HIV Vaccine Efficacy Trial Design: The Post-STEP Era

Steven G. Self, PhD
Director, HVTN Statistical Center

October 14, 2008
Imperatives for efficacy trial designs in post-STEP era

- More focused (test fewer hypotheses) (2 minutes)
- More sensitive (to detect real vaccine effects) (4 minutes)
- More safe (2 minutes)
- As reliable/interpretable! (2 minutes)
Previous TOC designs tested multiple primary and secondary hypotheses

- Trial populations and primary efficacy hypotheses
  - STEP (low Ad5 NAb, overall)
  - Phambili (men, women)
  - PAVE100 (3 human and viral populations)
- Secondary hypotheses
  - Correlates of protection
  - Sieve analyses

Test efficacy hypotheses in a single, biologically and epidemiologically efficient population
Previous TOC designs used co-primary endpoints

- STEP, Phambili and PAVE100
  - HIV VL Setpoint
  - HIV Infection

Test efficacy hypothesis only for VL endpoint
STOC Design is a more focused TOC Design

STOC requires smaller trial for VL test of same power

- **TOC**
  - Co-primary endpoints (Infection, VL)
  - Split Type I error between endpoints (eg .025 for each endpoint)
  - **40 evaluable VL endpoints required** for 80% power against ΔVL of 1 log (assume VE_S = 0)

- **STOC**
  - Single primary endpoint (VL)
  - Type I error all spent on single test (eg .05)
  - **34 evaluable VL endpoints required** for 80% power against ΔVL of 1 log (assume VE_S = 0)

Use STOC-type trial designs
Additional strategies to improve sensitivity of STOC designs

- Definition of VL endpoint
  - Focus
  - Reduce variability

- Alternative Type I error rates
  - Use Type I and Type II error rates appropriate for screening

- Use of vaccine response rate in design exercise
  - Predictive model of VL variation in vaccinees
  - Calibrate overall effect size
Primary Endpoint: “VL Setpoint”

- **Definition of evaluable VL endpoint**
  - Infection endpoint evaluable only if diagnosed subsequent to clinic visit of final immunization
    - Increase potential signal by considering infections after full immunization
  - VL endpoint the average of multiple pre-ART measurements within pre-specified window of quasi-stable VL
    - Reduce variation of endpoint measurement by restriction to epoch of quasi-stable VL
    - Reduce variation of endpoint by averaging multiple measurements
  - Consider ART-initiation guidelines in selection of study population
    - Rate of ART initiation within first year of Dx but without standard clinical/biomarker indicators
Type I and Type II Error Rates

• For screening trials of candidates in sparse pipelines
  – Strict control of Type I error rates (false positives) to very low levels is less critical than in TOC and pivotal trials
  – Strict control of Type II error rates (false negatives) to very low levels is more critical than in TOC and pivotal trials

• Consider less stringent alternatives to standard Type I error rates (eg, $\alpha = 0.10$ or higher)

• Consider more stringent alternatives to standard Type II error rates (eg, $\beta = 0.10$ or lower)
Minimum Detectable Effect Size at 90% Power* ($\Delta_{90}$)

Assume 36 evaluable VL endpoints (18:18)

<table>
<thead>
<tr>
<th>Type I Error (2-sided)</th>
<th>$\Delta_{90}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.025</td>
<td>1.21</td>
</tr>
<tr>
<td>.05</td>
<td>1.12</td>
</tr>
<tr>
<td>.10</td>
<td>1.00</td>
</tr>
<tr>
<td>.15</td>
<td>0.92</td>
</tr>
<tr>
<td>.20</td>
<td>0.87</td>
</tr>
</tbody>
</table>

*Wilcoxon rank sum test; 16,000 simulations
## Variability of VL Endpoint

### Standard deviations of VL endpoints in studies of MSM

<table>
<thead>
<tr>
<th>Study</th>
<th>Group</th>
<th>Number Infections</th>
<th>VL endpoint definition</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>MACS</td>
<td>Natural history</td>
<td>269</td>
<td>Lyles et al. 2000</td>
<td>.75</td>
</tr>
<tr>
<td>Vax004</td>
<td>Ph 3</td>
<td>179 Vx + Plc</td>
<td>Avg of Week 8 and 16 PD</td>
<td>.80</td>
</tr>
<tr>
<td>STEP</td>
<td>Ph 2b</td>
<td>44 Vx + Plc</td>
<td>Avg of Week 8 and 12 PD</td>
<td>.88</td>
</tr>
</tbody>
</table>
Vaccine Response Rate ("Take"!?)

- Is it unrealistic to expect vaccine effect on VL among vaccinees who show no evidence of meaningful response to vaccine?

