surveys

<table>
<thead>
<tr>
<th>Country</th>
<th>Year of survey</th>
<th>Type of survey</th>
<th>Age group</th>
<th>Sample size*</th>
<th>Type of specimen</th>
<th>Linkage of test results to individuals’ characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burkina Faso</td>
<td>2003</td>
<td>DHS</td>
<td>15–49</td>
<td>15-59</td>
<td>DBS</td>
<td>Linked</td>
</tr>
<tr>
<td>Benin</td>
<td>2002</td>
<td>Household survey</td>
<td>&gt;12 years</td>
<td>&gt;15 years</td>
<td>Venous</td>
<td>Linked</td>
</tr>
<tr>
<td>Cameroon</td>
<td>2004</td>
<td>DHS</td>
<td>15–49</td>
<td>15-59</td>
<td>DBS</td>
<td>Linked</td>
</tr>
<tr>
<td>Congo</td>
<td>2003</td>
<td>Household survey, restricted to urban areas</td>
<td>15–49</td>
<td>15-49</td>
<td>DBS</td>
<td>Linked</td>
</tr>
<tr>
<td>Equatorial Guinea</td>
<td>2004</td>
<td>Household survey</td>
<td>15–49</td>
<td>15-59</td>
<td>DBS</td>
<td>Linked</td>
</tr>
<tr>
<td>Ghana</td>
<td>2003</td>
<td>DHS</td>
<td>15–49</td>
<td>15-59</td>
<td>DBS</td>
<td>Linked</td>
</tr>
<tr>
<td>Guinea</td>
<td>2005</td>
<td>DHS</td>
<td>15–49</td>
<td>15-59</td>
<td>DBS</td>
<td>Linked</td>
</tr>
<tr>
<td>Kenya</td>
<td>2003</td>
<td>DHS</td>
<td>15–49</td>
<td>15-59</td>
<td>DBS</td>
<td>Linked</td>
</tr>
<tr>
<td>Lesotho</td>
<td>2004</td>
<td>DHS</td>
<td>15–49</td>
<td>15-59</td>
<td>DBS</td>
<td>Linked</td>
</tr>
<tr>
<td>Mali</td>
<td>2001-02</td>
<td>DHS</td>
<td>15–49</td>
<td>15-59</td>
<td>DBS</td>
<td>Unlinked</td>
</tr>
<tr>
<td>Niger</td>
<td>2002</td>
<td>Household survey</td>
<td>15–49</td>
<td>15-59</td>
<td>DBS</td>
<td>Linked</td>
</tr>
<tr>
<td>Republic of South Africa</td>
<td>2003</td>
<td>Household survey</td>
<td>15–24</td>
<td>15-24</td>
<td>Oral fluids</td>
<td>Linked</td>
</tr>
<tr>
<td>African RHU</td>
<td>2005</td>
<td>Household survey</td>
<td>&gt;2 years</td>
<td>&gt;2 years</td>
<td>DBS</td>
<td>Linked</td>
</tr>
<tr>
<td>Rwanda</td>
<td>2005</td>
<td>DHS</td>
<td>15–49</td>
<td>15-59</td>
<td>DBS</td>
<td>Linked</td>
</tr>
<tr>
<td>Senegal</td>
<td>2005</td>
<td>DHS</td>
<td>15-49</td>
<td>15-59</td>
<td>DBS</td>
<td>Linked</td>
</tr>
<tr>
<td>Sierra Leone</td>
<td>2005</td>
<td>Household survey</td>
<td>15–49</td>
<td>15-59</td>
<td>DBS</td>
<td>Linked</td>
</tr>
<tr>
<td>Tanzania</td>
<td>2004</td>
<td>AIL</td>
<td>15–49</td>
<td>15-49</td>
<td>Venous</td>
<td>Linked</td>
</tr>
<tr>
<td>Uganda*</td>
<td>2004-05</td>
<td>AIL</td>
<td>15–49</td>
<td>15-49</td>
<td>Venous</td>
<td>Linked</td>
</tr>
<tr>
<td>Zambia</td>
<td>2002</td>
<td>DHS</td>
<td>15–49</td>
<td>15-59</td>
<td>DBS from venous</td>
<td>Unlinked</td>
</tr>
<tr>
<td>Zimbabwe</td>
<td>2001-02</td>
<td>Household survey</td>
<td>15–29</td>
<td>15-29</td>
<td>DBS</td>
<td>Linked</td>
</tr>
</tbody>
</table>

DHS: Demographic and Health Survey (preliminary report for Guinea, Rwanda, Senegal); DBS: dried blood spots; AIL: AIDS Indicator Survey (preliminary report for Uganda); RHU: Reproductive Health Research Unit; HSRC: Human Sciences Research Council.

