Streamlining HIV Testing for HIV Preexposure Prophylaxis

Juan V. Guanira, Teri Leigler, Esper Kallas, Mauro Schechter, Usha Sharma, David Glidden, Robert M. Grant for the iPrEx Study Team

Investigaciones Médicas en Salud, Lima, Peru; University of California, San Francisco, California, USA; Gladstone Institutes, San Francisco, California, USA; University of Sao Paulo, Sao Paulo, Brazil; Praca Onze, Rio de Janeiro, Brazil; NIH/NIAID, Bethesda, Maryland, USA

HIV-testing algorithms for preexposure prophylaxis (PrEP) should be optimized to minimize the risk of drug resistance, the time off PrEP required to evaluate false-negative screening results, and costs and to expedite the start of therapy for those confirmed to be infected. HIV rapid tests (RTs) for anti-HIV antibodies provide results in less than 1 h and can be conducted by nonlicensed staff at the point of care. In many regions, Western blot (WB) testing is required to confirm reactive RT results. WB testing, however, causes delays in diagnosis and adds expense. The iPrEx study evaluated the safety and efficacy of daily oral emtricitabine-tenofovir disoproxil fumarate among HIV-seronegative men and transgender women who have sex with men: HIV infection was assessed with two RTs plus WB confirmation, followed by HIV-1 plasma viral load testing. During the iPrEx study, there were 51,260 HIV status evaluations among 2,499 volunteers using RTs: 142 (0.28%) had concordant positive results (100% were eventually confirmed) and 19 (0.04%) had discordant results among 14 participants; 11 were eventually determined to be HIV infected. A streamlined approach using only one RT to screen and a second RT to confirm (without WB) would have had nearly the same accuracy. Discrepant RT results are best evaluated with nucleic acid testing, which would also increase sensitivity.

Preexposure prophylaxis (PrEP) with oral emtricitabine/tenofovir (FTC/TDF) is safe and effective for preventing the acquisition of HIV infection in men and transgender women who have sex with men and heterosexual men and women. HIV testing prior to starting PrEP is required to avoid resistance that occurs if PrEP is started after HIV infection is already established. Optimal testing algorithms are needed to minimize the risk of drug resistance that occurs if PrEP is started after systemic infection has already occurred, the time off PrEP due to false-positive test results, and the testing costs and burden on the PrEP user and health care system.

Rapid tests (RTs) for HIV antibodies have several advantages. These tests provide an initial diagnosis within an hour; however, in many settings, regulatory authorities require that the results of rapid tests be confirmed by either Western blotting (WB) or indirect immunofluorescence assay (1, 2). These confirmatory tests are performed at laboratories away from the point of care, which delays diagnosis and so increases the time off PrEP (if the rapid tests are falsely reactive) or delays HIV treatment (if the rapid tests are truly reactive).

In 2009, the Association of Public Health Laboratories (APHL) issued a status report on algorithms for HIV testing. In this report, the APHL discussed the importance of point-of-contact (PoC) RTs, and they proposed algorithms with one to three RTs to test for HIV in areas of high and low prevalence (3). Another study showed that by definitively diagnosing HIV at the PoC, health care providers could more promptly link people with HIV to health care resources (4), to shorten the window before they begin treatment. We utilized specimens and test results from the iPrEx study, a randomized placebo-controlled trial of daily oral FTC/TDF PrEP that was conducted in Asia, Africa, South America, and North America.

MATERIALS AND METHODS

Study population. The iPrEx study was performed at 11 sites in 6 countries (Peru, Ecuador, Brazil, South Africa, Thailand, and the United States). The study population consisted of 2,499 men who have sex with men (MSM) and male-to-female transgender women (MFTW) who were at least 18 years old and at risk for acquiring HIV. This cohort was followed with monthly visits for an average of 1.7 years, as described previously (5).

The iPrEx study obtained ethical approval at each study site, within each site’s respective country, and centrally from the Committee on Human Research (CHR) at the University of California, San Francisco (UCSF). Written informed consent was obtained from all participants prior to enrollment in the study.