- If not, use vaccine response rate in two ways
  - Prediction of variability of VL measurements in vaccinees to be used in design calculations
  - Calibration of overall effect (vaccinees vs controls)

- Example:
  - Assume
    - vaccine response rate is \( p \)
    - variance of VL measurement is same among placebos and among vaccinees who did not response to vaccine \( (\sigma^2) \)
    - \( \log_{10} \) VL is lower among responders by \( \delta \)
  - Then
    - Overall vaccine effect on VL is \( \Delta = p\delta \)
    - Distribution of VL endpoint among vaccinees is mixture of 2 distributions and is more variable than that for controls
Rate of Vaccine Response (p) and Variability of VL among Vaccinees

Viral Load Set-Point (log10 copies/ml)

Assume VL distribution for vaccinees is mixture of vaccine “responders” (proportion p) and “non-responders”; is more variable than for non-vaccinees
## Effect of Vaccine Response Rate on Power and Calibration of overall Effect Size

Power to detect an overall mean $\log_{10}$ VL difference ($\Delta$) (mean VL difference in “takers” $\delta$)*

<table>
<thead>
<tr>
<th>Vaccine proportion who “take” $p$</th>
<th>$\Delta = 0.6$</th>
<th>$\Delta = 0.8$</th>
<th>$\Delta = 1.0$</th>
<th>$\Delta = 1.4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.55 (1.2)</td>
<td>0.69 (1.6)</td>
<td>0.77 (2.0)</td>
<td>0.82 (2.8)</td>
</tr>
<tr>
<td>0.6</td>
<td>0.58 (1.0)</td>
<td>0.76 (1.33)</td>
<td>0.85 (1.67)</td>
<td>0.92 (2.33)</td>
</tr>
<tr>
<td>0.7</td>
<td>0.60 (0.86)</td>
<td>0.80 (1.14)</td>
<td>0.90 (1.43)</td>
<td>0.97 (2.0)</td>
</tr>
<tr>
<td>0.8</td>
<td>0.62 (0.75)</td>
<td>0.82 (1.0)</td>
<td>0.92 (1.25)</td>
<td>0.99 (1.75)</td>
</tr>
<tr>
<td>0.9</td>
<td>0.63 (0.67)</td>
<td>0.84 (0.89)</td>
<td>0.94 (1.11)</td>
<td>1.0 (1.56)</td>
</tr>
<tr>
<td>1.0</td>
<td>0.64 (0.60)</td>
<td>0.85 (0.80)</td>
<td>0.96 (1.0)</td>
<td>1.0 (1.4)</td>
</tr>
</tbody>
</table>

*2-sided $\alpha = 0.10$; power = 90%; 36 evaluable endpoints (18 in each group) Wilcoxon rank sum test; 10,000 simulations
Safety Monitoring

• Monitor for higher HIV infection rate in vaccine group
  – Study population defined to ensure equipoise
  – Maximally vigilant “continuous” monitoring (after each confirmed WITT infection event)
  – Safety monitoring strategy developed for (Heyse et al, 2008) and used effectively in rotovirus vaccine efficacy trial

• Formally test for elevated HIV infection relative risk RR (vaccine group relative to placebo) after each observed infection
  – Harm indicated if \( H_0: \text{RR} \leq 1 \) is rejected in favor of \( H_1: \text{RR} > 1 \)
  – Control overall Type I error of tests for harm to no more than 0.05 (or 0.10) 1-sided
  – Use same statistical criterion for each test performed; do not use very conservative criteria for early tests as in sequential testing for efficacy
    • Overall “0.05” rule: Stop if individual 1-sided test* gives \( p \leq 0.015 \)
    • Overall “0.10” rule: Stop if individual 1-sided test* gives \( p \leq 0.031 \)

* Exact binomial test with unadjusted/nominal p-value
Stopping Boundaries for the Overall “0.05” and “0.10” Rules

Examples of Vx:Plc Infection splits that just hit stopping boundary:

0.05 rule:
7:0, 9:1, 16:4, 22:8

0.10 rule:
6:0, 9:1, 15:5, 21:9
“Power” of the “0.05” and “0.10” Rules for Correctly Stopping Early

Probability of stopping early (before the 45th WITT infection) for a range of possible true relative risks (RRs) of HIV infection (vaccine/placebo)

<table>
<thead>
<tr>
<th>Stopping Rule (overall FP rate)</th>
<th>RR=1</th>
<th>RR=1.5</th>
<th>RR=2.0</th>
<th>RR=2.5</th>
<th>RR=3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05 rule</td>
<td>.049</td>
<td>.29</td>
<td>.61</td>
<td>.82</td>
<td>.93</td>
</tr>
<tr>
<td>0.10 rule</td>
<td>.093</td>
<td>.42</td>
<td>.73</td>
<td>.90</td>
<td>.96</td>
</tr>
</tbody>
</table>
Reliability/Interpretability

• STOC design “unusual” for a randomized controlled trial as no aspect of primary analysis is based on entire randomized population
  – Comparison of highly-selected small subset of randomized participants who become infected can be subject to bias
  – Even if bias is due to measured confounders, small numbers of subjects with evaluable endpoints may thwart attempts for post-hoc statistical adjustment

• Sensitivity analyses can help assess extent to which observed effect may be due to selection bias
  – Depends on range of plausible values for RR
  – Estimate of RR from STOC trial highly variable

• Significant effects on VL in STOC trial may not be reliably attributed to direct biological effect on VL
Acknowledgements

- Peter Gilbert
- Devan Mehrotra
- Dean Follmann
- Jerry Sadoff