*Numbers tested; for DHS/AIL countries for the 15–49 age range, except Mali (men are 15–59).


Second generation surveillance systems are increasingly moving away from anonymous unlinked testing (UAT) from ANC clinic specimens (unlinked, no results returned to the individual) and towards data from Prevention of Mother to Child Transmission (PMTCT) programmes (linked, results returned to the individual) for HIV surveillance estimates. It is now thought that the ethical implications of UAT are insurmountable in this era of improved access to anti-retroviral therapy. While more biases will be introduced by using PMTCT data, it is a more ethical route to take and the potential sample sizes may be greater but one must take into account that this population may have a higher coverage of ARV.

3.5. Assay Validation Pathway

Any future candidate assay to detect recent HIV infection would be compared to the performance of other existing assays. Specimens collected from a longitudinal cohort would serve as excellent biological material to calibrate, then validate the performance each assay. These types of specimens are difficult to acquire and in many cases the specimen set from each individual is not collected for a sufficient time period to assess performance during the early phases of infection AND the long-term infection phase. A panel of biological specimens kept in a centralised facility could be a useful resource to maintain under the aegis of WHO or another similar body.
Biological specimens will be scarce and difficult to source, certain organizations will be reluctant to allow their specimens to be used in a global manner for the purpose of assay validation. The following table outlines a possible panel of specimens that it divided into subsets and only given in a triaged manner i.e. the assay must pass the first subset to receive the second subset, etc. We need to seek consensus on what should be the order of the different specimens panels. This will take into account that specimens are precious and difficult to procure. The most precious specimens should be used last when there is a greater likelihood that the assay is performing consistently. A less-than ideal assay would be $P_{NP} > 0.05$ and $E[\pi_U] < 0.25$ years, assays with performance lesser than these characteristics should not be considered for evaluation and validation.

Consensus was not reached on the exact sequence of the specimen set but the following Tables 3 and 4 were proposed for discussion.

Table 3: Panel Composition for Assay Qualification Process: Core Set Panels

<table>
<thead>
<tr>
<th>Recent (sensitivity)</th>
<th>Triage</th>
<th># of specimens</th>
<th>Pass/Fail Criteria</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assay precision including inter- and intra-run variability</td>
<td>1st stage</td>
<td>TBD</td>
<td>CV under 10-15%</td>
<td>Maybe 5% is too optimistic, set at 10% or 15%. Refer to industry standards for determining CV. e.g. CSLI, GHTF. Note that there will be different thresholds for different assays.</td>
</tr>
<tr>
<td>ICSS: the blinded reference panel</td>
<td>1st stage</td>
<td>TBD</td>
<td>90% concordance what if recent is defined differently for different assays?</td>
<td>ICSS (Incidence Core Sample Set) sent out to assay developer, set analysed by assay developer, results to come back to central lab &amp; results analysed to see if in ball park but only in the assay developers hands. Need to include a large proportion of specimens to test for false recents, this is the most important part.</td>
</tr>
<tr>
<td>Assay developer to provide QC material (reference panel) used to develop the assay</td>
<td>2nd stage</td>
<td>TBD</td>
<td>90% concordance</td>
<td>Run ICSS and assay developer QC on HIV IA in central lab.</td>
</tr>
<tr>
<td>Linear range</td>
<td>3rd stage</td>
<td>TBD</td>
<td>r value</td>
<td>Is this applicable for all assays? i.e. avidity</td>
</tr>
<tr>
<td>Recent Infection n=0 to n=6 months</td>
<td>3rd stage</td>
<td>TBD</td>
<td>90% concordance</td>
<td>Should be high?</td>
</tr>
<tr>
<td>Established n=&gt;6 months</td>
<td>3rd stage</td>
<td>TBD</td>
<td>90% concordance</td>
<td>Should be high?</td>
</tr>
<tr>
<td>Clade</td>
<td>3rd stage</td>
<td>TBD</td>
<td>Equivalence*</td>
<td></td>
</tr>
<tr>
<td>Cutoff (determines the WP or threshold for recent infection)</td>
<td>3rd stage</td>
<td>TBD</td>
<td>6 months (180 days)</td>
<td>Would like this to align with above subsets but unwilling to assign as 4 months or as 8 months.</td>
</tr>
</tbody>
</table>