HIV testing. Blood samples for HIV testing were obtained by either finger stick or venipuncture, depending on procedures that were specific for each visit. RTs and WB tests were performed at each study site. Specimens for testing plasma for HIV RNA were frozen at −70°C and shipped to a centralized laboratory. To ensure comparability across sites, the study team implemented a unified program for laboratory quality, which included a proficiency-testing program.

Rapid HIV antibody testing outside the United States was performed with a combination of Determine HIV 1/2 (Alere, Livermore, CA, USA) and SD Bioline HIV 1/2 3.0 (Standard Diagnostics, Inc., Gyeonggi-do, Republic of Korea). Within the United States, rapid HIV antibody testing was performed with a combination of OraQuick Advance rapid HIV 1/2 antibody test (OraSure Technologies, Inc., Bethlehem, PA, USA) and Clearview HIV 1/2 Stat-Pak (Alere, Livermore, CA, USA). Within 15 days of collecting samples, serology was confirmed by Western blotting (WB) with either the New LAV Blot I (Genetic Systems, Bio-Rad, Hercules, CA)
TABLE 1 Agreement between rapid test results during the iPrEx study follow-up period

<table>
<thead>
<tr>
<th>Rapid test 2</th>
<th>No. of results with rapid test 1 that were a:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>No. of negative results</td>
<td>50,099</td>
<td>13</td>
</tr>
<tr>
<td>No. of positive results</td>
<td>6</td>
<td>142</td>
</tr>
<tr>
<td>Total</td>
<td>50,105</td>
<td>155</td>
</tr>
</tbody>
</table>

a There were 19 (0.04%) discordant results between RTs, giving a correlation between RTs of r = 0.9374.

or Cambridge Biotech HIV-1 WB kit (Maxim Biomedical, Rockville, MD), depending on the site. To determine the visit of first evidence of infection, archived samples from preseroconversion visits underwent PCR testing for HIV plasma RNA using the Abbott RealTime HIV-1 assay (40 copies/ml limit of quantitation; Abbott Diagnostics, Des Plaines, IL) or Amplicor HIV-1 Monitor test (50 copies/ml limit of quantitation; Roche Applied Sciences, Pleasanton, CA).

iPrEx algorithm. The iPrEx algorithm for diagnosing HIV was conducted by the approved protocol algorithm. Briefly, two simultaneous rapid tests (RTs) from two different manufacturers were performed (step 1). If one or more RTs were reactive, a confirmatory WB was performed; otherwise the result was considered negative (step 2). If the WB result was negative, the participant was considered noninfected (step 3), and if positive, they were considered infected (step 4). If the WB result was indeterminate, a second sample was drawn to repeat the rapid testing and Western blot testing (step 5). If discrepancies occurred, a third sample was obtained to repeat the algorithm (step 6). If the discrepancies persisted, HIV RNA testing was performed on 3 time points as a tie breaker, and all test results were sent to an expert HIV acquisition adjudication committee for a final decision regarding HIV status (step 7) (see the supplemental material).

Theoretical algorithms. Seven theoretical testing algorithms were developed for PoC RT as proposed by the APHL (3) and were assessed for performance by analyzing receiver operating characteristics (ROC) curves.

Data management and statistical methods. Information regarding the results of HIV testing was collected on specific case report forms at each site. These forms were faxed to a centralized data management center, where the data passed through a quality control/quality assurance (QC/QA) process before being locked for analyses. The QA/QC program was implemented at each site under the supervision of the study team and the Division of AIDS (DAIDS). Data regarding RTs, WB, plasma HIV RNA, patient identification (PTID), date of visit, and date of seroconversion were extracted from the main database and analyzed separately. Follow-up visits from the randomized phase of the iPrEx trial from 2007 through November 2010 were selected for analysis. Two-by-two tables and correlation measures evaluated RT performance compared to the results for RNA testing as the gold standard.