For CV:1- 3 specimens from each subtype, tested 15 times in one run, on each of three days (total = 3 runs) in a row, using 2 batches. With a consensus on the different subtypes.

---

2 Assumes that the ability to detect clades is equivalent but is there any effect on the false recent rate? Can we assume this?
Table 4 - Desired Characteristics of Specimen Panels : 2nd Tier Panels

<table>
<thead>
<tr>
<th></th>
<th>False Recent (specificity)</th>
<th>False Established (specificity)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Priority</td>
<td># of samples</td>
</tr>
<tr>
<td>ARV</td>
<td>High</td>
<td>TBD</td>
</tr>
<tr>
<td>AIDS</td>
<td>High</td>
<td>TBD</td>
</tr>
<tr>
<td>Known false recent samples</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(as determined by known assays)</td>
<td></td>
<td>TBD</td>
</tr>
<tr>
<td>Different modes of transmission</td>
<td>High</td>
<td>TBD</td>
</tr>
<tr>
<td>(IDU vs. sexual)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probably not required</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Various age groups</td>
<td>Medium</td>
<td>TBD</td>
</tr>
<tr>
<td>(e.g. adolescents, adults)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probably not required</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both genders</td>
<td>Medium</td>
<td>TBD</td>
</tr>
<tr>
<td>Probable not required</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Different ethnicities (based</td>
<td>Low</td>
<td>TBD</td>
</tr>
<tr>
<td>on HLA types)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Different geographic regions</td>
<td>High</td>
<td>TBD</td>
</tr>
<tr>
<td>Pregnant women (e.g. currently</td>
<td>High</td>
<td>TBD</td>
</tr>
<tr>
<td>pregnant)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parity (post-partum)</td>
<td>Low</td>
<td>TBD</td>
</tr>
<tr>
<td>Individuals with co-</td>
<td>High</td>
<td>TBD</td>
</tr>
<tr>
<td>infections (e.g. TB, malaria,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>other infectious diseases)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interfering Substances?</td>
<td>Low</td>
<td>TBD</td>
</tr>
</tbody>
</table>

TBD: to be determined

Specifying specimen requirements: prospective specimen collection of serum and plasma and dried blood spot from each individual.

The table below would be a useful tool to assist programme implementers to understand the necessary sample sizes required to observed an increase or decrease in HIV incidence at the varying HIV prevalences.

Table 5 - Sample Size Calculations

<table>
<thead>
<tr>
<th>Sample size required</th>
<th>Prevalence</th>
<th>&lt; 1%</th>
<th>1 -3%</th>
<th>3-5%</th>
<th>5 -10%</th>
<th>10-15%</th>
<th>&gt; 15%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 - 2%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 - 3%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt; 3%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The sample sizes required for an accurate estimate of HIV incidence will depend heavily on the accepted co-efficient of variation (CV) for incidence estimates, the characteristics of the assay and the incidence of the population.
3.6. Policy-Relevant Questions about Assays for Recent HIV Infection

A country may want to know approximately has HIV incidence gone up/down rather than an exact measurement of by how much (%). This may require less rigor required for incidence estimates than an exact measurement.