RESULTS

Within the iPrEx study, we investigated the performance of two different RTs within four different algorithms, with or without WB confirmation. Two of these algorithms involved sequentially testing for HIV with both RTs. During the period for follow-up in the iPrEx study, we evaluated the HIV status of 2,499 participants at 51,260 time points. Each participant had a median of 25 evaluations during a mean of 21 months of follow-up (median of 20). Of the participants, 147 (5.9%) were diagnosed with HIV; 10 were seronegative at enrollment but were later found to be infected at enrollment using RNA testing, 131 were infected and diagnosed during the follow-up period, and 6 were infected and diagnosed at posttreatment visits. All infections were eventually confirmed by WB and HIV RNA detection, including laboratory evidence of HIV infection at 2 or more visits.

Globally, the Determine/Bioline and OraQuick/StatPak tests diagnosed 92.3% and 7.7% of the HIV infections, respectively. There were no differences in performance between these pairs. Of the results, 50,099 (99.68%) were concordant nonreactive, 142 (0.28%) were concordant reactive, and 19 (0.04%) were discordant (r = 0.94) (Table 1). Among concordant reactive results, WB results were negative for 4 (2.8%), positive for 107 (75.3%), and indeterminate for 28 (19.7%); 3 were not tested by WB because participants were either lost to follow up or refused phlebotomy. All concordant reactive results were confirmed to be HIV infected by HIV plasma RNA testing, performed according to algorithms defined in the protocol.

The 19 RT discordant results came from 14 participants, of whom 11 were eventually diagnosed with HIV by Western blotting and RNA detection. Among the 11 who were HIV infected at the visit when RT results were discordant, 7/11 (63.6%) had a positive WB, 2/11 (18.2%) had a negative WB, and 2/4 (50%) had...

FIG 1 Theoretical PoC algorithm for 2 parallel rapid tests without Western blot (algorithm A), using the 50,260 time point evaluations from the iPrEx study.
an indeterminate WB (Table 2). Three HIV-uninfected persons had discordant RT results at a total of 8 visits: WB was negative at 6 visits (75%), indeterminate at 2 visits (25%), and positive at no visit.

Analysis of the performance of individual RTs revealed that Determine produced 7 false-positive (FP) results in 3 participants and Oraquick produced 1 FP result, while Bioline and Stat-Pak did not produce FP results. We also performed retrospective analysis of preseroconversion specimens, which yielded 26 samples where both RTs had produced negative results (on whole blood) but
plasma HIV RNA was detected. Of these, 10 were already infected at study enrollment (5). All false-negative (FN) results were found with Determine/Bioline, and none were found with Oraquick/Stat-Pak.

The predicted performances of the alternative algorithms were considered. Algorithm A (simultaneous RT, no WB) misclassified 20.7% (37/179) of infected samples and 0.00% (0/50,081) of noninfected samples (Fig. 1). In algorithm B, where WB was added to test all samples with at least one reactive RT, 17.9% (32/179) of the HIV infections were missed and 0.004% (2/50,081) of time points at which the subject was HIV uninfected were misdiagnosed (Fig. 2). Algorithm C, where WB is used only at time points with discordant RT results, improves sensitivity slightly, to 15.6% (28/179), because 4 people with concurrent reactive RTs had negative WB results and would not have had WB testing in this algorithm. When the sequential algorithms were used, algorithm D misclassified 0/50,081 of noninfected samples and 31/179 of infected samples. Algorithms B, C, and D were further evaluated using viral load instead of WB to confirm reactive RT. Algorithm B correctly classified 2 FP and 2 FN cases missed by algorithm B, and algorithm C correctly classified one additional FP missed by algorithm C (Fig. 3). Algorithm D correctly classified one FP and 2 FN cases. Changes in sensitivity, specificity, positive predictive value, negative predictive value, and ROC area results were not significantly different from the results for the initial algorithms (Table 3).