Pragmatic guidance on the implementation of assays (or a testing algorithm) to detect recent HIV infection is required, certain questions will persist about the misclassification using the currently available assays with poor long-term specificity. For instance, if one would want to see a change in incidence overtime, it should be recommended that it is best to use the same assay (or testing algorithm) for recent HIV infection and be performed in the same laboratory, using the same protocol for any misclassification calculation. It may be possible to ask survey participants for if AIDS has been diagnosed, if they are currently on ARV (or have been in the past), etc. and these individuals could then be can excluded from the calculations, so that the long-term specificity can be accounted for. There was a suggestion to engage Abbott GmbH or other companies such as Johnson and Johnson (Ortho Clinical Diagnostics), based on evidence from the FHI market assessment, to take on the avidity "kit" as a formally recognized modified protocol for the AxSym, Architect, and Vitros assays.

4. Conclusions:

Alex Welte will present the outcomes of this meeting at the meeting organized by Family Health International to be held 14-15 May 2009 in Chapel Hill, USA.

The conclusions in brief:

1. The naïve incidence estimator is valid in the case that the false positive rate (false recent rate) of a recent infection testing algorithm is zero.

2. The meaning of ‘recently infected’ in such a case is defined by the assay itself, which is a conceptual break from a previously widely used definition not recently infected as defined by infection occurring within a specified time. This removes the need to speak of sensitivity and specificity of the assay in the usual diagnostic sense, as the assay is no longer a proxy for any other condition. Sensitivity and specificity would both be perfect (value 1).

3. The definition of ‘recently infected’ (the estimated number of individuals in this category being the numerator of the estimator) needs careful redefinition in the case of a non-zero false positive rate.

4. This new definition of recent infection is based on observing the distribution of times spent in the ‘assay-defined’ RITA-positive state, among individuals where this time is finite. Individuals who spent all time post infection in the RITA-positive state are formally assigned a time spent in the ‘recent’ state, drawn from the distribution of waiting times in the population where it is finite.
5. This recently infected state is not observable in all individuals and the assay is unable to
discern which individuals are correctly characterized. A false positive rate can
nevertheless be formally defined, and the number of recently infected individuals can be
estimated without bias, given the exact value of the false positive rate.

6. The general estimator, expressed in terms of the number of recently infected individuals,
is the same as the naïve estimator. It can be unpacked using the method for estimating
the number of recently infected individuals.

7. There is both bias and loss of statistical power that attends to the use of the general
formula required to deal with a non-zero false recent rate. This has been computed
explicitly for a number of scenarios. The calculations are complex, as statistical power is
dependant on numerous factors (value of incidence, mean window period, false recent
rate, the precision with which window period and false recent rate are known, and sample
size). The particular power calculations to be presented were agreed to, and calculations
performed by Thomas McWalter and Alex Welte.

8. A false recent rate of less than about 2% appears to be required to offer a substantial
advance over the current assays, which have false recent rates of about 5%.

9. The fundamental trade-off in assay design is to find biomarkers which persist in the RITA-
positive regime for as long as possible (making individuals in this category easy to count
with precision) while keeping the false recent rate low.

10. There is no meaningful incidence-related statistical analysis that can be done without
knowledge of the false recent rate. However, the window period is not strictly required to
be known in order to set up a test for variation of incidence between two surveys.

11. An estimate of the window period is required in order to produce a quantitative incidence
estimate.

12. Issues of calibration were discussed but there are several subtleties about both the
window period and false positive rate which cannot be addressed in a short presentation,
so while this was discussed, it was not treated in the presentation for Chapel Hill. It was
noted that the false recent rate is in principle directly estimatable from the proportion of
RITA-positive results in a sample of specimens with known non-recent infection, without
the need for follow up. Window period estimates have previously mainly been produced in
seroconverter cohorts, involving intensive follow-up. An alternative analysis of such data,
involving an additional ‘lower’ threshold, was discussed, which may successfully trade-off
a reduction in usable individual time series against removal of bias and uncertainty due to
substantial inter-donation intervals. It was also noted how data from long interval follow up
can be used to provide window period estimates, given an independent estimate of the
false recent rate.
This meeting report will be presented to the wider WHO Technical Working Group on HIV Incidence Assays, when they next meeting in Cape Town, South Africa on 16-17 July 2009.

5. Associated Technical Documents

1. Methodologic Guidance for Validation of Existing and Future HIV Incidence Assays
   (Draft Version 5.0, 13 October 2008)
2. Meeting Report from WHO Working Group on HIV Incidence Assays held in Mexico City
   (2-3 August 2008)

6. Bibliography