**DISCUSSION**

Two concordant positive RTs had 100% specificity. Adding a WB to confirm concordant RTs did not improve specificity. Indeed, the WB was falsely negative in 2.9% and indeterminate in 28% of time points, requiring additional testing that eventually confirmed infection in all cases. These data are consistent with other reports. Haukoos and colleagues found that, among an emergency room population, a sequential algorithm initially yielded five positive results, which were comparably likely to be due to false positive results (6). In another sequential algorithm at the New Jersey Department of Public Health, all discordant-positive results (355 samples) were confirmed by WB, and all discordant results (8 samples) were identified as FP by WB (7). Furthermore, a two-test algorithm conducted by the Centers for Disease Control (CDC) distinguished almost all true-positive results from false-positive results (8). These findings imply that two-test algorithms can correctly classify one FP missed by algorithm C (Fig. 3). Algorithm D correctly classified one FP and 2 FN cases. Changes in sensitivity, specificity, positive predictive value, negative predictive value, and ROC area results were not significantly different from the results for the initial algorithms (Table 3).

Two concordant positive RTs had 100% specificity. Adding a WB to confirm concordant RTs did not improve specificity. Indeed, the WB was falsely negative in 2.9% and indeterminate in 28% of time points, requiring additional testing that eventually confirmed infection in all cases. These data are consistent with other reports. Haukoos and colleagues found that, among an emergency room population, a sequential algorithm initially yielded five positive results, which were comparably likely to be due to false positive results (6). In another sequential algorithm at the New Jersey Department of Public Health, all discordant-positive results (355 samples) were confirmed by WB, and all discordant results (8 samples) were identified as FP by WB (7). Furthermore, a two-test algorithm conducted by the Centers for Disease Control (CDC) distinguished almost all true-positive results from false-positive results (8). These findings imply that two-test algorithms can correctly classify one FP missed by algorithm C (Fig. 3). Algorithm D correctly classified one FP and 2 FN cases. Changes in sensitivity, specificity, positive predictive value, negative predictive value, and ROC area results were not significantly different from the results for the initial algorithms (Table 3).

**TABLE 3** Sensitivity, specificity, positive predictive value, negative predictive value, and ROC area for each of the seven theoretical algorithms

<table>
<thead>
<tr>
<th>Parameter</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>B’</th>
<th>C’</th>
<th>D’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>85.47</td>
<td>81.82</td>
<td>84.36</td>
<td>82.68</td>
<td>85.47</td>
<td>85.47</td>
<td>82.68</td>
</tr>
<tr>
<td>Specificity</td>
<td>99.98</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>PPV</td>
<td>95.03</td>
<td>98.69</td>
<td>98.69</td>
<td>99.33</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>NPV</td>
<td>99.95</td>
<td>99.94</td>
<td>99.94</td>
<td>99.94</td>
<td>99.94</td>
<td>99.94</td>
<td>99.94</td>
</tr>
<tr>
<td>ROC area</td>
<td>0.93 (0.93–0.93)</td>
<td>0.91 (0.91–0.91)</td>
<td>0.92 (0.92–0.92)</td>
<td>0.91 (0.91–0.92)</td>
<td>0.93 (0.93–0.93)</td>
<td>0.93 (0.93–0.93)</td>
<td>0.91 (0.91–0.92)</td>
</tr>
</tbody>
</table>

*Values are in comparison to the results for the gold standard of HIV-1 viral load. PPV, positive predictive value; NPV, negative predictive value; ROC, receiver operating characteristics.*

The predicted performances of the alternative algorithms were considered. Algorithm A (simultaneous RT, no WB) misclassified 20.7% (37/179) of infected samples and 0.00% (0/50,081) of noninfected samples (Fig. 1). In algorithm B, where WB was added to test all samples with at least one reactive RT, 17.9% (32/179) of the HIV infections were missed and 0.004% (2/50,081) of time points at which the subject was HIV uninfected were misdiagnosed (Fig. 2). Algorithm C, where WB is used only at time points with discordant RT results, improves sensitivity slightly, to 15.6% (28/179), because 4 people with concurrent reactive RTs had negative WB results and would not have had WB testing in this algorithm. When the sequential algorithms were used, algorithm D misclassified 0/50,081 of noninfected samples and 31/179 of infected samples. Algorithms B, C, and D were further evaluated using viral load instead of WB to confirm reactive RT. Algorithm B correctly classified 2 FP and 2 FN cases missed by algorithm B, and algorithm C correctly classified one additional FP missed by algorithm C (Fig. 3). Algorithm D correctly classified one FP and 2 FN cases. Changes in sensitivity, specificity, positive predictive value, negative predictive value, and ROC area results were not significantly different from the results for the initial algorithms (Table 3).

**DISCUSSION**

Two concordant positive RTs had 100% specificity. Adding a WB to confirm concordant RTs did not improve specificity. Indeed, the WB was falsely negative in 2.9% and indeterminate in 28% of time points, requiring additional testing that eventually confirmed infection in all cases. These data are consistent with other reports. Haukoos and colleagues found that, among an emergency room population, a sequential algorithm initially yielded five positive results, which were comparably likely to be due to false positive results (6). In another sequential algorithm at the New Jersey Department of Public Health, all discordant-positive results (355 samples) were confirmed by WB, and all discordant results (8 samples) were identified as FP by WB (7). Furthermore, a two-test algorithm conducted by the Centers for Disease Control (CDC) distinguished almost all true-positive results from false-positive results (8). These findings imply that two-test algorithms can correctly classify one FP missed by algorithm C (Fig. 3). Algorithm D correctly classified one FP and 2 FN cases. Changes in sensitivity, specificity, positive predictive value, negative predictive value, and ROC area results were not significantly different from the results for the initial algorithms (Table 3).

**TABLE 3** Sensitivity, specificity, positive predictive value, negative predictive value, and ROC area for each of the seven theoretical algorithms

<table>
<thead>
<tr>
<th>Parameter</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>B’</th>
<th>C’</th>
<th>D’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>85.47</td>
<td>81.82</td>
<td>84.36</td>
<td>82.68</td>
<td>85.47</td>
<td>85.47</td>
<td>82.68</td>
</tr>
<tr>
<td>Specificity</td>
<td>99.98</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>PPV</td>
<td>95.03</td>
<td>98.69</td>
<td>98.69</td>
<td>99.33</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>NPV</td>
<td>99.95</td>
<td>99.94</td>
<td>99.94</td>
<td>99.94</td>
<td>99.94</td>
<td>99.94</td>
<td>99.94</td>
</tr>
<tr>
<td>ROC area</td>
<td>0.93 (0.93–0.93)</td>
<td>0.91 (0.91–0.91)</td>
<td>0.92 (0.92–0.92)</td>
<td>0.91 (0.91–0.92)</td>
<td>0.93 (0.93–0.93)</td>
<td>0.93 (0.93–0.93)</td>
<td>0.91 (0.91–0.92)</td>
</tr>
</tbody>
</table>

*Values are in comparison to the results for the gold standard of HIV-1 viral load. PPV, positive predictive value; NPV, negative predictive value; ROC, receiver operating characteristics.*
study, they estimated that 25% of people diagnosed with HIV were not linked to care in the United States, which increases the risk of their HIV infection progressing to AIDS or being transmitted to uninfected partners. Replacing WB confirmation with viral load testing could shorten the time to linkage to care and treatment by 2 weeks or more. A considerable number of people do not return to check the confirmatory report (4, 16, 17).

PrEP programs are most valuable in populations with a high incidence of HIV infection (18, 19). To evaluate these populations, we need to correctly diagnose HIV at PoC locations with testing algorithms that can be performed rapidly and effectively (20, 21), without expensive equipment or specialized laboratory technicians. Guidelines for PrEP (22) propose that people undergo HIV testing every 3 months while taking antiretrovirals. Based on our findings, we recommend that health care providers adopt a sequential approach for diagnosing HIV by first screening people with one RT and then confirming a positive diagnosis with a second RT. If the two RT results are discordant, the diagnosis is best confirmed by nucleic acid testing, with use of WB reserved for situations in which the nucleic acid testing is negative. With this algorithm, we can avoid delays in diagnosis and treatment and reduce the incidence of HIV transmission.

ACKNOWLEDGMENTS

We sincerely appreciate the iPrEx study participants, staff, and investigators that made this research possible.

The iPrEx study was sponsored by the U.S. National Institutes of Health under grant number 4U01 AI064002 and cofunded by the Bill and Melinda Gates Foundation under grant number 48162.

REFERENCES